UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA TÉCNICA DE TELECOMUNICACIONES: SONIDO E IMAGEN

PROYECTO FIN DE CARRERA

Predicción del coste sanitario en pacientes psiquiátricos mediante técnicas de aprendizaje estadístico

AUTOR: ALBERTO GARCÍA DURÁN
TUTOR: JOSÉ MIGUEL LEIVA MURILLO

8 de septiembre de 2008

Título: *Predicción del coste sanitario en pacientes psiquiátricos mediante técnicas de aprendizaje estadístico*

Autor: *Alberto García Durán*

Tutor *José Miguel Leiva Murillo*

La defensa del presente Proyecto Fin de Carrera se realizó el día 16 de Septiembre de 2008; siendo calificada por el siguiente tribunal:

Presidente: *Marcelino Lázaro Teja*

Secretario: *Ricardo Santiago Mozos*

Vocal *Isaac Seoane Pujol*

Habiendo obtenido la siguiente calificación:

Calificación:

Presidente **Secretario** **Vocal**

Agradecimientos

A mis padres, Mari Tere y Antonio, por haber confiado siempre en mis posibilidades y haber sido tan complacientes conmigo. También le dedico este proyecto a mi hermano Carlos.

A mis amigos Laura y José por haberme demostrado siempre su amistad.

A mi profesor José Miguel por haberme dado la oportunidad de hacer un proyecto que realmente me gustase, por todas las horas que me ha dedicado y por haber tenido siempre un rato para ayudarme.

A Enrique Baca García por habernos proporcionado la base de datos para llevar a cabo esta línea de investigación y habernos despejado las dudas que nos hemos ido encontrando.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Motivaciones del autor

El 11 de Marzo de 1997, el para muchos mejor ajedrecista de la historia, Garry Kasparov, cayó derrotado ante la supercomputadora Deep Blue. El perfeccionamiento de las rutinas de aprendizaje máquina, entre otros aspectos, inclinó la balanza a favor de los chips. Pero esto no es suficiente, el objetivo real no es conseguir que la máquina "venza" al ser humano, sino que se comporte de manera similar a éste. En palabras del Marvin Minsky (experto en IA):

> *"Una vez que contemos con programas que tengan capacidad real de autoaprendizaje, estará garantizado un rápido desarrollo. Ya que la máquina se mejorará a sí misma así como a su modelo, estaremos en posición de observar todos los fenómenos conectados con los conceptos de razón, inteligencia y consciencia. Aunque es difícil decir cuándo se producirá tal desarrollo, no cabe duda de que cambiará el mundo"* [15].

El empeño que muestran los científicos en intentar otorgar a las máquinas la característica posiblemente más distintiva de la inteligencia humana, el aprendizaje, está orientado hacia la consecución de programas que mejoren automáticamente con la experiencia. Un empeño que proporciona sus frutos cada día, pero que está muy lejos del objetivo definido por Marvin Minsky.

Sucesos como éste despertaron en mí un interés inusual que me llevaron a elegir un proyecto fin de carrera relacionado con esta temática, el aprendizaje máquina, que en este proyecto tiene un sentido más de aprendizaje estadístico.

Así que decidí buscar un PFC que me permitiera consolidar y aumentar mis conocimientos en este área, y poder aplicarlos a un caso práctico, un caso útil. Esta oportunidad se me presentó con el proyecto que he tenido la suerte de poder realizar:

Predicción del coste sanitario en pacientes psiquiátricos mediante técnicas de aprendizaje estadístico

Grosso modo, este estudio se basará, con la ayuda de diversas tareas que se pueden hacer con sistemas de aprendizaje; tales como predicción (clasificación y estimación), descripción o segmentación, en el reconocimiento de patrones clínicos en el comportamiento de pacientes de psiquiatría.

Partiendo de una base de datos, de la cual conoceremos múltiples datos tanto de los pacientes como de sus consultas, deberemos seleccionar cuidadosamente aquellos pacientes y sus variables que consideremos relevantes para el problema en cuestión. Este problema y otros tantos que se nos presentarán, serán abordados en capítulos posteriores.

1.2. ¿Qué pretende este proyecto?

El estudio del comportamiento de los eventos clínicos de distintos pacientes psiquiátricos proporcionados por la Fundación Giménez Díaz y el doctor en psiquiatría Enrique Baca García. Dicho estudio está orientado hacia la consecución de un objetivo final: identificar pacientes potencialmente caros, a los que se les es posible proporcionar una mejor atención primaria, permitiendo reducir su coste.

Trataremos de abordar el problema de la siguiente manera:

- Se parte de una base de datos recogidos en un psiquiátrico compuesta de multitud de eventos clínicos de una gran cantidad de pacientes. Cada evento clínico está descrito de manera muy completa, recogiendo información sobre el evento en sí y del paciente en cuestión.
- A partir de esa información deberemos construir un modelo que caracterice el comportamiento clínico de cada paciente cada cierto tiempo. Se deberá estudiar si este modelo debe describir sólo el comportamiento clínico en sí o incluir también datos del paciente, como pueden ser la edad o el sexo.
- Una vez establecido el modelo, deberemos obtener, en base a éste, todos los vectores que caractericen a cada paciente y el coste asociado a cada vector. De este modo, cada paciente estará descrito por un conjunto de vectores con sus costes asociados.
- Aplicar técnicas de clustering para intentar agrupar vectores con patrones similares y que, idealmente, deberán tener costes asociados similares.
- Utilizar un conjunto de entrenamiento/validación para, con la ayuda de técnicas de regresión/clasificación, obtener una máquina de regresión/clasificación para cada cluster.

- Evaluar la calidad de los resultados sobre un conjunto de test. Cada muestra de este conjunto será clasificada en uno de los clusters y le será aplicado la máquina de regresión/clasificación correspondiente.

- En base a criterios médicos, aplicaremos medidas sobre aquellos pacientes catalogados como caros, para darles un tratamiento estándar e intentar minimizar su coste. El hecho de identificar correctamente como caros a los pacientes que en realidad sí lo son es en donde radicará el éxito del proyecto.

El estudió terminará con una serie de conclusiones, en base a los resultados obtenidos con distintos tipos de regresores/clasificadores y el algoritmo de agrupamiento escogido, que determinarán la utilidad o no, de la aplicación de estas técnicas para conseguir el objetivo deseado.

1.3. Vista general del aprendizaje máquina

En esta sección intentaré proporcionar una visión general acerca del aprendizaje máquina, por tanto, omitiré explicaciones detalladas (las cuales podrán ser encontradas en otros capítulos).

¿Qué es el aprendizaje máquina? Es la pregunta que deberá ser contestada antes de adentrarnos en la taxonomía de los métodos de aprendizaje máquina (*Machine Learning, ML*). El ML define al conjunto de algoritmos o técnicas utilizados por sistemas que llevan a cabo tareas asociadas con la inteligencia artificial (*Artificial Intelligence, AI*) y que permite a estos sistemas aprender, es decir, mejorar automáticamente conforme aumenta su experiencia. Estas tareas suelen estar relacionadas con el reconocimiento, diagnóstico, predicción, etc.

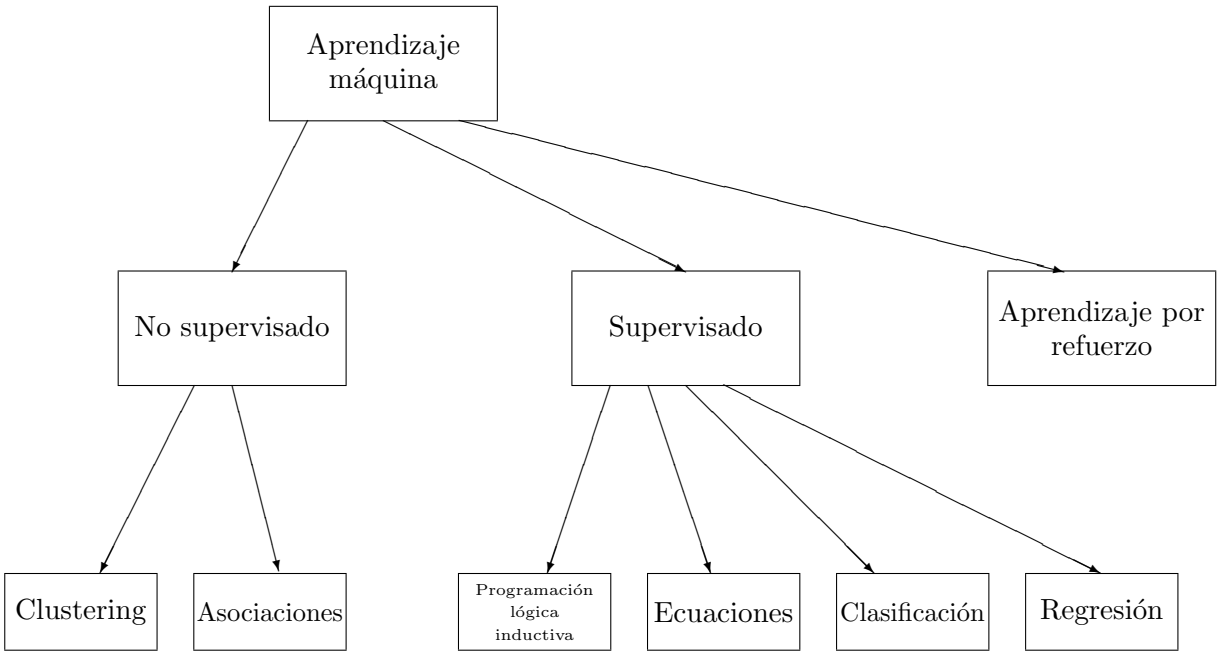
1.3.1. Una taxonomía del aprendizaje máquina

Como disciplina compleja, los distintos algoritmos del ML podrían ser clasificados de muy distinta manera. Aquí voy a presentar un posible árbol de clasificación que divide a las técnicas en 3 grupos principales.

1. Aprendizaje supervisado: aquellas técnicas a las cuales hay que presentar, en primer lugar, un conjunto de entrenamiento que incluye tanto las entradas como las salidas deseadas, permitiendo, de esta manera, aprender una función. La máquina deberá poder ser capaz de generalizar tanto para estos datos de entrenamientos como para ejemplos nuevos [8].

2. Aprendizaje no supervisado: aquí el problema a resolver es encontrar la estructura subyacente de un conjunto de datos [17].

3. Aprendizaje por refuerzo: en sistemas cuyas salidas son una secuencia de acciones. Lo que busca este tipo de aprendizaje es aprender de aquellas *políticas* que generaron secuencias de acciones que consiguieron lograr sus objetivos y evaluar nuevas *políticas* [25].



Taxonomía extraída de [13].

Para una información más detallada acerca de los distintos tipos de aprendizaje máquina consultar [1] y [13].

Introduciré de una manera breve aquellos tipos de técnicas que, posteriormente, utilizaré.

Regresión

En estos problemas tenemos un conjunto de objetos, descritos con algunos atributos. Estos atributos son variables observables independientes. La variable dependiente (salida del predictor) es continua y su valor está determinado por una función de las variables observables independientes. La tarea de los regresores es determinar el valor, que desconocemos, de la variable continua dependiente para un determinado objeto en cuestión.

Los algoritmos de aprendizaje para estos tipos de tareas intentan, por lo tanto, determinar una función que permita relacionar (mapear) el espacio de los atributos con los valores de predicción, a partir de un conjunto de entrenamiento. Está función permitirá posteriormente predecir valores para nuevos ejemplos.

Algunos de los más comunes regresores son: regresor lineal, máquina de vectores soporte o redes neuronales.

1.3.2. Clasificación

El término de clasificar podría cubrir cualquier contexto en el cual una decisión o predicción en base a cierta información disponible es llevada a cabo. El procedimiento de clasificar es un método más formal que permite tomar repetidamente decisiones para nuevos casos. En este proyecto, la tarea de clasificar estará relacionada con la de asignar a cada nueva secuencia de casos una clase de un conjunto de predefinidas. Esto se hará en base a las características observadas en cada nuevo caso.

La construcción de un procedimiento de clasificación a partir de un conjunto de datos para las cuales se conocen las posibles clases es conocido de distintas maneras: *reconocimiento de patrones*, *discriminación* o *aprendizaje supervisado*.

Usada en una gran variedad de contextos como la concesión o no de de créditos bancarios a clientes, en determinar si aplicar un determinado tratamiento a un paciente clínico o en cualquier mecanismo automático de reconocimiento de formas, colores, palabras...

Clustering

Los métodos de clustering son de los más populares de los métodos de aprendizaje no supervisado. La tarea de éstos es determinar subconjuntos coherentes dentro de la totalidad del conjunto de muestras para el aprendizaje. Lógicamente, estos subconjuntos deberán agrupar muestras similares.

El número de clusters puede ser o bien conocido o bien determinado por el algoritmo. Otro aspecto importante, que determinará el éxito o el fracaso del clustering será la elección de la medida de similitud entre las muestras.

Algunos de los más comunes algoritmos de clustering: k-means, mapas autoorganizados de Kohonen o ISODATA.

1.3.3. Otros aspectos

Vectores de entrada y Salidas

Los primeros están formados por componentes. Los valores de éstos pueden ser de 3 tipos (números reales, números discretos o valores categóricos), en nuestro caso serán de

tipo real.

Las salidas dependerán del tipo de tarea que estemos llevando a cabo. En el caso de clustering las salidas serán categorías o etiquetas. Cuando estemos utilizando regresión las salidas serán números reales que llamaremos estimaciones.

Ruido

Normalmente los vectores de entrada estarán afectados por ruido. Obviamente, una mayor cantidad de ruido en los datos hará más difícil que tengamos éxito a la hora de conseguir nuestros objetivos, ya que genera un mayor impacto en la interpretación de los datos, y en los modelos y decisiones tomados en base a éstos.

Hay dos tipos de ruido:

1. Ruido de clase: altera de manera aleatoria el valor de salida.
2. Ruido de atributos: altera de manera aleatoria el valor de los componentes del vector de entrada.

Régimen de entrenamiento

Según la manera en la que los datos de entrenamiento son presentados a la máquina:

1. Batch: todos los datos son dados a la máquina al principio del aprendizaje.
2. On-line: se les presentan los datos de uno en uno. Para cada entrada, da una estimación de la salida antes de conocer la salida deseada. Con cada nuevo ejemplo la máquina se irá actualizando.

Y según la interpretación de la salida obtenida:

1. Modelo generativo: obtiene el modelo que determina la probabilidad de la entrada conociendo la salida, lo que comúnmente se conoce como $p(\boldsymbol{x}|y)$ $(y \in -1, +1, ...)$.
2. Modelo discriminativo: el objetivo no es modelar a los datos analíticamente, sino tener un modelo capaz de etiquetar o clasificar a cada muestra en categoría/s.

Todos estos aspectos se encuentran detallados en [17].

Evaluación de los resultados

Es importante tener métodos para poder evaluar la calidad del aprendizaje. En los métodos supervisados, como regresión, se suele comprobar sobre un conjunto de validación. Posteriormente, los resultados finales son evaluados sobre un conjunto de test que, normalmente, está formado aproximadamente por el 20 por ciento del total del conjunto de muestras, siendo el restante 80 por ciento (conjunto de entrenamiento/validación) parte de la etapa de aprendizaje. Ver Figura 1.1.

Figura 1.1: Conjuntos de aprendizaje y test

1.3.4. Aplicaciones del aprendizaje máquina: diagnóstico médico

En la medicina actual, los cimientos del éxito en los tratamientos son unos diagnósticos correctos. Los diagnósticos son establecidos en función de los signos, síntomas y resultados de pruebas sobre los pacientes. Un problema similar a los diagnósticos son los pronósticos, los cuales determinan, en base a probabilidades, la historia natural de una enfermedad [13].

Basados en los historiales de los pacientes que han sido tratados en un mismo psiquiátrico (en nuestro caso), los métodos de aprendizaje estadístico pueden ser usados para diagnosticar (o pronosticar) nuevos pacientes.

1.4. Ciclo de vida del proyecto

Gráficamente, este proceso se podría ver como el siguiente ciclo (extraído de [13]):

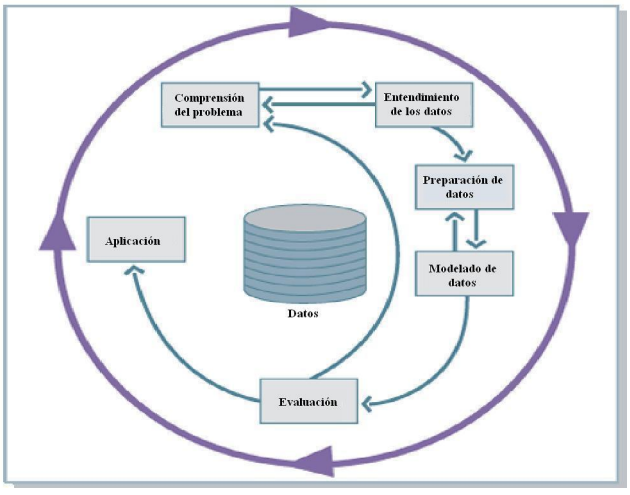


Figura 1.2: Ciclo de vida del proyecto

Como vemos, se trata de un proceso iterativo, en el cual las distintas fases son cíclicamente repetidas hasta conseguir el objetivo deseado. Cada iteración nos ha permitido llegar a conclusiones como qué variables son relevantes para describir el comportamiento de los pacientes, u otras referidas a qué técnicas utilizar para obtener mejores resultados, ejemplos de estas últimas han sido la toma de decisión de utilizar mecanismos de regresión/clasificación no lineales o de la utilización de un algoritmo de agrupamiento.

Las dos primeras etapas, consistentes en la comprensión del problema y de los datos, son las que serán abordadas en el siguiente capítulo.

1.5. Sumario

En el campo de la salud mental, la declaración explícita de valores y principios señalaría nuestras prioridades, aquellas cuestiones que deben orientar nuestro comportamiento y nuestros esfuerzos económicos, científicos y normativos.

Según [5], se estima que unos 450 millones de personas en el mundo padecen un trastorno mental o de comportamiento en un momento dado de su vida. En España no existen datos suficientes para valorar el coste económico exacto que las enfermedades mentales suponen, pero se estima que estará alrededor del 3 y el 4 % del PNB, siendo la causa más frecuente de enfermedad en Europa, por delante de las enfermedades cardiovasculares y del cáncer. Además, más de la mitad de las personas que necesitan tratamiento no lo reciben y, de las que están en tratamiento, un porcentaje significativo no recibe el adecuado.

Para promocionar la salud mental se puede actuar sobre la persona o sobre la población. A nivel individual, reforzando la resiliencia con intervenciones que incrementan la autoestima y dotan de destrezas para afrontar el estrés. A nivel de población, con intervenciones para incrementar el capital social, promover conductas sanas de crianza, mejorar la seguridad, reducir el estrés en las escuelas y en los lugares de trabajo.

Por otra parte, el gasto sanitario ha crecido de forma importante en las últimas décadas. Con el fin de controlarlo, las Administraciones sanitarias adoptaron criterios de gestión empresarial y de mercado, con el riesgo de anteponer la economía a cualquier otra consideración. Ante este hecho, ciertos documentos han alertado de tal situación y han propuesto un modelo sanitario que gire, ante todo, alrededor de teorías éticas de justicia distributiva y de la valoración de la salud.

Es de este interés, del de poder proporcionar un servicio psiquiátrico lo más adecuado posible para cada paciente y no realizar gastos innecesarios, de donde nace este proyecto, puesto que el objetivo principal es adelantarnos al desarrollo natural del comportamiento clínico del paciente para, de este modo, poder dedicarle la atención que requiera. Como es natural, unos pronósticos adecuados y una atención temprana traerán como consecuencia

el ahorro de gastos innecesarios en tratamientos que pueden ser reemplazados por otros que se adecuan de mejor manera a las necesidades de cada paciente.
Obviamente, no es tarea fácil establecer un modelo general de ahorro para todos los pacientes, puesto que las distintas patologías que sufren los pacientes no pueden ser tratadas de la misma manera. Sin embargo, y con la ayuda del doctor en psiquiatría Enrique Baca García, hemos conseguido una primera aproximación con sus simplificaciones que puede ser de utilidad como base para futuras líneas de investigación. Las mencionadas simplificaciones son explicadas y razonadas en el Capítulo 6.

Para una mayor información acerca de la estrategia seguida en la salud mental consultar en [5].

Como se ha comentado anteriormente, para tal objetivo nos hemos aprovechado de técnicas de aprendizaje estadístico. El proceso esquemático de las distintas etapas que conforman este estudio se puede apreciar en la Figura 1.3.

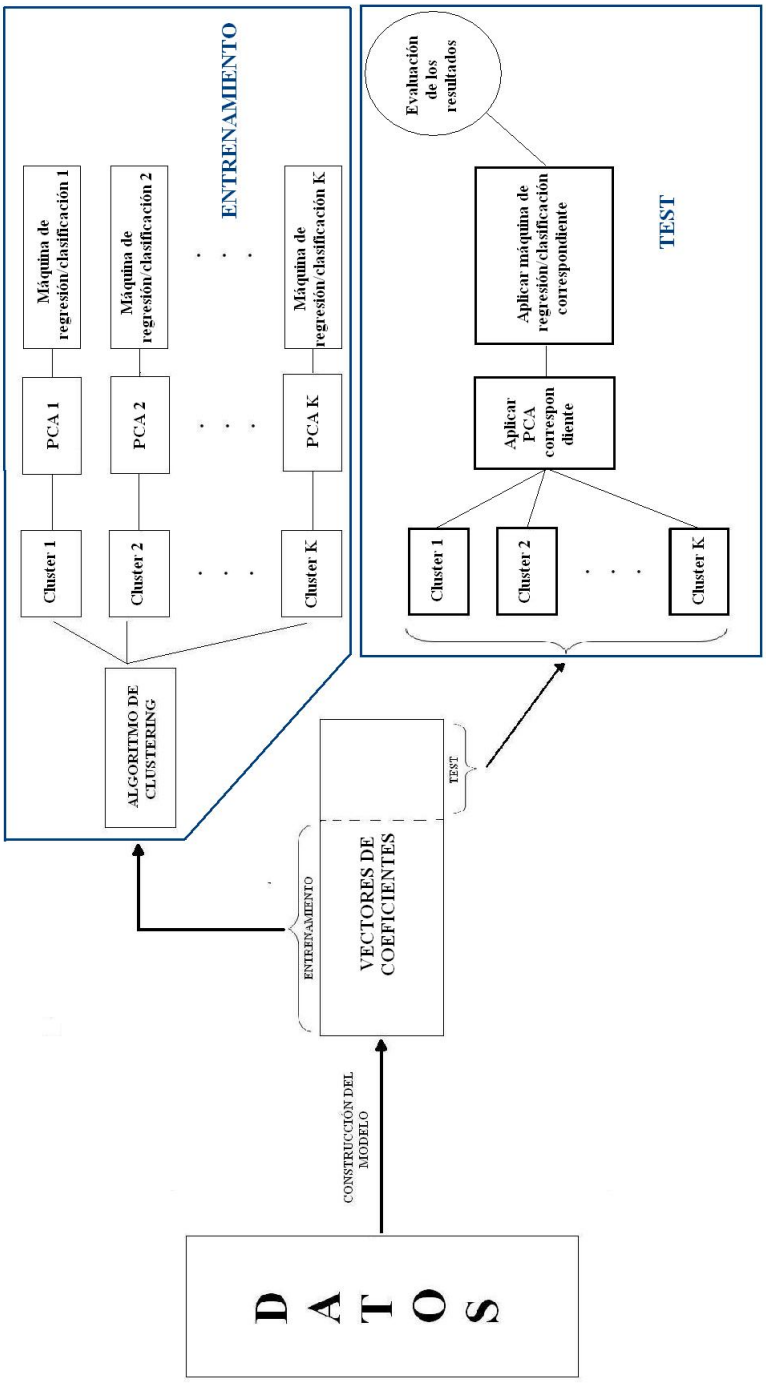


Figura 1.3: Esquema general del proyecto

Siguiendo el esquema de la Figura 1.3, la memoria del proyecto está estructurada de una manera cronológica a las diferentes etapas seguidas. De esta manera en:

1. El Capítulo 2 se presenta el modelo que caracteriza al comportamiento dinámico de los pacientes.

2. El Capítulo 3 se estudiarán el funcionamiento de las técnicas de clustering, y se presentarán dos que permitirán, por un lado, agrupar pacientes con comportamiento similares, y, por el otro obtener, una visualización en dos dimensiones de los coeficientes del modelo empleado.

3. El Capítulo 4 se introducirá la técnica de análisis de componentes principales, comentando el porqué esta herramienta es útil en el proyecto.

4. El Capítulo 5 se presentarán diferentes algoritmos de regresión y clasificación.

5. El Capítulo 6 se valorarán los resultados obtenidos con las diferentes estrategias y se comentarán otros puntos que pueden resultar de interés para el lector como una descripción más menuda de cómo se ha conseguido un funcionamiento lo más óptimo posible de las técnicas empleadas o del método de validación usado, entre otras.

6. El Capítulo 7 se resumen las principales conclusiones extraídas de los resultados mostrados en el Capítulo 6 y se propondrán futuras líneas de investigación.

x

Capítulo 2

Construcción del modelo

Esta primera parte constituye un punto crítico, ya que una mala selección de variables o un modelo que no contenga toda la información relevante para conseguir el objetivo final puede condicionar el problema, de modo que los esfuerzos realizados a partir de aquí puedan resultar inútiles.

Varias son las cuestiones que se deben resolver: de todos los datos que disponemos ¿cuáles utilizar?, ¿qué modelo representará mejor el comportamiento clínico de los pacientes? o si es necesario algún tipo de preprocesado para eliminar de nuestro modelo vector de entrada $\mapsto$ salida aquellas muestras que puedan resultar perjudiciales para conseguir un sistema que generalice de manera óptima.

Todo esto será respondido sin perder de vista el objetivo fundamental de esta primera parte. Relacionar un vector de entrada, determinado por el modelo establecido, con un valor o etiqueta de salida.

2.1. ¿Qué variables seleccionar?

Para conseguir unas aproximaciones óptimas se deben elegir cuidadosamente las variables a emplear. En el modelo se deben incluir variables predictoras que cumplan dos requisitos. Primero, que realmente sirvan para predecir la variable de salida y segundo, que no covaríen entre sí. En el caso contrario, en el que se añaden variables irrelevantes o con dependencia entre ellas puede provocar un sobreajuste en el modelo [30].

Un procedimiento sencillo, pero a la vez útil, para utilizar solamente aquellas variables, de todas las disponibles, más relevantes, consistiría en hacer un barrido de experimentos con distintos número y distintas combinaciones de variables. Obviamente, cuando el número de variables es grande el número de posibles combinaciones de modelos de distintos

números de variables y combinaciones de ellas aumenta exponencialmente, lo que hace que este barrido de distintas posibilidades sea muy costoso y por tanto, inviable.

Como hemos dicho, el hecho de incluir en el modelo variables dependientes entre sí o irrelevantes puede provocar un sobreajuste innecesario en el modelo. Otro tipo de fenómeno que queremos evitar son los denominados modelos subajustados, ya que en ambos casos la generalización del modelo no será buena, y de esta manera, las salidas predichas para nuevos ejemplos dará malos resultados.

- Modelos sobreajustados: Establecemos la hipótesis del modelo de regresión/clasificación $g(\cdot)$ con un conjunto de parámetros. Si está hipótesis es más compleja que la subyacente a la función real de los datos estaremos sobreajustando. Un ejemplo de esto sería asumir para un problema de comportamiento linear una $g(\cdot)$ cuadrática. [2]

- Modelos subajustados: En este caso si la hipótesis establecida es menos compleja que la subyacente a la función real de los datos estaremos subajustando. Un ejemplo de esto sería asumir para un problema de comportamiento cuadrático una $g(\cdot)$ linear. [2]

Cada entrada del historial de eventos clínicos contiene la siguiente información: sexo, edad, estado conyugal, nivel de estudios, situación laboral, sector en el que trabaja o estado civil.

De acuerdo a la caracterización de los pacientes que encontraremos en la sección 2.2, la única información que utilizaremos será la referida al tiempo que transcurre entre los distintos eventos clínicos.

2.2. Caracterización pacientes

Para describir el comportamiento clínico de los pacientes, donde hay 3 tipos de asistencia sanitaria (hospital, ambulatorio, urgencias), hemos supuesto un modelo de estados como muestra la Figura 2.1.

Como vemos en la Figura 2.1, necesitaremos un modelo de probabilidad temporal para caracterizar el comportamiento de un paciente durante un intervalo de 365 días, que estará determinado por un conjunto de transiciones entre los distintos estados (hospital, ambulatorio y urgencias). Para ello, se ha usado un modelo exponencial para cada transición, en concreto dispondremos de 9 modelos exponenciales por paciente y año. Simplificándolo, caracterizaremos a cada transición por un parámetro λ. Este parámetro es el que caracteriza al modelo exponencial:

$$f(t) = \lambda \exp(-\lambda t) \tag{2.1}$$

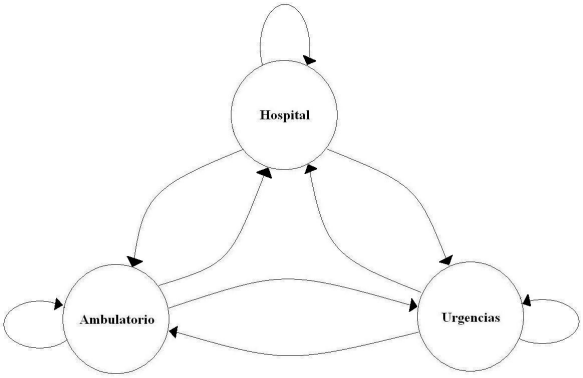


Figura 2.1: Modelo de un paciente

Modelo de Markov

Son sistemas en los que, en cualquier momento, nos encontramos en uno de los N estados que conforman el sistema: $S_1, S_2, ..., S_N$; y que se cumple el hecho de que la probabilidad de estar en un instante t_{n+1} en un determinado estado depende exclusivamente del estado en el que se encontraba en el instante t_n.

La definición formal de una cadena continua de Markov $X(t)$ es:

$$\begin{aligned} &Prob\big[X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, ..., X(t_1) = x_1\big] = \\ &= Prob\big[X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\big] \end{aligned} \tag{2.2}$$

Esto corresponde a decir que, dado el estado actual, el futuro es independiente del pasado. El equivalente matemático al dicho, "Hoy es el primer día del resto de tu vida".

Las cadenas de Markov están caracterizadas por diagramas de estados que describen los estados y las transiciones entre ellos permitidas. En el caso de una cadena de tiempo continuo, las transiciones pueden ocurrir en cualquier instante y están caracterizadas por medias de distribuciones exponenciales. Si por el contrario la cadena es de tiempo discreto, las transiciones solo pueden ocurrir en instantes dados y con probabilidades según distribuciones geométricas.

Para más información de los Modelos de Markov consultar [6], [1] y [24].

La Figura 2.1 representa un modelo de Markov de tiempo continuo con tres estados, en el cual, la probabilidad de que el paciente se mueva de un estado a otro pasado un tiempo t está determinada por una probabilidad obtenida mediante una función de densidad exponencial caracterizada por su parámetro λ_i.

Cada estado tendrá un coste asociado según se recoge en la siguiente tabla:

Cuadro 2.1: Gasto asociado a cada estado

	Ambulatorio	Urgencias	Hospital
Gasto	33.44€	73.763€	148.144€/día

Por tanto, el empleo de un modelo de Markov de tiempo continuo cubrirá la necesidad de tener un modelo de probabilidad temporal que caracterice el comportamiento de un paciente durante un intervalo de 365 días.

Nuestro objetivo ahora es calcular los parámetros (λ) que caracterizan a las transiciones que se producen en cada paciente cada año.

2.2.1. Estimación mediante Maximum likelihood

Tenemos un conjunto de muestras χ independientes e idénticamente distribuidas. Asumimos que estas muestras $\chi = \{x^t\}_{t=1}^N$ son muestras de alguna familia de densidad de probabilidad conocida $p(x|\theta)$, caracterizada por sus parámetros θ.

Queremos encontrar los parámetros θ que hacen que esa función de densidad de probabilidad se ajuste lo mejor posible al conjunto de muestras $\{x^t\}_{t=1}^N$. Como estas muestras son independientes, la probabilidad total es el producto de las probabilidades de las muestras individuales.

$$l(\theta|\chi) \equiv p(x^t|\theta) = \prod_{t=1}^{N} p(x^t|\theta) \tag{2.3}$$

Como hemos dicho, estamos interesados en encontrar los θ que maximiza la posibilidad de que ocurra χ. Los θ que consiguen esto serán denotados como θ_{ml}.

Otra manera de resolverlo, sobre todo en el caso de que la función de densidad contenga exponenciales, ya que simplifica los cálculos, sería aplicando logaritmos.

$$L(\theta|\chi) \equiv \log l(\theta|\chi) = \sum_{t=1}^{N} \log p(x^t|\theta) \tag{2.4}$$

Para hallar el θ que maximiza la expresión 2.3 o la 2.4 resolvemos:

$$\frac{\partial l(\theta|\chi)}{\partial \theta} = 0 \qquad \text{ó} \qquad \frac{\partial L(\theta|\chi)}{\partial \theta} = 0 \tag{2.5}$$

Y comprobamos si el θ estimado ($\hat{\theta}$) es el θ_{ml}, si cumple la siguiente condición:

$$\frac{\partial^2 l(\theta|\chi)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0 \tag{2.6}$$

Desarrollo extraído de [1].

2.2.2. Evaluación del estimador

Definiendo el estimador de θ como $d = d(\chi)$, evaluaremos su calidad mediante el cálculo de cómo de parecido o diferente es de θ, esto es, $(d(\chi) - \theta)^2$. Para variables aleatorias necesitamos promediar este valor. Obtendremos este valor mediante el error cuadrático medio $r(d, \theta)$ del estimador d definido como:

$$r(d, \theta) = E[(d(\chi) - \theta)^2] \tag{2.7}$$

Desarrollando la expresión:

$$\begin{aligned} r(d,\theta) &= E[(d - E[d] + E[d] - \theta)^2] \\ &= E[(d - E[d])^2 + (E[d] - \theta)^2 + 2(E[d] - \theta)(d - E[d])] \\ &= E[(d - E[d])^2] + E[(E[d] - \theta)^2] + 2E[(E[d] - \theta)(d - E[d])] \\ &= E[(d - E[d])^2] + (E[d] - \theta)^2 + 2(E[d] - \theta)E[d - E[d]]] \\ &= \underbrace{E[(d - E[d])^2]}_{Varianza} + \underbrace{(E[d] - \theta)^2}_{Sesgo^2} \end{aligned} \tag{2.8}$$

El primer término, como vemos, es la varianza y el segundo el sesgo al cuadrado. Nos gustaría que el estimador fuera:

- Insesgado; esto es, que su sesgo sea nulo. Definiendo el sesgo como $R(d, \theta) = E[(d(\chi) - \theta)^2]$, se puede observar que éste será nulo cuando la esperanza del estimador es igual al propio valor estimado. La ausencia de sesgo nos dice que si promediamos el valor numérico del estimador obtenido en un conjunto amplio de muestras , dicho promedio tenderá a coincidir con el verdadero valor del parámetro estimado.
- Consistente; cuando al aumentar el número de muestras $N \to \infty$, la varianza del estimador tiende a 0 $Var \to 0$. Esto significa que se aproxima al valor real del parámetro a medida que aumenta el número de muestras.

Se puede apreciar que si dispusiésemos de un estimador insesgado y de infinitas muestras, el error cuadrático medio sería 0.

2.2.3. Significado de las λ

El hecho de abordar el problema de esta manera, exige que determinemos una familia de densidad para las transiciones entre los distintos estados.

El modelo está caracterizado por una serie de probabilidades que dependen del tiempo. Estas probabilidades determinan cuán probable es, dado que mi último evento recogido pertenece al estado $X \in \{Hospital, Ambulatorio, Urgencias\}$, que pasado un tiempo t (en días), el paciente acuda a cada uno de los 3 estados definidos en 2.2.

Es razonable pensar que una función de densidad exponencial se ajusta de una manera adecuada al modelo que se acaba de explicar.

Para ilustrar esta decisión se van a mostrar dos casos concretos que permitirán entender mejor el modelo:

- En el caso concreto de que un determinado paciente haya "repetido" en numeras ocasiones durante un año la transición Ambulatorio → Hospital, podríamos interpretarlo como que la probabilidad de que pasado intervalos de tiempo t (en días) muy pequeños se repita esta transición es muy alta, y conforme pasa el tiempo esta probabilidad va bajando. Gráficamente se puede observar dicho fenómeno en la Figura 2.2 (línea azul).

- Si por el contrario, en un año concreto un determinado paciente ha "repetido" en escasas ocasiones esta misma transición, podríamos interpretarlo como que su comportamiento en dicha transición no sigue ninguna "pauta" y, por tanto, la probabilidad de que pasado un tiempo t (en días) se produzca esta transición suele ser algo más homogénea a lo largo del tiempo t. Este comportamiento se puede observar en la Figura 2.2 (línea roja).

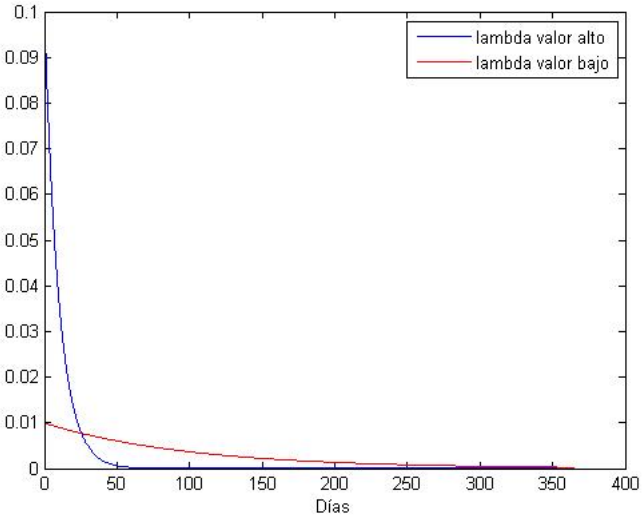


Figura 2.2: Ejemplos de valores de λ

También hay que destacar el caso en el que no haya pares de eventos clínicos referidos a una determinada transición, en cuyo caso el valor de la λ correspondiente será 0.

2.2.4. Estimación de las λ

Aplicaremos el algoritmo Maximum Likelihood descrito en 2.2.1 para poder estimar los valores de las λ, que corresponden a la medias de funciones de densidad exponenciales y que permiten describir a éstas de una manera completa.

$$f_i(t) = \lambda_i \exp(-\lambda_i t) \tag{2.9}$$

siendo i una de las 9 posibles transiciones.

Teniendo el conjunto de tiempos de las transiciones entre los distintos estados (disponible en la base de datos), tenemos la información necesaria para estimar las 9 λ_i.

$$l(\lambda|t) = \prod_{i=1}^{K} \lambda \exp(-\lambda t_i) \tag{2.10}$$

$$= \lambda^K \exp -(\lambda \sum_{i=1}^{K} t_i) \tag{2.11}$$

Siendo K el número total de tiempos correspondiente a una determinado transición y que hemos supuesto que están modelados por una función de densidad exponencial.

Aplicando logaritmos neperianos:

$$L(\lambda|t) = \ln(\lambda)^K + \ln(\exp(-\lambda \sum_{i=1}^{K} t_i)) \tag{2.12}$$

$$= K \ln(\lambda) - \lambda \sum_{i=1}^{K} t_i \tag{2.13}$$

Y por último hallamos la expresión del estimador de máxima verosimilitud de λ, λ_{ml}:

$$\frac{\partial L(\lambda|t)}{\partial \lambda} = \frac{K}{\lambda} - \sum_{i=1}^{K} t_i \tag{2.14}$$

$$0 = \frac{K}{\lambda} - \sum_{i=1}^{K} t_i \tag{2.15}$$

$$\lambda_{ml} = \frac{K}{\sum_{i=1}^{K} t_i} \tag{2.16}$$

Que como se aprecia es igual a la media aritmética invertida. Algunos autores escriben la función de densidad exponencial de la siguiente manera:

$$f(t) = \frac{1}{\lambda}\exp\left(\frac{t}{\lambda}\right) \tag{2.17}$$

"Provocando" que el parámetro λ a estimar mediante ML sea igual a la media aritmética.

Evaluación del estimador λ_{ml}

Como se vio en 2.2.2, para evaluar la calidad del estimador se deberá calcular el sesgo y la varianza de éste.

$$E[\lambda_{ml}] = E[\frac{K}{\sum_{i=1}^{K} t_i}] = \frac{K}{K * E[t]} = \frac{K}{K * \frac{1}{\lambda}} = \lambda \tag{2.18}$$

$$Var(\lambda_{ml}) = Var(\frac{K}{\sum_{i=1}^{K} t_i}) = \frac{K}{K^2 * Var(t)} = \frac{K}{K^2 * \frac{1}{\lambda^2}} = \frac{\lambda^2}{K} \tag{2.19}$$

Según 2.18 y 2.19, λ_{ml} se trata de un estimador insesgado y consistente.

Ejemplo

Analicemos el caso concreto del comportamiento clínico de un paciente durante un año:

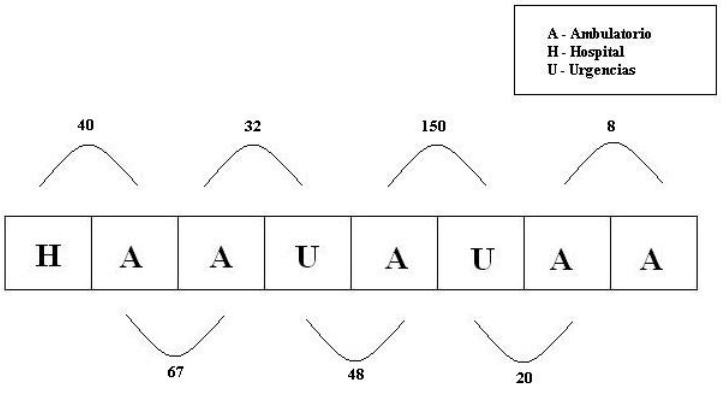


Figura 2.3: Ejemplo de las transiciones de un determinado paciente durante un año

Según lo observado en 2.3, el valor de las λ será:

$$\lambda_{A\mapsto A} = \frac{2}{67+8} = 0{,}02\widehat{6} \tag{2.20}$$

$$\lambda_{H\mapsto A} = \frac{1}{40} = 0{,}025 \tag{2.21}$$

$$\lambda_{U\mapsto A} = \frac{2}{48+20} = 0{,}0294 \tag{2.22}$$

$$\lambda_{A\mapsto U} = \frac{2}{32+150} = 0{,}0109 \tag{2.23}$$

$$\tag{2.24}$$

$$\lambda_{A\mapsto H} = \lambda_{H\mapsto U} = \lambda_{H\mapsto H} = \lambda_{U\mapsto U} = \lambda_{U\mapsto H} = 0 \tag{2.25}$$

Transición Hospital-Hospital

Esta transición tiene una pequeña particularidad. Un paciente podría permanecer varios días seguidos en el hospital, por lo que habría que tener en cuenta este aspecto.

Debido a la incapacidad de este modelo para representar la situación de permanecer más de un día en el hospital, es necesario encontrar alguna manera alternativa para codificar el valor de la λ correspondiente a esta transición, de modo que la función de densidad exponencial que representa esta situación sea capaz de reflejar, de cierta manera, tanto la frecuencia con que se produce esta transición como la permanencia de varios días en el hospital.

Analizaremos dos casos extremos para establecer un criterio para la codificación de esta λ.

- Si un paciente repite en numerosas ocasiones esta transición o no, pero en las ocasiones que ha acudido al hospital ha permanecido durante largos periodos ingresado, se considerará que el comportamiento clínico será similar (puesto que acarreará un coste parecido). En estos casos, $\lambda_{H\mapsto H}$ tendrá un valor alto y la función de densidad exponencial tendrá un aspecto parecido al de la Figura 2.2, línea azul.

- En el caso contrario, el gasto generado será mucho menor. Por tanto, si se repite esta transición poco o cuando acude al hospital su estancia es de corta duración, el valor de la λ deberá ser pequeño y el aspecto de la función de densidad exponencial que caracteriza esta transición será similar a la de la Figura 2.2, línea roja.

Ambas interpretaciones se conseguirán con:

$$\lambda_{H \mapsto H} = \frac{\text{Número de días que el paciente ha permanecido en el hospital}}{\text{365 días}} \tag{2.26}$$

2.3. Regresor

Un paciente podrá estar definido por uno o más vectores de parámetros ($\boldsymbol{\lambda}$), ya que estos vectores definen el comportamiento clínico de dicho paciente durante un periodo de 365 días y, por tanto, el número de vectores de un paciente estará determinado por el tiempo durante el cual se han recogido sus eventos clínicos en el historial del psiquiátrico.

Cada $\boldsymbol{\lambda}$ tendrá un coste asociado que corresponderá al coste generado por ese paciente en los 365 días posteriores. Obviamente esto tiene sentido, ya que nuestro objetivo es predecir el coste que va a generar ese paciente en los próximos 365 días en base a lo observado durante este año.

Es lógico pensar que dicha relación existe, puesto que un comportamiento clínico "grave" hace presagiar que el próximo año tendrá unos costes intermedios/altos, o un comportamiento sin muchos "incidentes" dará lugar a unos costes bajos/intermedios al año siguiente. Obviamente, el ser humano no sigue un comportamiento siempre previsible y puede que, en ocasiones, el comportamiento clínico de un año esté poco o nada relacionado con los costes generados en el siguiente periodo, pero es necesario asumir esta hipótesis.

Gráficamente sería:

$$\text{Paciente i-ésimo} \begin{cases} C_2 = g(\boldsymbol{\lambda}_1) \\ C_3 = g(\boldsymbol{\lambda}_2) \\ . \\ . \\ C_{j-1} = g(\boldsymbol{\lambda}_{j-2}) \\ C_j = g(\boldsymbol{\lambda}_{j-1}) \end{cases}$$

Siendo j el número de vectores de λ que tiene el paciente i-ésimo.

En el Capítulo 5 se abordarán técnicas de regresión que permitan obtener estas funciones.

2.4. Clasificación

En este caso, la diferencia con el regresor es que la salida no es el coste generado durante ese paciente en los 365 días posteriores, sino que la salida será una etiqueta que

indicará que ese coste ha sido considerado como caro o como barato.

Posibles etiquetas:

- +1: Paciente caro.
- -1: Paciente barato.

En el Capítulo 6 se explica en base a qué un paciente es considerado como caro o como barato.

2.5. Pre-procesado de la base de datos

Consiste en la preparación previa de los datos para ser usados en la construcción, entrenamiento y prueba de un modelo.

En nuestro caso, querremos que nuestro regresor/clasificador generalice de manera óptima ante cualquiera de las entradas posibles. El mayor problema que nos podemos encontrar en esta tarea es disponer de un "vector tipo" que predomina de manera muy destacable en la base de datos.

Tener muchos vectores iguales puede condicionar el problema, ya que la existencia de muchos casos idénticos puede provocar la estimación de un modelo de regresión/clasificación que generalice mal y obtener malos resultados para aquellas entradas que difieran de la predominante. Es por este motivo que es necesario establecer algún criterio para reducir en parte este problema.

No queremos eliminar esta predominancia, sino hacerla más "suave", de modo que siga siendo el vector tipo más existente en la base de datos, pero que no condicione demasiado el regresor/clasificador y permita generalizar de una manera más o menos óptima para cualquier tipo de entrada.

En este problema, se aprecia muchos vectores nulos. Esto significa que hay muchos pacientes con muy pocas entradas clínicas repartidas a lo largo de mucho tiempo, por lo que hay muchos años sin eventos clínicos y, por tanto, muchos vectores nulos. Hemos conseguido reducir la cantidad de estos vectores eliminando de la base de datos aquellos pacientes que no cumplen la siguiente condición:

$$\frac{\text{Número de sucesos clínicos del paciente}}{\text{Intervalo de tiempo en el que están registrados estos eventos clínicos}} > 0{,}01 \qquad (2.27)$$

Esto es equivalente a decir que solo consideraremos relevantes a aquellos pacientes que han tenido, en media, un evento clínico recogido cada 100 días.

Mencionar que previamente fueron eliminados aquellos pacientes cuyo historial recogía sucesos durante menos de 720 días (2 años), puesto que se necesitan por lo menos 720 días para poder entrenar el modelo $\boldsymbol{\lambda}_i \mapsto C_{i+1}$ para el caso de regresión o el modelo $\boldsymbol{\lambda}_i \mapsto Etiqueta \in \{+1, -1\}$ para clasificación..

Capítulo 3

Técnicas de clustering

En este capítulo se tratará de dar una visión general de los algoritmos de clustering o agrupamiento: el objetivo de éstos, una posible taxonomía de estas técnicas, las etapas que lo conforman... Para finalmente presentar dos algoritmos concretos, el k-means y los mapas auto-organizados.

3.1. Introducción

Clustering se refiere a la tarea de dividir o particionar datos no etiquetados en grupos o clusters significativos. Algo que resulta ser una útil aproximación en procesos de minería de datos para identificar patrones ocultos o revelar información subyacente en una larga colección de datos. Las áreas de aplicación de las técnicas de clustering incluye segmentación de imágenes, recuperación de información o clasificación de documentos.

Como se ha dicho, el principal objetivo de estas técnicas es particionar los datos en subconjuntos significativos. Matemáticamente hablando, una partición es una colección de subconjuntos disjuntos, esto es, no interseccionan entre ellos, y cuya unión es el conjunto entero. Si el algoritmo de clustering cumple esta condición hablaremos de *grisp* clustering. En el caso contrario, en el que los distintos clusters no sean disjuntos, hablaremos de *fuzzy* clustering.

El modelo obtenido para la detección de clusters puede ser sensible a muestras que se alejan mucho del "comportamiento" o tendencia general de los datos, lo que comúnmente se llaman *outliers*. Si el algoritmo es insensible a estos datos, se tratará de un método robusto.

Para una visión más exhaustiva de los algoritmos de clustering consultar [29] y [13].

3.1.1. Taxonomía de los algoritmos de clustering

Estos algoritmos pueden ser clasificados de diferentes maneras:

- ***Jerárquica o particional***. En los primeros inicialmente cada muestra se considera un cluster y, tras unas pocas iteraciones de unión de clusters "cercanos", obtenemos el conjunto de clusters finales. Mientras tanto, en los segundos el proceso es inverso; el conjunto de muestras de la base de datos es inicialmente un solo *cluster* y en cada iteración se va dividiendo en otros más pequeños que engloban muestras parecidas según la medida de similitud utilizada.

- ***Exclusivo o superpuesto***. La diferencia radica en que en los algoritmos exclusivos de clustering una muestra pertenece exclusivamente a un solo cluster, mientras que en los segundos una muestra puede pertenecer a varios, eso sí, cada uno con un cierto grado de "pertenencia".

- ***Determinista o estocástico***. Esta cuestión es la más relevante en aproximaciones particionales, y está diseñada para optimizar la función de error cuadrática. Esta optimización puede lograrse usando técnicas tradicionales o a través de una búsqueda aleatoria del espacio de estados componiendo todas las posibles asignaciones de etiquetas.

- ***Incremental o no incremental***. Cuando el tamaño del conjunto de muestras a las que se quiere aplicar este tipo de técnica afecta a la arquitectura del algoritmo o no.

Taxonomía extraída de [29].

3.1.2. Una vista general

Un proceso de clustering puede ser visto como la realización de las etapas mostradas en la Figura 3.1 [29]:

1. Representación de patrones, opcionalmente puede incluir extracción y/o selección de características,

2. Definición de una medida de proximidad entre patrones y

3. Clustering o agrupamiento de los datos de acuerdo al patrón de representación y medidas de proximidad elegidos.

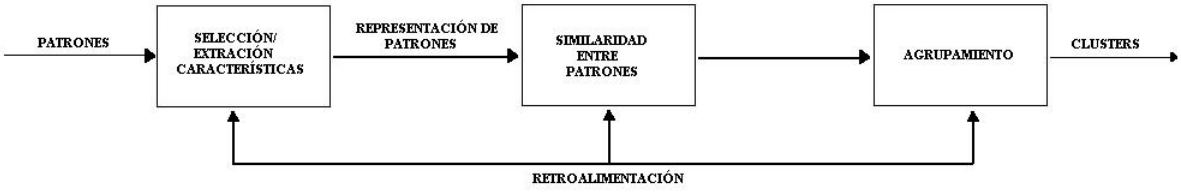


Figura 3.1: Una típica secuencia de una actividad de clustering

Representación de patrones, selección y extracción de características

Referida a la observación y abstracción del problema de aprendizaje, incluyendo el tipo y número de características, el número de patrones o el formato de representación de las características.

La selección de características es definida como la tarea de identificar un conjunto de variables obtenidas directamente o mediante transformación de las originales que resulten ser las más representativas posibles del conjunto disponible originalmente, para ser usado por la maquina. La extracción de características tiene como tarea transformar las variables o características seleccionadas en un "formato" que permita ser entendido por el algoritmo y trabajar con él.

La representación de patrones es considerada como la base del aprendizaje máquina, y la elección de un sistema de coordenadas adecuado puede llevar a la obtención de mejores resultados. Como se muestra en la Figura 3.2(A), el uso de un sistema de coordenadas cartesianas permitirá al algoritmo de clustering distinguir los 4 clusters distintos que hay. En cambio en la Figura 3.2(B), para otro conjunto de datos en el sistema cartesiano, el algoritmo de clustering no diferenciará los 4 subconjuntos que se pueden apreciar, ya que no son fácilmente separables en términos de distancia Euclídea. En su lugar, la utilización de un sistema de coordenadas polares probablemente nos guiase a un mejor resultado, ya que los datos en cada cluster están cercanos el uno respecto al otro en términos de ángulo polar.

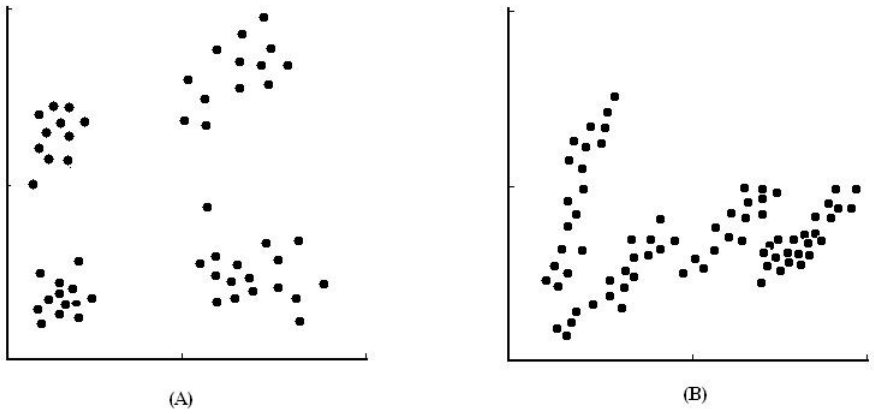


Figura 3.2: Representación en sistema de coordenadas de dos conjuntos de datos distintos

En este proyecto, la selección y extracción de características, así como la elección del sistema de coordenadas ha sido elaborada en una etapa anterior, concretamente en el Capítulo 2.

Medida de proximidad de patrones

Referidas a las métricas que se encargan de evaluar la similitud entre patrones, o lo que es lo mismo, indican como se parecen los unos a los otros.

Centrándonos en el ejemplo de la Figura 3.2(A), la distancia Euclídea es adecuada y suficiente para identificar los clusters en el conjunto de datos. En cambio para el caso de la figura (B), la distancia Euclídea no será la medida de similitud que dé lugar a la identificación de los clusters que son fácilmente identificables a simple vista por el ser humano.

A continuación se va a mostrar un ejemplo que va a demostrar como, siguiendo diferentes criterios de similitud, se pueden obtener distintas soluciones de agrupamiento, siendo la óptima aquella que mejor se ajuste a la consecución de la intención del usuario.

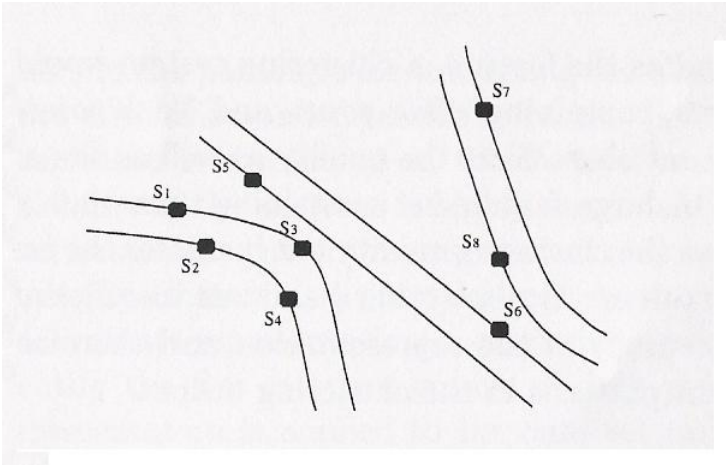


Figura 3.3: Usando distintos medidas de proximidad, las ocho cámaras en las tres carreteras pueden ser agrupadas de diferentes maneras

Medida	Resultados del clustering
Distancia Geométrica	$C_1 = \{S_1, S_2, S_3, S_4, S_5\}$ $C_2 = \{S_6, S_8\}$ $C_3 = \{S_7\}$
Conectividad	$C_1 = \{S_1, S_2, S_3, S_4\}$ $C_2 = \{S_5, S_6\}$ $C_3 = \{S_7, S_8\}$
Densidad	$C_1 = \{S_1, S_2, S_3, S_4, S_5\}$ $C_2 = \{S_6, S_7, S_8\}$

Algoritmos de clustering

En esencia, el objetivo de estos algoritmos es conseguir dos cosas:

- Maximizar la similitud entre las muestras pertenecientes a un mismo cluster y,
- Maximizar la diferencia entre los distintos clusters.

A pesar del gran número de algoritmos disponibles de clustering, no hay un método que sea capaz de dar buenos resultados para cualquier problema de agrupamiento. Por ese motivo, la elección de un determinado algoritmo para una determinada tarea será fundamental en el resultado de la misma. Además, normalmente el funcionamiento de los algoritmos, y por tanto de sus resultados, depende de una configuración óptima de sus parámetros internos, los cuales son habitualmente decididos en base a una batería de pruebas empíricas sobre el específico conjunto de datos.

3.2. K-means

Es uno de los más simples algoritmos de aprendizaje no supervisado para solucionar el conocido problema de clustering. Según las posibles clasificaciones de estos algoritmos vista en el apartado 3.1.1, se trata de un algoritmo de clustering particional y exclusivo. Además, como los clusters resultantes son disjuntos entre sí estamos ante un algoritmo de *grisp* clustering.

En primer lugar, los objetos que queremos agrupar deben ser representados como un conjunto de variables numéricas (vector). Además el usuario tiene que especificar el número de grupos (lo que nos referimos con K) que él desea identificar.

Cada objeto, por lo tanto, es representado como un vector de variables en un espacio de n dimensiones, siendo n el número de variables usadas para describir cada objeto. Entonces, el algoritmo elige aleatoriamente K puntos en el espacio de vectores, sirviendo esos puntos como los centroides iniciales de los clusters. Posteriormente, todos los objetos son asignados cada uno a su centroide más cercano. Normalmente la medida de similitud que determina cuán lejos o cerca (según esa métrica) está cada objeto de cada centroide y, por tanto, responsable de la asignación de los objetos a los distintos clusters, es elegida por el usuario. Como es lógico, la elección de una u otra medida de similitud condicionará los resultados que se obtengan. En el Cuadro 3.1 se pueden observar algunas conocidas métricas.

Después, para cada cluster un nuevo centroide es calculado promediando todos los vectores de variables que han sido asignados al cluster en cuestión. Este proceso de asignar objetos a los distintos clusters y recalcular centroides es repetido hasta que el proceso converja. Se puede demostrar que este algoritmo siempre converge tras un número finito de iteraciones.

Algunos aspectos relacionados con las medidas de similitud, la elección de los centroides iniciales, el recálculo de los nuevos centroides o el número de clusters (k) son, debido a su importancia, objeto de estudio, ya que influirán de una manera importante en el resultado final.

Formalmente, el objetivo del algoritmo es minimizar la siguiente función:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{m} d(x_i^{(j)}, c_j) \tag{3.1}$$

Siendo m el número de objetos a agrupar y los c_j los centroides.

En pseudocódigo:

1. **begin initialize:** $c_1, c_2, ..., c_k$

2. **do:** clasificar los m objetos según su centroide más cercano
   recalcular los $\boldsymbol{c_j}$
3. **volver al punto 2 (until)**, hasta que no haya cambios en los $\boldsymbol{c_j}$
4. **return** los $\boldsymbol{c_j}$
5. **end**

Euclídea	$d(i,l) = [\sum_{j=1}^{p} \|x_{ij} - x_{lj}\|^2]^{1/2}$
Manhattan	$d(i,l) = [\sum_{j=1}^{p} \|x_{ij} - x_{lj}\|]$
Angular	$d(i,l) = \arccos[\frac{x_i^t x_l}{\|\|x_i\|\|\|x_l\|\|}]$
Distancia de Mahalanobis	$d(i,l) = [(\boldsymbol{x_i} - \boldsymbol{x_l})^t \Sigma^{-1} (\boldsymbol{x_i} - \boldsymbol{x_l})]^{1/2}$

Cuadro 3.1: Distancias de similitud

Más información sobre el k-means en [1] y [29].

3.2.1. Evaluar Clusters K-means

Cada subconjunto o cluster está formado por un conjunto de muestras que son, de alguna manera, más similares entre ellas que con las muestras de los otros clusters.

Nuestra intención es medir la calidad de todas las particiones de las muestras. Para dicho objetivo se pueden utilizar diferentes criterios, de los cuales presentaré dos de ellos a continuación.

Criterio de la suma del error cuadrático

Siendo n_i el número de muestras en el cluster D_i, c el número de clusters y m_i la media de cada cluster,

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x. \tag{3.2}$$

Entonces la suma del error cuadrático queda definido como:

$$J_e = \sum_{i=1}^{c} \sum_{x \in D_i} ||x - m_i||^2 \tag{3.3}$$

El valor de J_e depende de cómo estén las muestras agrupadas y del número de clusters. Una forma de reducir J_e sería aumentando el número de clusters, aunque esto podría llevar a un modelo de agrupación no óptimo.

Este criterio, J_e, es óptimo para medir la calidad de problemas en donde las muestras se presentan como nubes compactas y éstas están claramente separadas entre ellas.

Criterio de la mínima varianza

$$J_e = \frac{1}{2}\sum_{i=1}^{c} n_i \bar{s_i}, \tag{3.4}$$

donde n_i es el número de muestras del cluster i-ésimo y

$$\bar{s_i} = \frac{1}{n_i^2}\sum_{x\in D_i}\sum_{x'\in D_i} ||x - x'||^2 \tag{3.5}$$

Pudiendo reemplazar $\bar{s_i}$ por la media, la mediana, o quizás la máxima distancia entre dos puntos del mismo cluster.

Otros criterios para evaluar la calidad de los clusters en [28].

3.3. Mapas auto-organizados

3.3.1. Introducción

Los mapas auto-organizados (*Self-Organizing maps*, SOM) son una técnica de visualización de datos inventada por el Profesor Teuvo Kohonen consistentes en una red neuronal que lleva a cabo aprendizaje no supervisado para reducir la dimensión de los datos. El problema que pretende solucionar es que los humanos no pueden visualizar datos de alta dimensionalidad. Otras dos técnicas que reducen la dimensionalidad de los datos son IsoMap y Locally Linear Embedding. Estas técnicas son explicadas con una mayor profundidad en [20] y [26]. La manera en la que los SOMs reducen la dimensionalidad generando un mapa, normalmente, de una o dos dimensiones es "trazando" las similitudes de los datos agrupando datos con características similares juntas. Así, las SOMs llevan a cabo dos tareas, la reducción de la dimensionalidad y la exposición de similitudes.

La Figura 3.4 da una idea de la apariencia de un SOM para un ejemplo en el que se quieren organizar diferentes colores. Se puede observar cómo los colores con tonalidades parecidas se van agrupando en regiones cercanas entre ellas.

Figura 3.4: Ejemplo de SOM

3.3.2. Componentes

Conjunto de datos

La idea de los mapas auto-organizados es proyectar los datos de n-dimensiones en algo que puede ser entendido mejor visualmente, habitualmente una representación de una o dos dimensiones.

Para demostrar visualmente el proceso seguiremos con el ejemplo de los colores, que pueden ser representados con vectores de 3 dimensiones si, como en este caso, usamos un sistema de codificación RGB. En este ejemplo, esperaríamos que en el mapa auto-organizado final el azul oscuro acabase cerca del gris o el amarillo claro cerca del blanco.

Centroides

Una parte primordial de las SOM son los centroides. Cada centroide tiene dos componentes. La primera componente son sus datos, que son de la misma dimensión que la de los conjuntos de datos, y su segunda componente es su localización.

Siguiendo con el ejemplo de los colores:

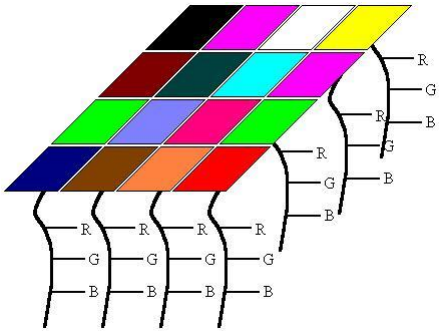


Figura 3.5: Mapa de centroides

Como vemos en la Figura 3.5, el mapa final está caracterizado por un conjunto de centroides, los cuales tienen su vector de datos que definen sus componentes RGB y su posición dentro del mapa.

Algoritmo

La manera en que las neuronas de las SOMs se organizan a sí mismos es compitiendo entre ellas para la representación de las muestras. Se permite que las neuronas vayan evolucionando aprendiendo de las muestras para conseguir ser los más parecidas a ellas y así, cada vez que se entrene con una nueva muestra del conjunto, aumentar la esperanza de ser la más parecida a esta nueva muestra. Este proceso de selección y aprendizaje hace que los centroides se organicen a sí mismos en un mapa representando similitudes.

Grosso modo, lo que hace el algoritmo es, tras inicializar los centroides, seleccionar una muestra aleatoriamente y buscar en el mapa qué centroide es el que le representa mejor. Tanto el centroide ganador, es decir el que mejor representa a la muestra, como los centroides vecinos aprenden de esta muestra para intentar parecerse algo más a ella. Para conseguir que con el paso del tiempo el mapa se vaya estabilizando se hace, ya se verá cómo, que tanto el número de vecinos que tienen un aprendizaje significativo, como la "cantidad" que aprenden vaya decreciendo con el paso de las iteraciones. Este proceso se repite iterativamente para todas las muestras.

A continuación se explicará detalladamente cada etapa del algoritmo.

3.3.3. Etapas

Inicialización de los centroides

Hay un gran número de formas de posibles inicializaciones de centroides. En función de cada caso y del conocimiento que se tiene sobre los datos puede provocar que unos buenos valores iniciales hagan que se consiga generar un buen mapa en un menor número de iteraciones, lo que provoca un ahorro de tiempo al usuario.

La Figura 3.6 muestra tres posibles inicializaciones para el caso de los colores.

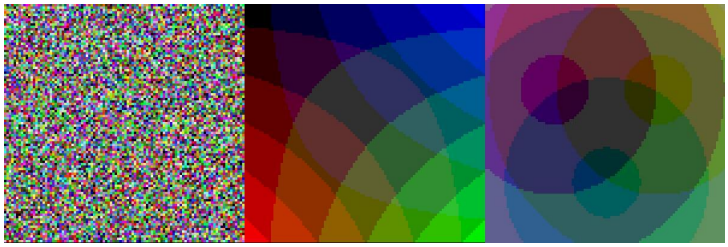


Figura 3.6: Tres posibles inicializaciones

Obtener el Best Matching Unit(BMU)

Este es un paso muy simple, lo único que hay que hacer es calcular la distancia de la muestra a todos y cada uno de los centroides. El centroide con la distancia más pequeña a la muestra es el ganador, al cual se denomina *Best Matching Unit* (BMU). Si hubiera más de un centroide con la misma distancia más pequeña, el centroide ganador es elegido aleatoriamente entre ellos. Existen diferentes métricas para medir la distancia de la muestra a cada centroide, siendo la métrica más común la distancia euclídea:

$$\sqrt{\sum_{i=0}^{n}(x_i - c_i)^2} \tag{3.6}$$

donde x_i es la nueva muestra de datos, c_i un centroide y n es el número de dimensiones de los datos, que, como se ha dicho, coincide con la dimensión de los centroides.

Escalado de vecinos

Consistente en dos partes: determinar qué centroides son considerados como vecinos y cuánto aprenden del vector muestra.

Determinación de vecinos Los vecinos del centroide ganador pueden ser determinados de muy distintas maneras. Habitualmente, que es la que utiliza la librería SOM Toolbox para Matlab, se determinan usando una función gaussiana.

Visualmente este proceso se puede apreciar si nos imaginásemos que colocásemos una gaussiana "encima" del mapa y centrada en el centroide ganador, provocando que lo que aprende cada vecino está ponderado por el valor de la gaussiana en su posición en el dominio espacial. Por tanto, cuanto más lejos se encuentre el vecino del peso ganador, menos aprenden. Ver Figura 3.7.

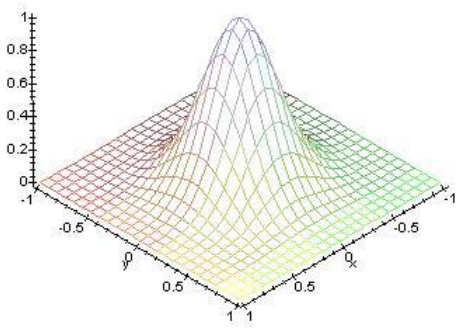


Figura 3.7: Función gaussiana que pondera lo que aprenden los vecinos

Como queremos que conforme pase el tiempo (iteraciones), la solución que se vaya obteniendo se estabilice, para ello, además de reducir lo que aprenden de cada nueva muestra (ver siguiente apartado), se hace que la varianza de la gaussiana vaya disminuyendo para que los vecinos que se ponderan por un valor a tener en cuenta, determinado por la gaussiana, sean cada vez menos.

Aprendizaje El centroide ganador es premiado siendo más parecido a la nueva muestra. Los vecinos también son recompensados aumentando su parecido con la nueva muestra, pero como se ha dicho antes, este aprendizaje está ponderado por su distancia al peso ganador.

El nuevo valor del centroide ganador (para no perder generalidad será referenciado como k) de la iteración t, será igual a:

$$Centroide_k^t = Centroide_k^{t-1} * (1 - \alpha(t)) + x^t * \alpha(t) \tag{3.7}$$

Así, en la primera iteración (t igual a 1) el valor de $\alpha(t)$ será igual a 1, por lo que el primer BMU tendrá un valor exactamente igual a la primera muestra (x^1).

Conforme disminuye $\alpha(t)$, se puede ver en 3.7 que lo que aprenden de la nueva muestra es menor.

El siguiente pseudocódigo nos puede dar una visión general de todas las etapas que conforman esta técnica de aprendizaje no supervisado:

1. Inicializar los centroides.
2. Dar un valor a $\alpha(t)$ de 0 a 1. Normalmente al principio del algoritmo $\alpha(t)$ toma un valor igual o cercano a 1.

3. Para cada muestra del conjunto:

   a) Seleccionar aleatoriamente una de las muestras.

   b) Obtener el Best Matching Unit.

   c) Escalado de vecinos: determinación de vecinos y aprendizaje.

   d) Disminuir el valor de $\alpha(t)$ una cantidad pequeña.

3.3.4. Determinación de la calidad del SOM

Siguiendo con el ejemplo de los colores, probablemente el SOM final mostrado en 3.4 es o esté muy cerca del SOM ideal para este caso. Pero como ya se ha comentado, salvo que siempre se inicialice igual los vectores de pesos, los SOMs resultantes pueden tener aspectos muy dispares. Otro posible SOM para este caso podría ser el mostrado en la Figura 3.8.

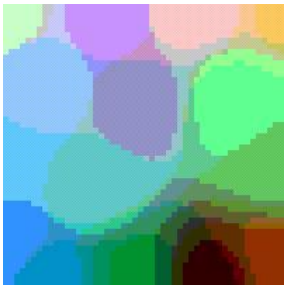


Figura 3.8: Otro posible SOM

Como vemos en la Figura 3.8, a primera vista podría parecer que este otro mapa es bastante acertado, ya que los colores similares son agrupados juntos. Sin embargo, hay algunos colores rodeados por otros que, aunque tengan cierta relación, no serían sus vecinos en el caso de que el mapa fuera elaborado por una persona.

Entonces, ¿cómo sabemos que dos centroides cercanos entre sí son realmente parecidos entre sí y no ha sido cuestión de azar? Un método sencillo consiste en calcular la distancia de cada centroide a sus vecinos para posteriormente promediar estas distancias. Para observar gráficamente las similitudes entre vecinos, si el valor promediado es alto entonces significa que presenta diferencias notorias con sus vecinos y asignaremos el color negro en la localización de ese peso. Si por el contrario esta distancia es baja, un color más claro será asignado. En la Figura 3.9 se muestra la aplicación del mencionado método de comprobación de calidad al mapa 3.8.

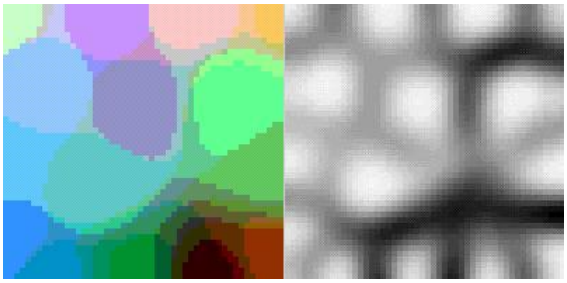


Figura 3.9: Calidad del SOM

Mirando al SOM en blanco y negro podremos obtener una primera aproximación de la calidad del mapa auto-organizado obtenido.

3.3.5. Conclusiones

Ventajas

Probablemente uno de los mejores aspectos de los mapas auto-organizados es su simplicidad tanto teórica como a la hora de ser manejados por un usuario de una manera efectiva. Otro gran aspecto es que funcionan bastante bien, además de que se puede evaluar fácilmente la calidad del mapa.

Desventajas

El mayor problema de los mapas auto-organizados es que conforme aumenta la dimensión de los datos se incrementa de manera notable el coste computacional, lo cual puede ser un gran inconveniente.

La información presente en este apartado sobre los SOM y más, puede ser encontrada en [7], [11], [12] y [21].

Capítulo 4

Análisis de componentes principales

Esta técnica es una de las más famosas dentro de la rama del análisis multivariante para la visualización de datos de alta dimensionalidad y para el pre-procesado de los datos. Presentado por Pearson en 1901 y desarrollado posteriormente por Hotelling, no fue ampliamente usada hasta la llegada de los ordenadores.

El Análisis de Componentes Principales (*Principal Component Analysis*, PCA) es un método no supervisado en el que, como sucede en este tipo de técnicas, no se usa la información de salida, en nuestro caso el coste C.

Usada en una amplia variedad de aplicaciones, la idea central de esta técnica es reducir, mediante una transformación lineal, la dimensionalidad de un conjunto de datos en el que, supuestamente, ciertas variables están relacionadas entre sí o "miden" lo mismo bajo distintos puntos de vista, manteniendo tanta varianza (o información) como sea posible en el nuevo conjunto de datos. De este modo, si queremos representar unos datos de dimensionalidad k en uno de n ($n < k$) esta técnica proporciona el menor error de reconstrucción en el espacio n-dimensional.

Considerando el PCA como el análisis espectral de la matriz de covarianza, me centraré en la explicación teórica y matemática de este análisis desde dos puntos de vista o propósitos de aplicación.

4.1. Reducción de la dimensionalidad

Consideremos el caso sencillo en el que recogemos 3 variables para definir cada muestra de un conjunto de interés.

Este es un caso irreal, y en el que raramente se utilizará PCA para intentar reducir la

dimensionalidad del problema, pero a la vez suficientemente simple para demostrar cómo se puede reducir la complejidad de un problema transformando las variables originales. Además, tiene la ventaja de que al tratarse de un problema tridimensional podemos seguir el proceso gráficamente.

Tenemos la siguiente "nube" de datos:

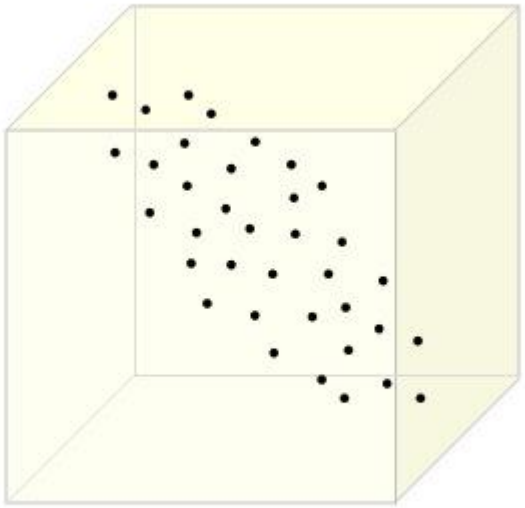


Figura 4.1: Datos Originales

En este caso estamos interesados en conseguir "mapear" este espacio de 3 dimensiones a uno nuevo de menor dimensión con la menor pérdida de información. En este ejemplo lo transformaremos a un espacio bidimensional.

Como vemos, se trata de un espacio con 3 ejes (x_1,x_2,x_3). Nuestro objetivo es encontrar unos componentes principales (w_1,w_2,,w_3) tales que la primera componente principal w_1 sea la dirección que maximiza la varianza de la proyección de los datos sobre ella y que el resto de componentes w_2 y w_3 maximice también la varianza de la proyección de los datos sobre ellas sujetas a que sean ortogonales entre sí y a la componente principal.

Siendo la proyección de $\boldsymbol{x}$, cuya matriz de covarianza es $Cov(x) = \boldsymbol{\Sigma}$, sobre las componentes principales $\boldsymbol{w}$ igual a $\boldsymbol{z} = \boldsymbol{w}^t\boldsymbol{x}$ resulta que:

$$z_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 \tag{4.1}$$
$$z_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 \tag{4.2}$$
$$z_3 = w_{31}x_1 + w_{32}x_2 + w_{33}x_3 \tag{4.3}$$

Sería lógico pensar que para maximizar la varianza de las proyecciones en las direcciones de los componentes principales $\boldsymbol{w}$ podríamos otorgar a estos w_{ji} valores muy altos. Y así es, pero se impone que el módulo de cada $\boldsymbol{w}_j^t = (w_{j1}, w_{j2}, w_{j3})$ sea igual a una constante de normalización, esto es, $||\boldsymbol{w}_j|| = cte$, donde en la mayoría de las ocasiones esta constante es igual a 1. Además, se impone que las componentes $\boldsymbol{z}$, (z_1, z_2, z_3), sean incorreladas entre sí, o lo que es lo mismo, $Cov(z_i, z_j) = 0$ para toda i distinta de j.

Por tanto, el objetivo es maximizar de manera independiente la varianza de cada componente z_i sujeto a que $||\boldsymbol{w}_j|| = \boldsymbol{w}_j^t\boldsymbol{w}_j = 1$ y a que sean ortogonales entre sí las diferentes componentes, lo cual puede ser expresado y solucionado con el operador de Lagrange.

Para la componente z_1 tenemos:

$$Var(z_1) = \boldsymbol{w}_1^t\boldsymbol{\Sigma}\boldsymbol{w}_1 \qquad\qquad \text{siendo } \boldsymbol{w}_1^t\boldsymbol{w}_1 = 1 \tag{4.4}$$

$$\arg\max_{\boldsymbol{w}_1} \ \boldsymbol{w}_1^t\boldsymbol{\Sigma}\boldsymbol{w}_1 - \alpha(\boldsymbol{w}_1^t\boldsymbol{w}_1 - 1) \tag{4.5}$$

Derivando respecto a $\boldsymbol{w}_1$ e igualando a cero:

$$\begin{aligned} 2\boldsymbol{\Sigma}\boldsymbol{w}_1 - 2\alpha\boldsymbol{w}_1 = 0 \\ \boldsymbol{\Sigma}\boldsymbol{w}_1 = \alpha\boldsymbol{w}_1 \end{aligned} \tag{4.6}$$

Llegados a este punto, y sabiendo que $\boldsymbol{\Sigma}$ al ser una matriz de covarianza es cuadrada $(m \times m)$, nos percatamos que $\boldsymbol{\Sigma}$ debe cumplir $\boldsymbol{\Sigma}v_i = \gamma_i v_i$, siendo v_i un autovector de $\boldsymbol{\Sigma}$ y γ_i el correspondiente autovalor.

Identificando términos en 4.6, podemos llegar a la conclusión que $\boldsymbol{w}_1$ es igual a un autovector de $\boldsymbol{\Sigma}$ y α a su correspondiente autovalor. Como queremos que esta primera componente principal maximice la varianza de la proyección de los datos sobre ella, $\boldsymbol{w}_1$ será igual a aquel autovector de $\boldsymbol{\Sigma}$ con el mayor autovalor asociado.

La segunda componente principal $\boldsymbol{w}_2$ deberá maximizar también la varianza de la proyección de los datos sobre ella $(\boldsymbol{w}_2^t\boldsymbol{\Sigma}\boldsymbol{w}_2)$, tener módulo 1 $(\boldsymbol{w}_2^t\boldsymbol{w}_2 = 1)$ y que sea ortogonal

a $\boldsymbol{w}_1$ ($\boldsymbol{w}_2^t \boldsymbol{w}_1 = 0$).

Al igual que antes, esto puede ser expresado y solucionado con el operador de Lagrange.

$$\arg\max_{\boldsymbol{w}_2} \ \boldsymbol{w}_2^t \boldsymbol{\Sigma} \boldsymbol{w}_2 - \alpha(\boldsymbol{w}_2^t \boldsymbol{w}_2 - 1) - \beta(\boldsymbol{w}_2^t \boldsymbol{w}_1 - 0) \tag{4.7}$$

Derivando respecto a $\boldsymbol{w}_2$ e igualando a 0 obtenemos la siguiente expresión:

$$2\Sigma \boldsymbol{w}_2 - 2\alpha \boldsymbol{w}_2 - \beta \boldsymbol{w}_1 = 0 \tag{4.8}$$

Multiplicando por $\boldsymbol{w}_1^t$ la expresión tenemos:

$$2\boldsymbol{w}_1^t \boldsymbol{\Sigma} \boldsymbol{w}_2 - 2\alpha \boldsymbol{w}_1^t \boldsymbol{w}_2 - \beta \boldsymbol{w}_1^t \boldsymbol{w}_1 = 0 \tag{4.9}$$

Sabiendo que $\boldsymbol{w}_1^t \boldsymbol{w}_2 = 0$ por ortogonalidad, y que $\boldsymbol{w}_1^t \boldsymbol{w}_1 = 1$ por imposición, nos queda:

$$2\boldsymbol{w}_1^t \boldsymbol{\Sigma} \boldsymbol{w}_2 = \beta \tag{4.10}$$

Lo cual tiene que ser 0 ya que $Cov(z_1, z_2) = 0$, por tanto $\beta = 0$.

Reduciendo la expresión 4.8 a $2\boldsymbol{\Sigma} \boldsymbol{w}_2 - 2\alpha \boldsymbol{w}_2 = 0$, tenemos el mismo caso que para la componente $\boldsymbol{w}_1$, donde en este caso $\boldsymbol{w}_2$ corresponderá al autovector de $\boldsymbol{\Sigma}$ con el segundo autovalor asociado más grande.

De manera análoga, el resto de componentes principales corresponderán a los autovectores de la matriz de covarianza con autovalores decrecientes.

Gráficamente obtenemos:

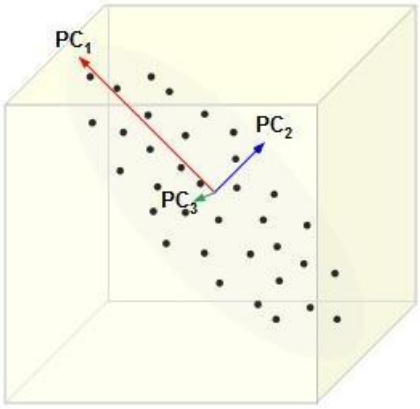


Figura 4.2: Componentes Principales

Cuya matriz de covarianza será:

$$\mathbf{\Sigma}_z = \begin{pmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{pmatrix} \tag{4.11}$$

Siendo $\alpha_i \geq \alpha_{i+1}$.

Como se puede observar en la Figura 4.2, es la primera componente principal ($\boldsymbol{w}_1$) la componente sobre la cual radica una mayor varianza de los autovectores existentes. Si quisiésemos reducir la dimensionalidad del problema, por motivos tales como que algunos algoritmos presentan problemas al manejar vectores de gran tamaño o sencillamente para "resumir" unos datos en su información más importante, deberíamos seleccionar las i-ésimas primeras componentes de las p disponibles (i≤p). Es en este momento cuando se produce una pérdida de información y no antes, ya que el conjunto de todas las componentes principales contiene exactamente la misma información que el vector de variables original.

Si decidimos reducir el problema del ejemplo a un problema bidimensional nos quedamos exclusivamente con las proyecciones de los datos sobre las dos primeras componentes principales, resultando:

$$\begin{aligned} z_1 &= w_{11}x_1 + w_{12}x_2 + w_{13}x_3 \\ z_2 &= w_{21}x_1 + w_{22}x_2 + w_{23}x_3 \end{aligned} \tag{4.12}$$

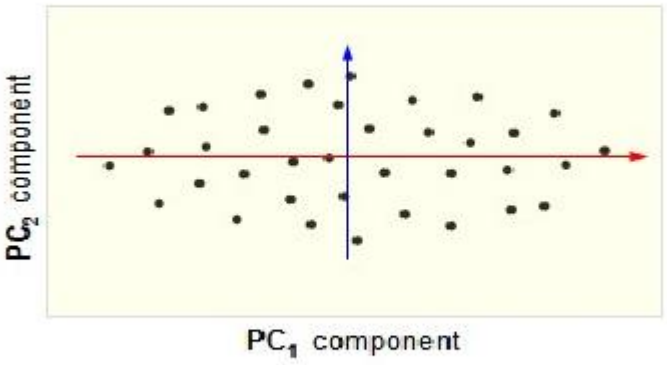


Figura 4.3: Redución de dimensionalidad

Bibliografía de interés del PCA [10], [7] y [14].

4.2. "Blanqueado" de los datos

Este proceso es bastante similar al del caso anterior, la única diferencia es que, como veremos más adelante, se multiplica a los componentes principales por una determinada matriz que nos permite normalizar la matriz de covarianza de $\boldsymbol{z}$.

Siendo $Cov(x) = \boldsymbol{\Sigma}$, se puede descomponer en autovalores y autovectores de la manera: $\boldsymbol{\Sigma} v_i = \gamma_i v_i$. Sabiendo además que esta matriz de covarianzas es de $d \times d$, tenemos d autovectores con sus d autovalores asociados. Englobando estos en un caso general nos queda:

$$\boldsymbol{\Sigma V} = \boldsymbol{VD} \tag{4.13}$$

donde:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 | & \boldsymbol{v}_2 | & \dots & | \boldsymbol{v}_d \end{bmatrix}, \ \boldsymbol{D} = \begin{bmatrix} \gamma_1 & 0 & \dots & 0 \\ 0 & \gamma_2 & \dots & 0 \\ \dots & \dots & \ddots & 0 \\ 0 & 0 & 0 & \gamma_d \end{bmatrix} \tag{4.14}$$

Despejando Σ queda:

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1} \tag{4.15}$$

Como sabemos que $\boldsymbol{\Sigma}$ es una matriz simétrica, sus autovectores $\boldsymbol{v}_i$ son ortogonales entre sí y el determinante de $\boldsymbol{V}$ es 1; por tanto, se cumple que la matriz inversa de $\boldsymbol{V}$ es igual a su traspuesta ($\boldsymbol{V}^{-1} = \boldsymbol{V}^t$).

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^t \\ E[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^t] &= \boldsymbol{V}\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{V}^t \end{aligned} \tag{4.16}$$

Llamando, por comodidad, a $\boldsymbol{V}\boldsymbol{D}^{1/2} = \boldsymbol{Q}$ y a $\boldsymbol{D}^{-1/2}\boldsymbol{V}^t = \boldsymbol{Q}^{-1}$ tenemos que:

$$E[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^t] = \boldsymbol{Q}\boldsymbol{Q}^t \tag{4.17}$$

Para diagonalizar y normalizar la matriz de covarianza, esto es, convertirla en la matriz unidad $\boldsymbol{I}$, debemos multiplicar en ambos lados de la igualdad por $\boldsymbol{Q}^{-1}$ y $\boldsymbol{Q}^{-t}$.

$$\begin{aligned} \boldsymbol{Q}^{-1}E[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^t]\boldsymbol{Q}^{-t} &= \boldsymbol{Q}^{-1}\boldsymbol{Q}\boldsymbol{Q}^t\boldsymbol{Q}^{-t} \\ E[\boldsymbol{Q}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^t\boldsymbol{Q}^{-t}] &= \boldsymbol{I}\boldsymbol{I} \\ E[(\boldsymbol{z} - \boldsymbol{\mu}_z)(\boldsymbol{z} - \boldsymbol{\mu}_z)^t] &= \boldsymbol{I} \end{aligned} \tag{4.18}$$

Por tanto, en este caso los componentes principales serán iguales a, como en el caso anterior, los autovectores de la matriz de covarianza de X por una matriz $\boldsymbol{D}^{-1/2}$ responsable de la normalización de la matriz de covarianza.

Resultando:

$$\boldsymbol{Z} = \boldsymbol{W}^t\boldsymbol{X} = \boldsymbol{D}^{-1/2} \cdot \underbrace{\boldsymbol{V}^t\boldsymbol{X}}_{Caso\ anterior} \tag{4.19}$$

Con $\boldsymbol{\Sigma}_z = \boldsymbol{I}$ y $\boldsymbol{\mu}_z = E[\boldsymbol{D}^{-1/2}\boldsymbol{V}^t\boldsymbol{X}]$.

La normalización de la matriz de covarianza puede resultarnos de interés cuando, como en este proyecto, vamos a utilizar máquina de vectores soporte como técnica de regresión/clasificación. Este tipo de máquinas requieren, para un correcto funcionamiento, la "configuración" de determinados hiperparámetros. Blanqueando los datos conseguimos normalizar todas las dimensiones de los datos, de manera que si representásemos los datos en el espacio de dimensión n formarían una nube en el que los datos estarían distribuidos de una manera más predecible, y el barrido de los hiperparámetros se haría sobre una serie de valores más "normales". Es decir, a la hora de hacer barridos de hiperparámetros el

hecho de conocer a priori cuál va a ser la distribución de los datos nos beneficia.

Gráficamente lo que se consigue es lo que muestra la Figura 4.4, pero en este proyecto en cuestión sobre un espacio de dimensión 9.

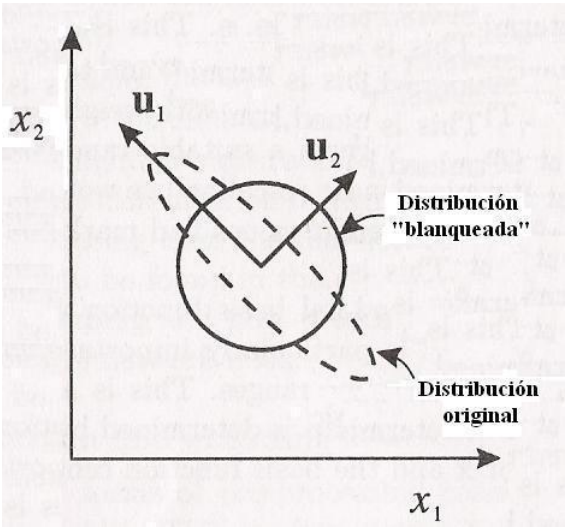


Figura 4.4: Ilustración del blanqueado de un conjuntos de datos bidimensionales

Capítulo 5

Técnicas de regresión y clasificación

El objetivo de este capítulo es dar al lector una idea de qué pretenden y cómo funcionan las técnicas de regresión y clasificación para luego adentrarnos en éstas y presentar las máquinas de vectores soporte tanto para regresión como para clasificación, y una de las más simples técnicas de regresión, la regresión lineal.

5.1. Introducción a la clasificación

El problema de la clasificación es uno de los primeros que aparecen en la actividad científica y constituye un proceso consustancial con casi cualquier actividad humana, de tal manera que en la resolución de problema y en la toma de decisiones la primera parte de la tarea consiste precisamente en clasificar el problema o la situación, para después aplicar la metodología correspondiente y que en buena medida dependerá de esa clasificación.

Cuando hablamos de clasificar a un sujeto en un grupo determinado, a partir de los valores de una serie de parámetros medidos u observados, esa clasificación tiene un cierto grado de incertidumbre, por tanto, como es lógico, se deberá cuantificar esta incertidumbre de alguna manera, habitualmente en tanto por ciento.

5.1.1. Algunos aspectos de la teoría de aprendizaje estadístico

En el diseño de clasificadores habitualmente se utiliza el criterio de riesgo mínimo. Este criterio está basado en encontrar una función g que haga mínimo el *funcional de riesgo*. Es decir, si tenemos un problema de clasificación compuesto por l observaciones, y cada observación consiste en:

- un conjunto de atributos $\mathbf{x} \in \Re^n$, denominado vector de observación o entrada,

- una etiqueta $y \in \{1, ..., k\}$ que define la clase a la que pertenece el vector de observación y
- una función de densidad conjunta $f_{\mathbf{X},Y}(\mathbf{x}, y)$,

encontrar la función $g : \Re^n \mapsto \kappa$ que hace mínimo el funcional de riesgo, definido como:

$$R(g) = E_{\mathbf{X},Y}\{L(g(\mathbf{x}, y))\} = \int L(g(\mathbf{x}), y)\partial f_{\mathbf{X},Y}(\mathbf{x}, y) \tag{5.1}$$

Siendo $L(g(\mathbf{x}, y))$ una función de coste.
El problema radica en que habitualmente se desconoce $f_{\mathbf{X},Y}(\mathbf{x}, y)$. Se puede solucionar estimándola a partir del conjunto de muestras $\{\mathbf{x_i}, y_i\}\ \forall i$, pero esta tarea puede resultar más compleja que el propio diseño del clasificador. Otra opción sería asumir que $f_{\mathbf{X},Y}(\mathbf{x}, y)$ pertenece a una función de densidad conocida y, por tanto, sólo tendríamos que calcular los parámetros que caracterizan a dicha función. Pero corremos el riesgo de asumir una función de densidad que en realidad no se ajusta a la verdadera $f_{\mathbf{X},Y}(\mathbf{x}, y)$.

Es por ese motivo, por el cual se suele optar por estimar directamente el clasificador mediante la sustitución del funcional de riesgo por su versión muestral, conocida como riesgo empírico:

$$R_{emp}(g) = \frac{1}{l}\sum_{i=1}^{l} L(g(\mathbf{x_i}), y_i) \tag{5.2}$$

Escogiendo aquel clasificador g que minimice el riesgo empírico.

Como vemos en 5.2, el riesgo empírico es obtenido sobre un conjunto de entrenamiento de longitud l. Este error no será el mismo si aplicamos el clasificador sobre otro conjunto de muestras generados por la misma $f_{\mathbf{X},Y}(\mathbf{x}, y)$. La cota máxima de error sobre este otro nuevo conjunto (u otro cualquiera) estará determinada por la denominada cota de Vapnik:

$$R(g) \leq R_{emp}(g) + \sqrt{\frac{h\big(\log\frac{2l}{h} + 1\big) - \log\frac{\eta}{4}}{l}} \tag{5.3}$$

donde η es el intervalo de confianza y h es la llamada de dimensión Vapnik-Chervonenkis (dimensión VC) del clasificador g.

Por tanto, con probabilidad $1 - \eta$ (η toma valores entre 0 y 1) se cumple la cota de Vapnik (5.3).

Sobre esta cota se pueden apreciar 3 puntos a destacar:

1. Es independiente de $f_{\mathbf{X},Y}(\mathbf{x}, y)$.

2. No se puede calcular el valor exacto del lado izquierdo de la ecuación.

3. Conociendo la dimensión VC se puede hallar la cota máxima.

Además, la diferencia máxima entre el riesgo empírico y el verdadero, dependiente tanto del número de muestras como de la dimensión VC, es menor, para un número fijo de muestras, cuanto menor sea h.

Visto esto, si dispusiéramos de un conjunto de máquinas de aprendizaje, y eligiendo un η suficientemente pequeño, deberíamos escoger aquella que minimice el lado derecho de la ecuación 5.3.

Información extraída de [19].

Dimensión de Vapnik-Chervonenkis

Definimos la dimensión VC para un conjunto de funciones $g(\alpha)$ como el número máximo de muestras que pueden ser separables por $g(\alpha)$ sea cual sea el etiquetado de éstas.

En el caso en el que nos encontremos en $\Re^2$ y queramos analizar la dimensión VC de un clasificador lineal binario, se puede observar como la dimensión de VC es 3. Esto es así puesto que sea cual sea el etiquetado de estas 3 muestras siempre habrá una posible frontera que separe las muestras de un tipo de las de otro (en la Figura 5.1 se puede apreciar las 2^3 posibles formas de etiquetado). Si intentásemos clasificar 4 muestras, un clasificador lineal no es suficiente para poder separar las 2^4 formas posibles de etiquetado (ejemplo en la Figura 5.2).

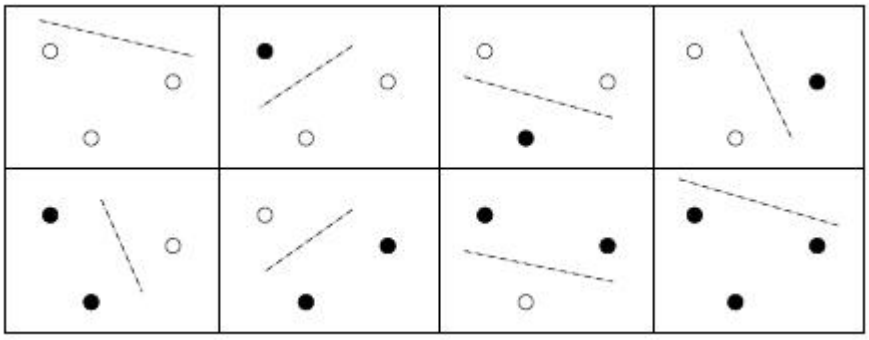


Figura 5.1: Posibles etiquetados de 3 muestras en $\Re^2$

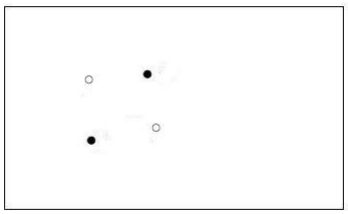


Figura 5.2: No existe clasificador lineal que pueda separar este caso

Puede demostrarse que, en el caso general de un espacio n-dimensional, la dimensión VC de la familia de los clasificadores binarios lineales (hiperplanos en el espacio n-dimensional) es $h = n + 1$.

En el caso de regresión también se puede apreciar la influencia de la dimensión VC en los resultados obtenidos. Si queremos ajustar un conjunto de l muestras a un polinomio de grado k, podemos apreciar que conforme aumenta el grado del polinomio el riesgo empírico disminuye, hasta que se hace nulo cuando el polinomio es de grado $l - 1$. Pero también, conforme aumenta el grado k aumenta la dimensión VC del polinomio (h) con el consiguiente aumento de la cota del riesgo verdadero debido al segundo término de la cota Vapnik (5.3). Por lo que se debe buscar un polinomio de grado intermedio que haga que la suma de los dos términos de la cota de Vapnik hagan mínima a ésta.

Información extraída de [19] [27].

5.1.2. Máquinas de vectores soporte para clasificación

Las máquinas de vectores soporte (*Support Vector Machine*,SVM) son un conjunto de métodos de aprendizaje supervisados usadas para clasificación y regresión.

Una propiedad especial de las SVM es que consiguen, simultáneamente, minimizar el error empírico de clasificación y maximizar el margen geométrico (explicado posteriormente), de ahí que también sean conocidos como clasificadores (en el caso de clasificación) de máximo margen.

Caso separable

Consideremos que tenemos el conjunto de muestras de entrenamiento $\{ (x_i, y_i) \}_{i=1}^{N}$, en donde $\mathbf{x}_i$ es el vector de entrada i-ésimo y y_i es la correspondiente salida deseada, pudiendo tomar los valores {-1,1}.

En el caso separable, la clase representada por el subconjunto de muestras con salida $y_i = +1$ es separable linealmente de la clase representada por el subconjunto de muestras con salida $y_i = -1$. La ecuación del hiperplano que hace de frontera entre ambos subconjuntos será de la forma:

$$\mathbf{w}^t\mathbf{x} + b = 0 \tag{5.4}$$

Siendo $\mathbf{x}$ el vector de entrada, $\mathbf{w}$ un vector de pesos ajustables y b el sesgo. Por tanto:

$$\begin{aligned} \mathbf{w}^t\mathbf{x} \ + \ b \ \geq \ 0 \quad & \text{para } y_i = +1 \\ \mathbf{w}^t\mathbf{x} \ + \ b \ \leq \ 0 \quad & \text{para } y_i = -1 \end{aligned} \tag{5.5}$$

Al ser un problema separable linealmente, las posibles soluciones, es decir, el número de posibles hiperplanos que son capaces de separar las muestras de ambas clases, son infinitas. En la Figura 5.3 se pueden apreciar 4 de las infinitas posibles soluciones.

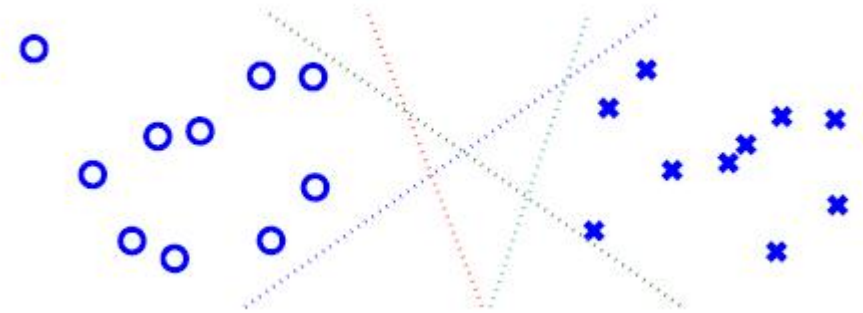


Figura 5.3: Cuatro posibles soluciones para el caso de separación lineal

Entonces, ¿qué solución elegir? Las SVM proporcionan la solución que maximiza el margen de separación entre las muestras positivas y negativas. A dicha superficie de decisión se le conoce como el *hiperplano óptimo*. Además esta frontera produce una tasa de error, vista anteriormente que es igual a la suma de la tasa de error del conjunto de entrenamiento más un término que depende de la dimensión de Vapnik-Chervonenkis (5.3), en el caso separable, igual a cero (riesgo empírico) más el valor del otro término dependiente de la dimensión VC.

La Figura 5.4 nos muestra una ilustración de la idea del hiperplano óptimo para patrones separables linealmente.

Geométricamente se puede demostrar que la distancia del origen al hiperplano (en la Figura 5.4 representado por una línea azul) es igual a $-b/||\mathbf{w}||$.

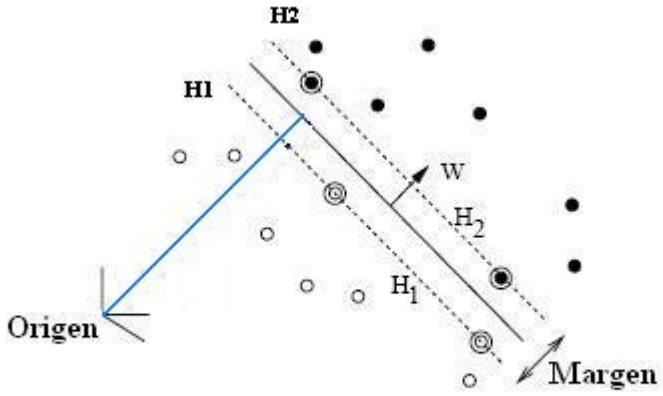


Figura 5.4: Solución de la SVM para el caso separable

Con la intención de calcular este hiperplano óptimo sujeto a las restricciones 5.5, expresamos a éstas en una sola inecuacion de la siguiente manera:

$$\begin{aligned} y_i(\mathbf{w}^t\mathbf{x}_i \ + \ b) \ &\geq \ M \quad \forall_i \\ y_i(\mathbf{w}^t\mathbf{x}_i \ + \ b) - M \ &\geq \ 0 \quad \forall_i \end{aligned} \tag{5.6}$$

Donde M es un margen que imponemos y que podemos asumir, sin pérdida de generalidad, de valor $M = 1$.

Volviendo a la Figura 5.4, el margen es igual a la distancia entre los hiperplanos $H1$ y $H2$. O lo que es lo mismo, considerando las muestras que cumplen:

$$\begin{aligned} \mathbf{w}^t\mathbf{x}_i \ + \ b = 1 \quad & para \ y_i = +1 \\ \mathbf{w}^t\mathbf{x}_i \ + \ b = -1 \quad & para \ y_i = -1 \end{aligned} \tag{5.7}$$

se puede demostrar geométricamente, que las muestras que cumplen la primera igualdad están a una distancia perpendicular del origen igual a $|1 - b|/||\mathbf{w}||$, y las muestras que cumplen la segunda igualdad están a una distancia perpendicular del origen igual a $|-1-b|/||\mathbf{w}||$. De ahí que la distancia del hiperplano óptimo respecto a cada plano $H1$ y $H2$ sea igual a $1/||\mathbf{w}||$, y por tanto el margen será igual a la suma de ambas distancias, $2/||\mathbf{w}||$.

Habiéndose demostrado que este margen es igual a $2/||\mathbf{w}||$, si queremos maximizar el margen para encontrar el hiperplano óptimo, entonces deberemos minimizar la norma Euclídea del vector de pesos $\mathbf{w}$, $||\mathbf{w}||$, pero siempre con la restricción 5.6.

Como se ve en la Figura 5.4, la solución se encuentra a una distancia igual de los

hiperplanos $H1$ y $H2$, y estos dependen exclusivamente de unas pocas de las muestras de entrenamiento, los denominados vectores soporte, cuya presencia es la que determinan a los planos $H1$ y $H2$ y, por tanto, al hiperplano óptimo.

Este es un problema que puede ser planteado y resuelto usando el método de los multiplicadores de Lagrange que, como ya sabemos, nos permite trabajar con funciones de varias variables que nos interesa maximizar o minimizar sujeto a determinadas restricciones.

Construyendo la función Lagrangiana:

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^t\mathbf{w} - \sum_{i=1}^{l} \alpha_i[y_i(\mathbf{w}^t\mathbf{x}_i + b) - 1] \tag{5.8}$$

El factor escalar $1/2$ que multiplica a la norma Euclídea de $\mathbf{w}$ está incluida aquí solo por conveniencia de presentación, pero no afecta de ningún modo a la solución final. Además,los multiplicadores de Lagrange, como sabemos, deben ser mayores que 0.

La solución al problema de optimización restringido está determinado por un punto de equilibro de $J(\mathbf{w}, b, \alpha)$ en donde $J(\mathbf{w}, b, \alpha)$ está minimizado respecto a $\mathbf{w}$ y b y maximizado con respecto a α, ya que así se demuestra que la solución obtenida es la óptima. Dicha demostración se puede encontrar en [22].

Derivando $J(\mathbf{w}, b, \alpha)$ respecto a $\mathbf{w}$ y respecto a b tenemos:

$$\begin{aligned} \textit{Condición } 1: &\quad \frac{\partial J(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = \mathbf{0} \\ \textit{Condición } 2: &\quad \frac{\partial J(\mathbf{w}, b, \alpha)}{\partial b} = \mathbf{0} \end{aligned} \tag{5.9}$$

Obteniendo para cada condición:

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \\ \mathbf{0} &= \sum_{i=1}^{l} \alpha_i y_i \end{aligned} \tag{5.10}$$

Con estas dos condiciones, más las 3 ya conocidas:

$$\alpha_i \geq 0 \tag{5.11}$$

$$y_i(\mathbf{w}^t\mathbf{x}_i + b) - 1 \leq 0 \tag{5.12}$$

$$\alpha_i(y_i(\mathbf{w}^t\mathbf{x}_i + b) - 1) = 0 \tag{5.13}$$

Tenemos las denominadas condiciones de Karush-Kuhn-Tucker (KKT) [22], condiciones necesarias y suficientes para que la solución de un problema de programación no lineal (como 5.8) sea óptima, que nos permiten definir la solución. Nótese que se trata de un sistema de ecuaciones de 3 incógnitas ($\alpha, \mathbf{w}$, b) que puede ser resuelto de una manera sencilla.

Analizando la condición 5.13, podemos apreciar como las muestras pueden ser agrupadas en dos subconjuntos:

1. Muestras que cumplen $(y_i(\mathbf{w}^t\mathbf{x}_i) + b) > 1$, y que por tanto, para que se cumpla 5.13 deben tener $\alpha_i = 0$.

2. Muestras que cumplen $(y_i(\mathbf{w}^t\mathbf{x}_i) + b) = 1$, y que por tanto, sus α_i no tienen por que ser obligatoriamente iguales a 0, ya que de todas maneras se cumple 5.13, por lo que sus $\alpha_i \geq 0$.

De modo que solo las muestras pertenecientes al segundo subconjunto ($\alpha_i \geq 0$), los denominados vectores soporte, son los que determinan la solución del hiperplano óptimo $\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$. Pudiendo reescribirse como:

$$\mathbf{w} = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i \tag{5.14}$$

El valor del sesgo b, al igual que $\mathbf{w}$, se obtiene teniendo solamente en cuenta los vectores soporte ($\alpha_i \geq 0$) y cuyo valor se puede obtener despejando b en la ecuación 5.13 de KKT.

Bibliografía usada para la SVM lineal de clasificación para el caso separable: [7], [16], [28], [22], [4] y [3].

Teoría de la Dualidad

Según el teorema de la dualidad, cada problema de programación lineal tiene un segundo problema asociado a él. Uno se denomina primal y el otro dual. Los dos poseen propiedades muy relacionadas, de tal manera que la solución óptima a un problema proporciona información completa sobre la solución óptima para el otro.

Las relaciones entre el primal y el dual se utilizan para reducir el esfuerzo de cómputo en ciertos problemas y para obtener información adicional sobre las variaciones en la solución óptima debidas a ciertos cambios en los coeficientes y en la formulación del problema.

En un caso general, sea un problema primal a maximizar sujeto a determinadas restricciones, el dual se puede obtener del primal (y viceversa) de la siguiente manera:

1. Cada restricción de un problema corresponde a una variable en el otro (Ver tabla de Tucker, Figura 5.5).
2. Los términos independientes de las restricciones de uno son los coeficientes en la función objetivo del otro.
3. Un problema busca maximizar y el otro minimizar.
4. El problema de maximización tiene restricciones $\leq$ y el de minimización tiene restricciones $\geq$.
5. Las variables en ambos casos son no negativas.

MAXIMIZACION		MINIMIZACION.	
	$\leq$	$\geq$	
RESTRICCIONES	$\geq$	$\leq$	VARIABLES
	=	><	

	$\geq$	$\geq$	
VARIABLES	$\leq$	$\leq$	RESTRICCIONES
	><	=	

Figura 5.5: Tabla de Tucker

Más información de la teoría de la dualidad en [16].

Para postular el problema dual a nuestro problema primal, primero expandiremos la expresión 5.8,

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^t\mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{w}^t \mathbf{x}_i - b\sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i \tag{5.15}$$

Usando la información de las restricciones del problema primal, observamos que el tercer término es igual a 0 debido a la segunda condición de optimización ($\sum_{i=1}^{l} \alpha_i y_i = 0$). Además, recordando la primera condición de optimización ($\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i$) del problema

primal podemos comprobar como los dos primeros términos son idénticos entre sí e iguales a:

$$\mathbf{w}^t\mathbf{w} = \sum_{i=1}^{L} \alpha_i y_i \mathbf{w}^t \mathbf{x}_i = \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t \mathbf{x}_j \tag{5.16}$$

Reformulando la función objetivo a maximizar (recordamos que en este caso, ya que en el problema primal se buscaba minimizar, el objetivo es maximizar la expresión), puesto que ahora solo depende de los multiplicadores de Lagrange α

$$\begin{aligned} Q(\alpha) &= \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t \mathbf{x}_j - \sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t \mathbf{x}_j + \sum_{i=1}^{l} \alpha_i \\ &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t \mathbf{x}_j \end{aligned} \tag{5.17}$$

Si comparamos esta función objetivo a maximizar con la función objetivo a minimizar del problema primal (5.8), donde estaba sujeto a las 5 condiciones de KKT, observamos que ahora solo tenemos que tener en cuenta dos restricciones:

$$\begin{aligned} \sum_{i=1}^{l} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \end{aligned} \tag{5.18}$$

Lo que computacionalmente es mucho más sencillo y dando lugar a la misma solución que en el caso primal.

Caso no separable

Hasta ahora nos hemos centrado en el caso linealmente separable. En este apartado consideraremos un caso más difícil, el caso de patrones no separables, en donde, dado un conjunto de entrenamiento, no nos será posible "construir" un hiperplano de separación sin encontrar errores de clasificación. No obstante, nos gustaría encontrar un hiperplano óptimo que minimice la probabilidad de error de clasificación en el conjunto de entrenamiento.

En este caso, al margen de separación entre clases lo denominamos *suave* ya que hay una o más muestras $(\mathbf{x}_i, y_i)$ que violan la siguiente condición:

$$y_i(\mathbf{w}^t\mathbf{x}_i + b) \geq +1 \quad \forall_i \tag{5.19}$$

Esta violación se puede presentar de dos maneras diferentes:

- La muestra $(\mathbf{x}_i, y_i)$ cae en el lado correcto de la superficie de decisión, pero dentro de la superficie que se encuentra entre el hiperplano óptimo y el correspondiente $H1/H2$. Ilustrado en la Figura 5.6 (A).

- La muestra $(\mathbf{x}_i, y_i)$ cae en el lado incorrecto de la superficie de decisión como se puede ver en la Figura 5.6 (B).

Nótese que en el primer caso estamos estamos clasificando correctamente, mientras que en el segundo caso no.

Para el tratamiento de este problema, por tanto, introduciremos un nuevo conjunto de variables escalares no negativas, $\{\xi_i\}_{i=1}^{l}$ en la definición del hiperplano de separación:

$$y_i(\mathbf{w}^t\mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall_i \tag{5.20}$$

Los ξ_i son llamados variables "flojas"(*slack variables*), y miden la desviación del dato a la condición ideal de separabilidad. Para $0 \leq \xi_i \leq 1$, los datos se encuentran en la situación descrita en la Figura 5.6 (A). Para $\xi_i > 1$, los datos caen en el lado incorrecto del hiperplano de separación como muestra la Figura 5.6 (B). Los vectores soporte son aquellos que, al igual que en el caso separable, conforman los hiperplanos $H1$ y $H2$.

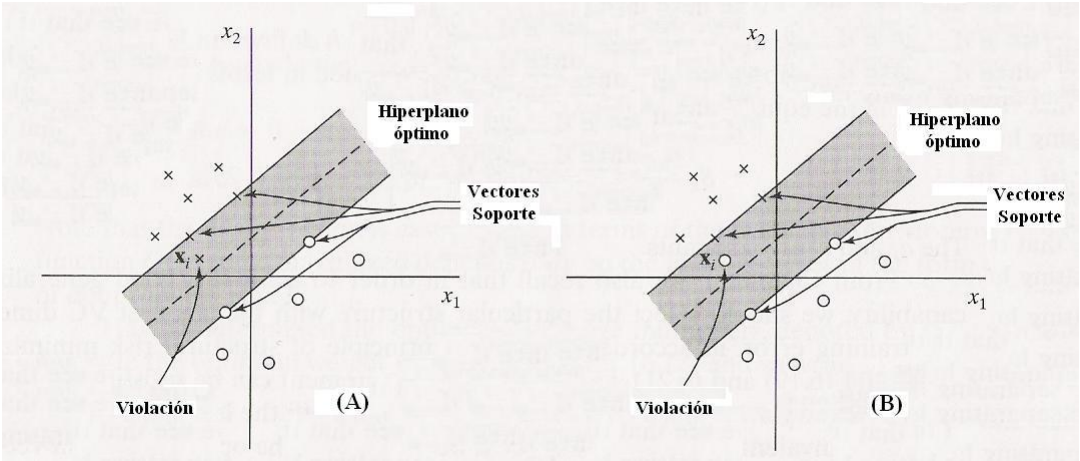


Figura 5.6: Posibles violaciones que ocurren en el caso no separable

Nuestro objetivo es encontrar un hiperplano de separación que alcance un compromiso entre la maximización del margen (minimización de $||\boldsymbol{w}||$) y la minimización de los errores. Por tanto, la función objetivo a minimizar será:

$$\frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i \tag{5.21}$$

El parámetro C sirve para buscar un compromiso entre la complejidad de la máquina y el número de puntos no separables, por lo que puede ser visto como una forma de regularización. Un coste alto implicará que damos prioridad a que se consiga una máquina con pocos errores en el conjunto de entrenamiento, aunque provoque que ésta sea compleja y no generalice correctamente para otro conjunto. Además, este parámetro C tiene que ser elegido por el usuario, lo que puede ser hecho en una de las dos siguientes maneras:

1. Experimentalmente, realizando una batería de pruebas sobre un conjunto de entrenamiento y su posterior validación en un conjunto de test o realizando validación cruzada.

2. Analíticamente, estimando la dimensión VC y usando límites en las prestaciones de las máquina basadas en está dimensión de VC estimada.

Generalizando la expresión 5.21 para incluir otras funciones de coste la expresamos como:

$$\frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + C\sum_{i=1}^{l} L(\xi_i) \tag{5.22}$$

siendo $L(\xi_i)$ la función de coste elegida. En la Figura 5.7 se muestras distintas funciones de coste. En verde la función de coste aplicada en 5.21, que, por comodidad, es la que utilizaremos para todo el desarrollo.

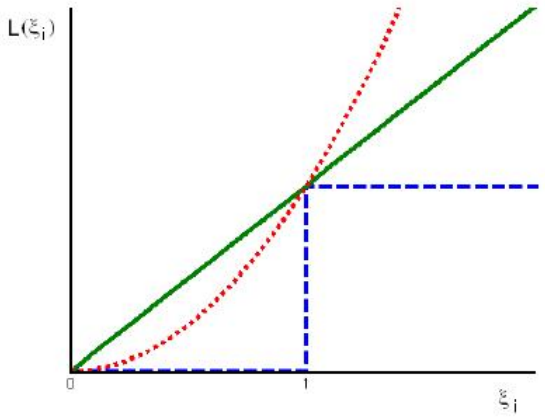


Figura 5.7: Tres funciones de coste: escalón lineal y cuadrática

Cada coste presenta una serie de características que da lugares a distintas soluciones y a máquinas de distinta complejidad.

De cualquier manera, la función objetivo a minimizar 5.21 está sujeta a las restricciones descritas en la ecuación 5.20 y a que $\xi_i \geq 0$. Construyendo la función Lagrangiana para este problema nos queda:

$$J(\boldsymbol{w}, \alpha, b, r) = \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i(y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{l} r_i\xi_i \tag{5.23}$$

Donde $\alpha_i \geq 0$ y $r_i \geq 0$, son multiplicadores de Lagrange, y por tanto no negativos. Los r_i son introducidos para asegurar la positividad de los ξ_i.

Ya se ha planteado el problema de optimización para el caso de patrones no separables. Se puede observar que el problema de optimización para patrones separables linealmente es un caso especial, concretamente cuando los $\xi_i = 0$.

La solución, al igual que en el caso separable, queda definida por las condiciones de KKT. En este caso se obtienen encontrando al mismo tiempo el mínimo de 5.23 respecto a $\boldsymbol{w}$, ξ, b y el máximo respecto a todos los multiplicadores de Lagrange, es decir derivando respecto a las mencionadas variables e igualando a 0. Lo que nos queda:

$$\begin{aligned}
\frac{\partial J(\boldsymbol{w}, \alpha, b, r)}{\partial \boldsymbol{w}} = 0 \longrightarrow \boldsymbol{w} = \sum_{i=1}^{l}\alpha_i y_i \boldsymbol{x}_i \\
\frac{\partial J(\boldsymbol{w}, \alpha, b, r)}{\partial b} = 0 \longrightarrow 0 = \sum_{i=1}^{l}\alpha_i y_i \\
\frac{\partial J(\boldsymbol{w}, \alpha, b, r)}{\partial \xi} = 0 \longrightarrow C - \alpha_i - r_i = 0
\end{aligned} \tag{5.24}$$

Más las ya conocidas condiciones 5.25 conforman las condiciones de KKT, con lo que la solución óptima queda definida. Ya solo hay que resolver el sistema de ecuaciones con 5 incógnitas ($\boldsymbol{w}, \alpha, b, \xi$ y r).

$$\begin{aligned}
y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b) &\geq 1 - \xi_i \\
\xi_i, \alpha_i, r_i &\geq 0 \\
\alpha_i(y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b) - 1 + \xi_i) &= 0 \\
r_i\xi_i &= 0
\end{aligned} \tag{5.25}$$

Al igual que en el caso separable, se puede seguir un razonamiento similar analizando algunas de las condiciones de KKT anteriores y agrupar a las muestras en 3 subconjuntos diferentes de la siguiente manera:

1. Muestras que cumplen $y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b) > 1 \rightarrow \alpha_i = 0$, $\xi_i = 0$, $r_i = C$
2. Muestras que cumplen $y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b) = 1 \rightarrow 0 \leq \alpha_i \leq C$, $\xi_i = 0$, $r_i = C - \alpha_i$
3. Muestras que cumplen $y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b) < 1 \rightarrow \alpha_i = C$, $\xi_i = 1 - y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b)$, $r_i = 0$

Así que, en este caso, los vectores soporte serán aquellos pertenecientes a los dos últimos subconjuntos y determinan el hiperplano óptimo:

$$\boldsymbol{w} = \sum_{i \in SV} \alpha_i y_i \boldsymbol{x}_i \tag{5.26}$$

Procediendo en una manera similar al del caso separable, podemos desarrollar el problema dual para el caso de patrones no separables, siendo la función a maximizar:

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^t \boldsymbol{x}_j \tag{5.27}$$

Sujeto a:

$$\begin{aligned} \sum_{i=i}^{l} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{aligned} \tag{5.28}$$

Como se aprecia, el problema dual para el caso de patrones no separables es bastante similar al del caso separable excepto por una pequeña, pero importante, diferencia. La función objetivo $Q(\alpha)$ a maximizar es la misma en ambos casos, pero difiere el uno del otro en que la restricción $\alpha_i \leq 0$ es reemplazada por una restricción más exigente $0 \leq \alpha_i \leq C$.

Bibliografía utilizada para SVM lineal de clasificación para caso no separable: [7], [16], [28], [22], [4] y [3].

5.2. Introducción a la regresión

El término regresión se utiliza usualmente para describir una situación de retroceso. En estadística el término de regresión proviene del trabajo del genetista Francis Galton (1822-1911). Su estudio se centró en la descripción de los rasgos físicos de los descendientes a partir de los de sus padres, concluyendo que los hijos de los padres más grandes tendían a tener un tamaño promedio menor que el de sus padres, mientras que los hijos de padres pequeños tendían a tener una mayor estatura que la de sus padres. Galton denominó a este fenómeno "regresión hacia la mediocridad". En la actualidad, el término de regresión se utiliza siempre que se busca predecir una variable en función de otra [30].

El análisis de regresión tiene por objetivo estimar el valor promedio de una variable, variable dependiente, con base en los valores de una o más variables adicionales, variables explicativas. Este análisis ha cobrado popularidad debido al gran número de paquetes estadísticos que lo incluyen y por ser un proceso robusto que se adapta a un sinfín de aplicaciones científicas y ejecutivas que permite la toma de decisiones.

Tipos de regresión:

- Atendiendo al número de variables independientes que se consideren en el análisis.
    1. Una variables independiente: el análisis se denomina Análisis de regresión Simple.
    2. Varias variables independientes: el análisis se denomina Análisis de regresión múltiple.
- Atendiendo al tipo de relación funcional entre variables.
    1. Lineal: La dependencia de la variable dependiente respecto a las variables independientes que la describen sigue una relación lineal.
    2. No lineal: La dependencia de la variable dependiente respecto a las variables independientes que la describen sigue una relación no lineal.

5.2.1. Regresión lineal

La importancia de las técnicas de regresión lineal radica en que nos permite observar como una variable influye sobre la otra. En el caso de que esta dependencia realmente sea lineal, esto es, el aumento o disminución de la variable X supone un aumento o disminución proporcional en la variable Y, este tipo de regresores funcionaran de una manera más o menos óptima. Ejemplos de estos pueden ser: el aumento de las temperaturas (variable X) provoca un aumento proporcional en el consumo de líquidos por parte de las personas (variable Y), o la disminución de las temperaturas (variable X) provoca un aumento en el consumo eléctrico en los hogares en calefacción (variable Y). La Figura 5.8 muestra un ejemplo de situación en el que este regresor tendrá un comportamiento aceptable.

Esta función de regresión requiere el cálculo de dos parámetros: la pendiente w y el origen de coordenadas de la recta de regresión w_0:

$$\hat{y} = wx + w_0 \tag{5.29}$$

En el caso de la regresión lineal múltiple nos encontramos ante la misma situación, solo que únicamente la variable dependiente Y dependerá de más de una variable independiente. Por lo tanto, esta función requerirá del cálculo de un vector de pesos $\boldsymbol{w}$.

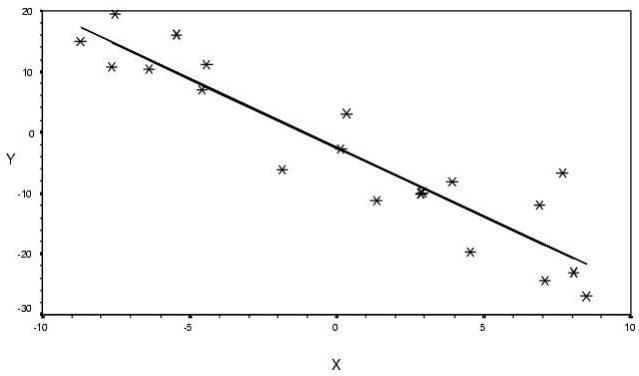


Figura 5.8: Recta de regresión

$$\hat{y} = w_0 + w_1 x_1 + ... + w_p x_p = \begin{pmatrix} w_0 & w_1 & ... & w_p \end{pmatrix} \begin{pmatrix} 1 \\ x_1 \\ ... \\ x_p \end{pmatrix} = \boldsymbol{w}^t \boldsymbol{x} \tag{5.30}$$

En el modelo anterior w_0 representa el factor constante del modelo lineal. Mientras que los demás parámetros w_i (i=1,2,...,p) indican el incremento de la variable dependiente Y por aumento unitario de la i-ésima variable independiente, suponiendo fijas el resto de variables.

El objetivo ahora es obtener la expresión que permita calcular el valor de $\boldsymbol{w}$ minimizando una determinada función de coste.

Definiendo la función de coste como C, $\boldsymbol{w}$ será aquel que:

$$min_{\vec{w}}\ C(y, \hat{y}) \tag{5.31}$$

En nuestro caso la función de coste será cuadrática (Fig. 5.9).

Por tanto:

$$min_{\boldsymbol{w}}\ C(y, \hat{y}) = min_{\boldsymbol{w}}\ (y - \hat{y})^2 = min_{\boldsymbol{w}}\ (y - \boldsymbol{w}^t x)^2 \tag{5.32}$$

Derivando respecto a $\boldsymbol{w}$ e igualando a 0 obtendremos la solución:

$$\frac{\partial C}{\partial \vec{w}} = 0 \tag{5.33}$$

$$2(y - w^t x)x^t = 0 \tag{5.34}$$

$$yx^t = w^t x x^t \tag{5.35}$$

$$w^t = yx^t(xx^t)^{-1} \tag{5.36}$$

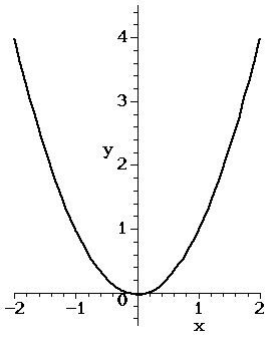


Figura 5.9: Función de coste cuadrática

Donde:

$$x = \begin{bmatrix} 1 & 1 & ... & 1 \\ x_1^1 & x_1^2 & ... & x_1^p \\ ..... & & & \\ x_n^1 & x_n^2 & ... & x_n^p \end{bmatrix}, \; y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} \tag{5.37}$$

Siendo n el número de muestras que componen nuestro conjunto de entrenamiento y p el número de variables independientes que describen a la variable dependiente Y.

Para un mayor detalle consultar [30], [1] y [18].

5.2.2. Máquinas de vectores soporte para regresión

La SVM fue en primer lugar desarrollada para clasificación. Pero para cuando se quiso que se generalizará para el caso de regresión, donde se requiere la estimación de una función que genere valores reales continuos en lugar de la determinación de ± 1 como en el caso de reconocimiento de patrones, se siguió un criterio similar al de clasificación para poder construir la función regresora.

En ambos casos, reconocimiento de patrones y regresión, se quiere minimizar la norma de $||\boldsymbol{w}||^2$. En el primer caso, como ya se demostró, para obtener una frontera de decisión que maximizara la distancia entre las muestras de las dos clases (en el caso binario), y en el segundo caso un pequeño valor de $||\boldsymbol{w}||^2$ corresponde a una función regresora sencilla.

El concepto de margen visto en clasificación se asimila en regresión mediante la introducción de la denominada función de pérdidas ϵ-insensitiva, la cual tiene la forma mostrada en la Figura 5.10.

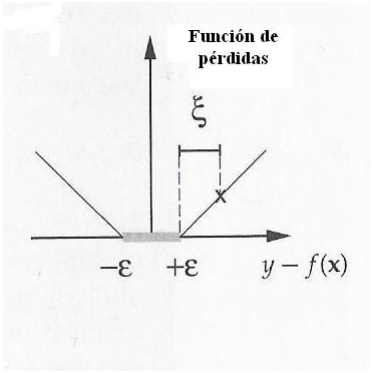


Figura 5.10: Función de pérdidas ϵ-insensitiva

Y "crea" un margen alrededor de la recta (en el caso lineal) generada por la función de regresión, en donde no se penaliza el error entre el valor real y el predicho por el regresor (5.11). De este modo, fijado el ϵ, el algoritmo obtiene aquel regresor para dicho margen que busca un compromiso, determinado por C, entre la complejidad del regresor y la importancia que se le da a los errores (en valor absoluto) mayores que ϵ.

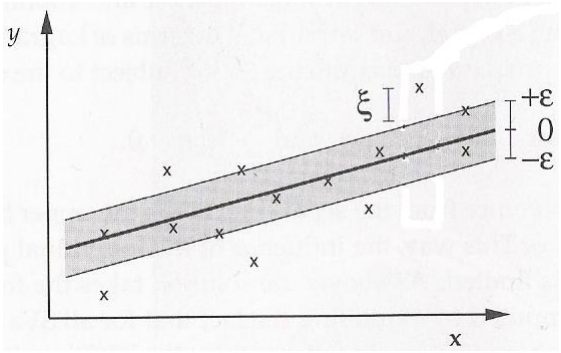


Figura 5.11: Los valores dentro del tubo de regresión de radio ϵ son considerados dentro del límite de bien predichos

Así, el algoritmo de regresión de la SVM, al cual llamaremos a partir de ahora $\varepsilon - SVR$, busca estimar la función de regresión

$$f(\boldsymbol{x}) = \boldsymbol{w}^t \boldsymbol{x} + b \tag{5.38}$$

suponiendo que el conjunto de pares $(\boldsymbol{x}_i, y_i)$ han sido originados de manera independiente e idénticamente distribuidos, que minimiza:

$$\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{l}|y_i - f(\boldsymbol{x}_i)|_\epsilon \tag{5.39}$$

Introduciendo *variables flojas*, podemos expresar este problema como un Lagrangiano similar al que se construye para resolver el problema de clasificación para el caso no separable. En este caso, necesitamos dos tipos de variables flojas, una para cuando $f(\boldsymbol{x}_i) - y_i > \epsilon$, a la que denotamos $\boldsymbol{\xi}$, y otra para cuando $f(\boldsymbol{x}_i) - y_i < \epsilon$, a la que denotamos $\boldsymbol{\xi}^*$; y colectivamente nos referiremos a ellas como $\xi^{(*)}$.

Por tanto, la función objetivo a minimizar es:

$$\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{5.40}$$

Sujeto a:

$$\begin{aligned}(\boldsymbol{w}^t\boldsymbol{x} + b) - y_i &\leq \epsilon + \xi_i \\ y_i - (\boldsymbol{w}^t\boldsymbol{x} + b) &\leq \epsilon + \xi_i^* \\ \xi_i^{(*)} &\geq 0\end{aligned} \tag{5.41}$$

Procediendo de una manera análoga al caso de clasificación, el problema primal puede ser planteado y resuelto construyendo el Lagrangiano adecuado y con la ayuda de las condiciones de KKT, las cuales permiten obtener la solución óptima. Igualmente, podemos transformar el problema primal en el dual sustituyendo en éste las relaciones obtenidas en el primal conseguidas derivando el Lagrangiano respecto a las variables a minimizar y, finalmente, maximizar la expresión resultante del dual, lo cual resulta ser un problema a resolver más sencillo computacionalmente que el primal.

Bibliografía de las SVM lineales de regresión: [7] y [22].

5.3. Máquinas de vectores soporte no lineales

Para muchos problemas de clasificación o regresión, una solución lineal no da buenos resultados. En esos casos es necesaria una aproximación no lineal.

Los datos que inicialmente pertenecen al espacio $\Re^d$, donde no pueden ser separados sin error por un hiperplano, se mapean, en primer lugar, a otro espacio (posiblemente de dimensión infinita) mediante un conjunto de transformaciones no lineales $\boldsymbol{\Psi(x)}$, donde ahora sí, en este nuevo espacio, los datos pueden ser separados (en el caso de clasificación) con menos errores por un hiperplano. La Figura 5.12 muestra gráficamente el objetivo que se desea alcanzar:

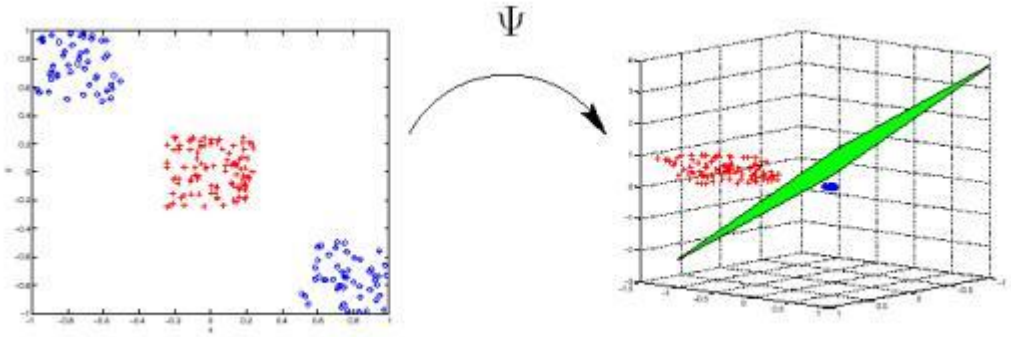


Figura 5.12: Transformación de un espacio de dimensión 2 a uno de espacio 3

Por tanto, el hiperplano que actúe como superficie de decisión será:

$$\sum_{j=1}^{m_1} w_j \Psi_j(\boldsymbol{x}) + b = 0 \tag{5.42}$$

donde los $[w]_{j=1}^{m_1}$ son el conjunto de pesos lineales, m_1 la dimensión del espacio de características y b el sesgo. Podemos simplificarlo escribiendo:

$$\sum_{j=0}^{m_1} w_j \Psi_j(\boldsymbol{x}) = 0 \tag{5.43}$$

asumiendo que $\Psi_0(\boldsymbol{x})$=1 y, por tanto, siendo w_0 el sesgo.

El desarrollo matemático es absolutamente idéntico a los casos anteriores, únicamente teniendo en cuenta esta transformación no lineal sobre los datos. Adaptando la expresión 5.27 a maximizar a este tipo de máquinas nos lineales obtenemos:

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \Psi^t(\boldsymbol{x}_i) \Psi(\boldsymbol{x}_j) \tag{5.44}$$

El problema radica en que está transformación Ψ nunca está explícita, esto es, es desconocida. Este contratiempo se soluciona mediante el denominado *truco del Kernel*, que nos permite expresar el producto de dos puntos en este espacio transformado con una función kernel que sí conocemos.

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{x}_i) &= \boldsymbol{\Psi}^t(\boldsymbol{x}) \boldsymbol{\Psi}(\boldsymbol{x}_i) \\ &= \sum_{j=0}^{m_1} \Psi_j(\boldsymbol{x}) \Psi_j(\boldsymbol{x_i}) \end{aligned} \tag{5.45}$$

Quedando la función a maximizar:

$$Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{5.46}$$

Sujeto a:

$$\begin{aligned} \sum_{i=i}^{l} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{aligned} \tag{5.47}$$

Como vemos, la complejidad de este tipo de máquinas es bastante similar a las lineales. La única diferencia radica en que se reemplaza el producto escalar $(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$ en la dimensión en la que viven los datos por $K(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$ en todas las partes del algoritmo de entrenamiento.

La máquina de vectores soporte llevará a cabo una separación (en el caso de clasificación) lineal en este nuevo espacio, denominado espacio de Hilbert $\mathcal{H}$, el cual es dimensionalmente grande.

Para más información consultar [22], [7], [28] y [9].

Teorema de Mercer

Se demuestra en [22] que estas funciones kernel, las cuales representan el producto punto en un espacio de Hilbert, son todas aquellas que cumplen la denominada condición de Mercer.

Una función es kernel

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i} \Psi(\boldsymbol{x})_i \Psi(\boldsymbol{y})_i \tag{5.48}$$

si y solo si $\int f(\boldsymbol{x})^2 d\boldsymbol{x}$ es finito, entonces:

$$\int K(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x}) f(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \geq 0 \tag{5.49}$$

Muchas veces el problema radica en que demostrar que se verifica 5.49 para una determinada función f($\boldsymbol{x}$) no es una tarea sencilla.

Algunas de las funciones Kernels conocidas son:

1. Polinomial de grado p:

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \cdot \boldsymbol{y} + 1)^p \tag{5.50}$$

2. Función base radial gaussiana (RBF):

$$K(\boldsymbol{x}, \boldsymbol{y}) = e^{\frac{||\boldsymbol{x} - \boldsymbol{y}||^2}{2\sigma^2}} \tag{5.51}$$

3. Red neuronal sigmoidal de dos capas

$$K(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\kappa \boldsymbol{x} \cdot \boldsymbol{y} - \delta) \tag{5.52}$$

4. Curva de plato delgado de grado n:

$$K(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x}, \boldsymbol{y}||^{2n} \ln ||\boldsymbol{x}, \boldsymbol{y}|| \tag{5.53}$$

Existen muchos otro tipos de kernels mucho más complejos que pueden ser estudiados en la bibliografía [22], [28], [8] y [23].

Capítulo 6

Evaluación de los resultados

Este capítulo está orientado a la evaluación y comparación de los resultados obtenidos siguiendo diferentes estrategias a partir de los datos proporcionados por la Fundación Giménez Díaz y que fueron descritos en el Capítulo 2.

El capítulo presenta la siguiente estructura:

1. La presentación de los conjuntos que han sido utilizados para entrenamiento/validación y el de test sobre el cual se ha comprobado si se ha conseguido el objetivo.
2. La obvia demostración de que los gastos generados durante un año están ligados al historial de entradas clínicas recogidas durante ese año.
3. La argumentación de en base a qué establecemos las simplificaciones llevadas a cabo para, en una primera aproximación, considerar a un paciente como susceptible de generar un gasto innecesario que puede ser reducido con una temprana y adecuada atención primaria.
4. El listado de los distintos parámetros a refinar para cada SVM, y cómo se ha conseguido este refinamiento gracias a la validación cruzada en base a distintos criterios. La explicación de cuál es la función de coste que nos determinará la calidad de los procedimientos usados y el por qué es así.
5. La justificación de las distintas técnicas usadas y los resultados obtenidos con las diferentes estrategias

6.1. Conjunto de entrenamiento y de test

El total del conjunto de entradas-salidas ($\boldsymbol{\lambda} \rightarrow C$ o $\boldsymbol{\lambda} \rightarrow \mathit{Etiqueta} \in \{+1, -1\}$) lo dividimos en dos conjuntos disjuntos con la intención de utilizar uno para entrenar las diferentes técnicas utilizadas (PCA, máquina de vectores soporte...) y el otro sobre el

cual utilizar las técnicas ya entrenadas en la fase de entrenamiento para evaluar la calidad/precisión de la utilización de estas técnicas para la consecución del objetivo final.

El conjunto de entrenamiento está formado por alrededor del 80 por ciento del total de entradas existentes, y el restante 20 por ciento (aproximadamente) formará el conjunto de test. En este proyecto esta división puede ser realizada de dos maneras distintas:

1. Cogiendo el 80 por ciento de las entradas de manera aleatoria.

2. Cogiendo estrictamente las primeras entradas de cada paciente que forman el 80 por ciento del total que tiene dicho paciente.

La diferencia entre ambas técnicas es el realismo, mientras que con la primera hacemos que los algoritmos estén aprendiendo de casos que suceden en instantes posteriores a un momento dado, en el segundo las técnicas aprenden de una manera más natural, siguiendo la evolución temporal de los gastos psiquiátricos del paciente.

Es por ese motivo por el cual definitivamente optamos por la segunda táctica para llevar a cabo las pruebas.

6.2. Correlación de los costes y $\boldsymbol{\lambda}$

Aunque sea un problema trivial, antes de iniciar la tarea de predecir el coste de un paciente durante un año en base a su historial psiquiátrico durante el año anterior, comprobamos la obvia relación del vector de parámetros ($\boldsymbol{\lambda}_i$) con el coste psiquiátrico generado ese año C_i.

Gráficamente:

$$\text{Paciente i-ésimo} \begin{cases} C_1 = f(\boldsymbol{\lambda}_1) \\ C_2 = f(\boldsymbol{\lambda}_2) \\ . \\ . \\ C_{j-1} = f(\boldsymbol{\lambda}_{j-1}) \\ C_j = f(\boldsymbol{\lambda}_j) \end{cases}$$

Para realizar dicha comprobación se llevará a cabo con un "simple" regresor lineal, y medir el posterior error absoluto medio definido como:

$$Error\ Medio = \frac{1}{l}\sum_{i=1}^{l} |f(\lambda_i) - C_i| \tag{6.1}$$

Siendo l el número de muestras de test.

Sin realizar clustering ni ningún otro tipo de técnica que nos permitiera mejorar los resultados, únicamente con regresión lineal y utilizando la primera división de conjuntos de entrenamiento-test descrita en el apartado 6.1, el error medio absoluto generado es igual a 157.0205 €.

La gráfica de los costes predichos contra los costes reales es la Figura 6.1:

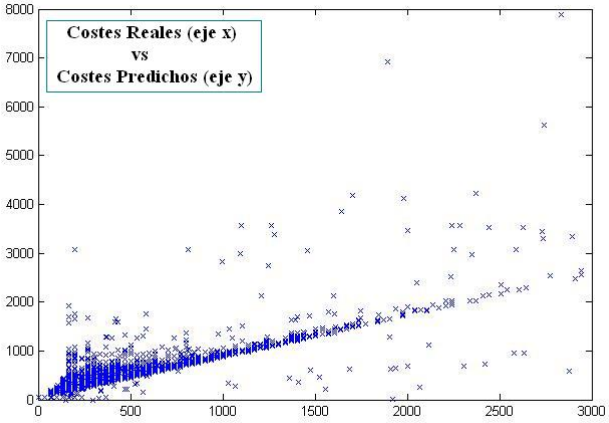


Figura 6.1: Gráfica de valores reales contra predichos

Si esta gráfica tiene forma de bisectriz como en este caso (nótese que la Figura 6.1 sería así si se normalizaran los ejes), significa que el regresor predice de una manera muy acertada el coste real. Obviamente en este problema toda la información del coste generado en un año está totalmente relacionado con el vector de parámetros de ese año ($\boldsymbol{\lambda}$), y por ese motivo la gráfica de la Figura 6.1 tiene la forma que desearíamos en el problema verdadero a resolver en este trabajo.

6.3. Referencia para la evaluación de los resultados

Según la Estrategia en Salud Mental del Sistema Nacional de Salud [5]:

"*En España no se dispone de información sobre la carga de enfermedad. En cuanto a la repercusión económica, el coste total en el año 198 fue estimado en 3.005 millones de euros. Amplios estudios, como el recogido en el Libro Blanco Estudio Socioeconómico, establecen el coste social de los trastornos de salud mental en España(1998) en 3.373,47 € por trastorno. Los costes directos representarían el 38.8 % (1.311,69 €) y los indirectos el 61 % (2.061,77 €). Dentro de los primeros, la hospitalización supone el 20.6 % (696,97 €), correspondiendo a hospitalización prolongada un 17.7 % (597,66 €) y un 3 % (98,31 €) a*

hospitalización breve; las consultas ambulatorias son un 10.4 % (352,22 €) y los dastos de farmacia el 7.8 % (263,50 €). Dentro de los costes indirectos, la invalidez representaría el 21.8 % (733,82 €), la mortalidad prematura el 21.6 % (730,12 €), la baja productividad el 9 % (303,33 €) y la incapacidad temporal el 8.7 % (294,50 €). Los trastornos mentales son la causa del 10.5 % de días perdidos por incapacidad temporal, y en torno al 6.8 % de los años de vida laboral perdidos por invalidez permanente"

Con esta información y el resto disponible en [5], debemos establecer algún tipo de criterio a partir del cual podemos considerar a un paciente como caro y por tanto, proporcionarle un tratamiento o cuidados psiquiátricos que permita por un lado tratar adecuadamente al paciente y por el otro, ahorrar dinero a la sanidad pública, ya que una temprana atención puede prevenir el que un paciente acuda de manera intensiva y frecuente en el futuro a urgencias, que es el tipo de atención clínica más caro.

Establecer dicho criterio es una tarea compleja y dependerá de cada patología, pero para una primera aproximación y con la ayuda del doctor en psiquiatría Enrique Baca García, hemos establecido el límite a partir del cual un paciente es considerado como caro en 1311.69 €, que corresponde a la media de los costes directos de un tratamiento, como se ha indicado anteriormente en el párrafo extraído del libro de la Estrategia en Salud Mental del Sistema Nacional de Salud

Habiendo detectado a los pacientes que sobrepasan este límite debido a una atención ineficiente en la mayoría de los casos y sabiendo que el coste que generarán será grande, nos centraremos en el tratamiento a proporcionarles en el instante actual para en el futuro evitar este gran coste. Este tratamiento a proporcionarles tendrá en media un coste igual, como en un caso normal, a la media del coste directo de un tratamiento e igual a 1311.69 €.

6.4. Refinamiento de parámetros de las SVM

Para cada problema de regresión/clasificación en que se emplee máquina de vectores soporte se requiere un refinamiento de hiperparámetros para obtener los mejores resultados posibles. El no refinamiento de estos hiperparámetros implica que la SVM empleará unos por defecto que probablemente no sean los óptimos y que, por consiguiente, los resultados obtenidos puedan ser mejorables.

En función de si se trata de un problema de regresión o de clasificación el número de hiperparámetros a refinar cambiará. Por ejemplo, el parámetro ϵ, como ha sido explicado anteriormente, es exclusivo de regresión.

6.4.1. Clasificación SVM

En función de si utilizamos una superficie de decisión que sea un hiperplano, esto es, que sea lineal, o que sea curva, esto es, que sea no lineal, provocará que se refinen unos parámetros o no, ya que el segundo caso implica la utilización de un kernel, el cual tiene hiperparámetros a configurar.

Para clasificación se debe reetiquetar a cada $\boldsymbol{\lambda}$ que, hasta ahora, estaba caracterizado por una salida que consistía en un coste en euros. Ahora debemos categorizar a éstos en dos grupos:

- Etiquetados como -1: aquellos $\boldsymbol{\lambda}$ cuyo coste asociado es menor que el precio standard (1311.69 €).
- Etiquetados como +1: aquellos $\boldsymbol{\lambda}$ cuyo coste asociado es mayor que el precio standard (1311.69 €).

Caso lineal

Si intentamos clasificar a las muestras mediante un hiperplano óptimo apreciamos como se trata de un problema de clasificación no linealmente separable. Por tanto, como se ha explicado en el tema de máquinas vectores soporte, la función objetivo a minimizar (en el problema primal) es:

$$\frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{6.2}$$

Como vemos, el único parámetro a refinar en este caso es C. Este parámetro regula la relación entre la complejidad de la máquina y el intento de establecer una frontera de separación con el menor error de clasificación posible.

La librería de Matlab de las SVM permite establecer un coste diferente para cada tipo de clase (+1 ó -1), de modo que podemos dar más importancia a la detección de las muestras pertenecientes a una clase que a las de otra. Esto puede ser requerido cuando el equivocarse en la clasificación de un tipo de muestras genera un coste o un gasto mayor que si nos equivocásemos en la clasificación de las muestras de la otra clase.

Para realizar el barrido estableciendo un coste para cada clase, lo cual es de interés para este trabajo, se realiza fijando el coste de una de las dos clases y realizando un barrido del otro coste, ya que lo realmente importante no son los valores que toman estos costes, sino la relación que hay entre ellos, es decir, denominando al coste de aquellos $\boldsymbol{\lambda}$ cuyo coste asociado es mayor que 1311.69 € como C^+ y a los que su coste asociado es menor que 1311.69 € como C^-, un par de valores $C^+ = 10$ $C^- = 1$ dará unos resultados parecidos al par de valores $C^+ = 1$ $C^- = 0{,}1$. Es por ese motivo por el que es suficiente con fijar un valor y hacer un barrido del otro parámetro.

Caso no lineal

En esta oportunidad, la diferencia del clasificador del caso anterior respecto a éste es la del uso de una función kernel, la cuál debe ser elegida por el usuario y que permite la transformación de los datos a un espacio de mayor dimensión en el que se conseguirá un menor error de clasificación en el conjunto de entrenamiento. La explicación extendida matemática y teórica del caso no lineal lo encontraremos en el Capítulo 5.

Utilizando la función kernel RBF:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}} \tag{6.3}$$

Además de los parámetros C^+ y C^- a refinar, tenemos un nuevo parámetro que aparece en la función kernel que debemos refinar. Este nuevo parámetro es σ y está relacionado con la anchura de las gaussianas. Indica el "radio de acción" de cada vector soporte influyendo de esta manera en la forma de la superficie de separación. Para una explicación más extensa consultar [14].

6.4.2. Regresión SVM

Al igual que en el caso de clasificación, la utilización de un regresor lineal o no lineal implicará la no o sí utilización respectivamente de un kernel que implicará el barrido de un nuevo hiperparámetro. Además en regresión hay que configurar otro nuevo parámetro exclusivo de la función de coste que utiliza.

Para regresión se mantiene para cada $\boldsymbol{\lambda}$ el coste generado durante el año siguiente. Y el modelo del regresor será:

$$\hat{C}_{i+1} = \mathbf{w}^t \boldsymbol{\lambda}_i + w_0 \tag{6.4}$$

Siendo $\hat{C}_{i+1}$ la predicción del coste para la muestra i-ésima.

Caso lineal

Recordando la función de coste ϵ-insensitiva:

Como se aprecia, este nuevo parámetro a configurar es ϵ, que penaliza exclusivamente aquellos errores de predicción que cumplan $|C_i - \hat{C}_i| > \epsilon$. En este problema ϵ tendrá unidades de euros, y significa que otorgamos al regresor cierta flexibilidad para fallar, como máximo, con un error absoluto de ϵ euros sin que estos fallos intervengan en la elaboración final del modelo de regresión de la SVM.

Además, siendo la función objetivo a minimizar:

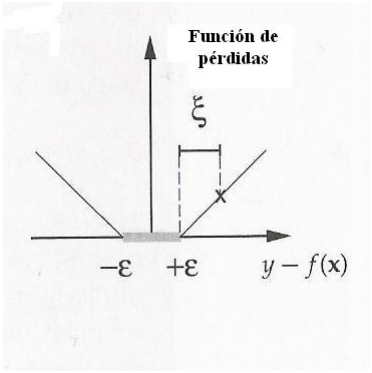


Figura 6.2: Función de perdidas ϵ-insensitiva

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{6.5}$$

Se debe refinar el parámetro C que, a igual que en clasificación, es la encargada de buscar una solución que busque un compromiso entre la complejidad de la máquina y la importancia que se le da a los errores, en este caso, de predicción.

A modo de ejemplo en este caso, para observar la importancia de cómo tener unos valores "malos" de hiperparámetros pueden afectar gravemente en la consecución del objetivo final expondré el siguiente caso: para el problema que tratamos de solucionar, en el que los costes habitualmente oscilan entre los 0 € y los 2500 €, establecer un ϵ de 3000 € y un C pequeño, que de poca importancia a los escasos casos en los que se genera un gasto mayor de 3000 €, supondría que el algoritmo convergería a una solución en lo que primordial sería minimizar la norma de $\mathbf{w}$ sujeto a unas condiciones muy poco restrictivas, provocando un regresor bastante malo si nuestra intención es predecir con bastante exactitud el coste generado.

Además, el barrido del hiperparámetro C está referenciado al de ϵ, ya que si ϵ es alto, entonces los ξ_i serán bajos, y para que en el funcional de la SVM el segundo término $(C\sum_{i=1}^{l}(\xi_i+\xi_i^*))$ siga teniendo peso,el valor de C también debe ser alto. Es por ese motivo por el cual el valor de los barridos de los parámetros C y ϵ están relacionados, si aumenta uno el otro también y si disminuye uno el otro también.

Caso no lineal

Si utilizamos una función kernel RBF, el número de parámetros a refinar será C, σ y ϵ. Teniendo C y ϵ el significado explicado en el punto anterior y σ, al igual que en el caso

de clasificación, afecta al "radio de acción" de los vectores soporte y por tanto a la forma de la función de regresión. Una explicación más extensa puede ser encontrada en [14].

6.5. Validación cruzada

Para llevar a cabo este refinamiento, el barrido de los hiperparámetros se debe evaluar sobre un conjunto de muestras del problema. Esto ha sido realizado de dos maneras:

1. Realizando validación cruzada sobre el conjunto de entrenamiento y eligiendo aquellos conjuntos de hiperparámetros que nos proporcionan el mayor ahorro posible tanto en el caso de clasificación como en el de regresión para este conjunto de muestras, es decir, aquellos hiperparámetros que dan los mejores resultados utilizando la función de coste que se ve en la sección 6.6.

2. Realizando validación cruzada sobre el conjunto de entrenamiento y eligiendo aquellos conjuntos de hiperparámetros que dan lugar a las mayores tasas de acierto en el caso de clasificación, o al menor error cuadrático medio en el caso de regresión.

Para la primera de las validaciones se realizará una serie de barridos de hiperparámetros sobre el funcional de coste explicado en el Capítulo 5 y evaluando posteriormente estas SVM generadas sobre la función de coste 6.1 (sección 6.6).

Para conjunto de muestras de un tamaño moderado, como en el problema que estamos tratando, el procedimiento que adoptamos es la validación cruzada. En su forma más elemental, la validación cruzada consiste en dividir el conjunto de datos de entrenamiento en m subconjuntos. Cada subconjunto de los m se utiliza para evaluar la máquina con sus correspondiente hiperparámetros entrenada sobre los $(m-1)$ subconjuntos restantes. Finalmente el error calculado es el promediado de los errores obtenidos tras evaluar sobre cada uno de los m subconjuntos las máquinas entrenadas en los $(m-1)$ subconjuntos restantes.

Una importante dificultad con el uso de la validación cruzada radica en que si se usa con algoritmos intensivos computacionalmente, la repetición de los m ciclos de aprendizaje puede requerir demasiado esfuerzo computacional.

En la Figura 6.3 podemos ver gráficamente la validación cruzada con m igual a 4 que ha sido empleada para el refinamiento de los hiperparámetros de las distintas SVM usadas.

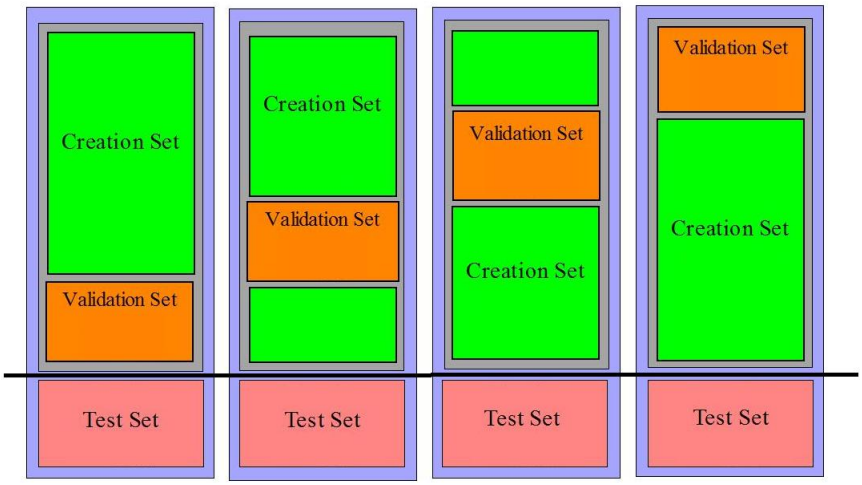


Figura 6.3: Validación cruzada con m igual a 4

Información sobre la validación cruzada en [1].

6.6. Evaluación de los resultados (medida de calidad)

Según lo visto en la sección 6.3, establecer un criterio de calidad de resultados en base al porcentaje de aciertos en el caso de clasificación o en base al error cuadrático medio en regresión seguramente no sea la opción más adecuada. Fijándonos de nuevo en el objetivo que perseguimos nos percatamos de que:

- En clasificación nos interesa más detectar los pacientes que generan un coste alto que en general tener un porcentaje de acierto bastante alto, pero no demasiado bueno detectando los paciente "caros".

- En regresión preferimos tener un regresor que aunque no prediga con muy buena exactitud el coste real que generará el paciente al año siguiente, pero que sí prediga un coste que determine con gran acierto si nos encontramos ante un paciente "caro" o "barato".

La función de coste será la siguiente:

Coste estimado / Coste real	Barato	Caro
Barato	0	C_i - Coste de referencia
Caro	0	C_i - Coste de referencia

Cuadro 6.1: Función de coste

El coste de referencia es el descrito anteriormente (sección 6.3) e igual a 1311.69 €.

Los casos son los siguientes:

1. Coste real barato - Coste predicho barato: si nos encontramos ante esta situación no hay gasto innecesario ni ahorro, ya que estamos ante un caso en el que no intervenimos en el tratamiento natural del paciente, y por tanto, como se ha dicho, no tendremos la posibilidad de ahorrar ni de gastar de más.

2. Coste real barato - Coste predicho caro: si un paciente que va a generar un gasto pequeño le detectamos como lo contrario, como que su gasto va a ser grande, le aplicaremos un tratamiento standard de coste 1311.69 €, mayor que lo que en realidad va a ser y, por tanto, produciremos un gasto innecesario igual a C_i - precio standard (nótese que si esta resta, como en este caso, da igual a un número negativo significa que se produce un gasto innecesario).

3. Coste real caro - Coste predicho barato: si lo detectamos como barato no intervendremos en el tratamiento que recibirá este paciente, y por tanto no habrá un gasto innecesario ni ahorro. De todas maneras, si lo detectásemos como caro sí podríamos ahorrar una cantidad igual a C_i - precio standard (cifra positiva). Es un caso en el que podríamos ahorrar dinero a la sanidad pública pero no lo hacemos.

4. Coste real caro - Coste predicho caro: detectamos correctamente a un paciente caro, con lo que el ahorro que experimentamos es igual a C_i - precio standard (cifra positiva).

6.7. Visualización de los mapas auto-organizados

Entrenando esta técnica con el conjunto de $\boldsymbol{\lambda}$ y sus C asociados, y usando una máscara que haga que los mapas se organicen en base a los $\boldsymbol{\lambda}$ y no de sus costes, los mapas resultantes son los mostrados en la Figura 6.4.

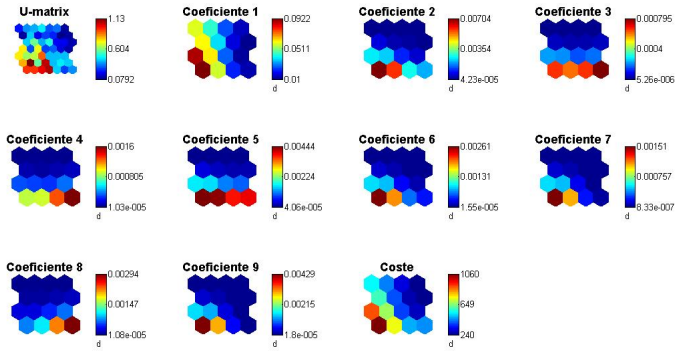


Figura 6.4: Visualización de los coeficientes y de su coste asociado

Como vemos, se justifica la hipótesis que asumimos al principio del proyecto, el coste de un año está íntimamente relacionado con el historial clínico del paciente durante los 365 días anteriores. Esto se puede ver porque, debido a la máscara utilizada, el mapa se autoorganiza en base a los $\boldsymbol{\lambda}$, y el mapa del coste se organiza en base al mapa resultante de sus $\boldsymbol{\lambda}$ asociados. De esta forma, se aprecia que la organización de los costes sigue una estructura muy lógica, teniendo cada celda del mapa de costes vecinos con costes bastantes parecidos entre sí.

6.8. Justificación de las técnicas usadas

En esta sección se justificará experimentalmente el uso de todas y cada una de las técnicas empleadas en el proyecto. Para ello se compararán estadísticos o datos que sirven para demostrar la valía de la introducción de una determinada técnica.

6.8.1. K-means

El uso de un algoritmo de clustering puede interesar para agrupar vectores de entrada ($\boldsymbol{\lambda}$) que presentan características comunes. Si los distintos clusters de $\boldsymbol{\lambda}$ presentan en media costes asociados suficientemente diferentes entre sí, significa que esta primera etapa nos permite agrupar pacientes que en general se comportan de una manera diferente a la de los otros clusters y, de esta manera, iniciar un estudio de cada cluster de manera independiente que debiera ser más preciso que un "tratamiento" global.

Usando el K-means con K=3 obtenemos 3 clusters con las características mostradas en el Cuadro 6.2, y la distribución de costes en cada cluster según la Figura 6.5.

	1º Cluster	2º Cluster	3º Cluster
Número de muestras	10202	3890	454
Coste medio	274.0734 €	603.46 €	1608.5 €
Desviación standard	1207.3 €	1300.6 €	1630.8 €
Coste máximo	65828 €	57433 €	15333 €
Coste mínimo	0 €	0 €	0 €
Porcentaje de pacientes caros	1.8 %	7 %	54 %

Cuadro 6.2: Información de los clusters

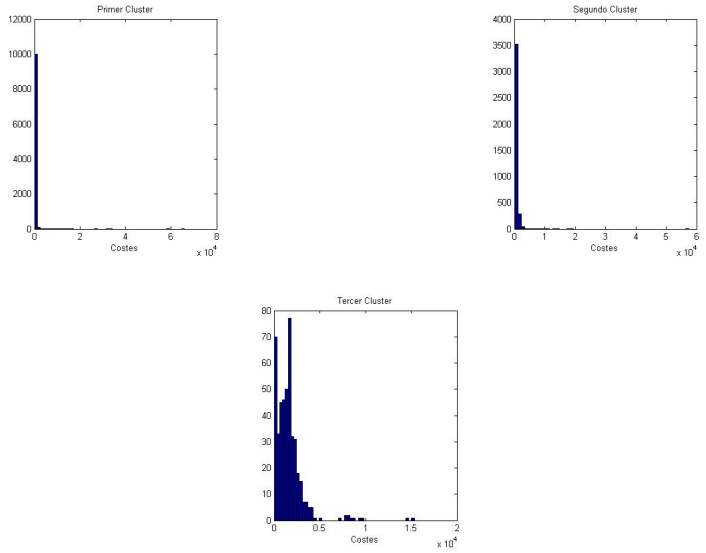


Figura 6.5: Histograma de los costes de los distintos clusters para el conjunto de entrenamiento

A la vista de los resultados en la Tabla 6.2, tenemos 3 clusters que pueden ser catalogados, debido a sus costes medios, como caro, intermedio y barato. Además y con la ayuda de la función de coste a utilizar vista en la sección 6.6, vistos sus costes medios podemos extraer las siguientes conclusiones:

1. El cluster caro es en donde, en principio y con el correcto funcionamiento de las etapas posteriores, deberíamos conseguir la mayor cantidad de ahorro por paciente.

2. En el cluster intermedio si bien es cierto que la mayoría serán detectados como baratos, habrá un número importante de pacientes sobre los que sí podríamos ahorrar.

3. Siendo el coste medio del cluster barato pequeño en comparación con el precio a

partir del cual los pacientes son catalogados como caros, queda claro que la cantidad posible a ahorrar o lo que es lo mismo, el número de pacientes susceptibles de ser caros, será muy pequeño.

Si no se hubiera realizado clustering la tabla de estadísticas sería la mostrada en el Cuadro 6.3.

	Datos
Número de muestras	14546
Coste medio	403.81 €
Desviación standard	1274.8 €

Cuadro 6.3: Información de los datos sin clustering

Para justificar el uso del algoritmo de clustering se realizará una prueba para comparar los resultados. En la sección 6.9, se verá como el ahorro medio conseguido con regresión lineal es igual a 104.85 €. Realizando la misma regresión sin distinciones de clusters, es decir, tratando al conjunto de datos como un solo cluster, el ahorro medio obtenido es igual a 83.04 €.

6.8.2. Análisis de componentes principales

El hecho de aplicar PCA está justificado porque, como se explicó en el Capítulo 4, nos interesa normalizar la varianza de los datos para conocer a priori la distribución de los datos y así realizar un barrido de valores de forma más controlada de los hiperparámetros de las SVM.

Esta técnica se aplica cluster por cluster porque cada grupo de muestras tendrá una distribución que, en principio, debería diferir bastante respecto a las otras. Si se tratase todos los datos de una manera conjunta, obtendríamos unos autovalores y autovectores que representarían con menos fidelidad la distribución de los datos de cada cluster que si lo hiciéramos cluster por cluster.

En la Figura 6.6 se muestra un ejemplo de cómo es notable la diferencia de los autovectores de cada cluster y, por tanto, es justificable el uso de PCA de manera independiente en cada cluster.

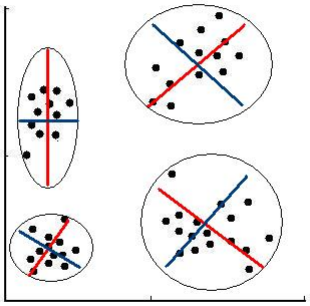


Figura 6.6: En rojo y azul los autovectores correspondientes a cada cluster

6.8.3. Regresión y clasificación

La regresión ha sido utilizada con la intención de predecir el coste generado por el paciente durante el año siguiente.

El objetivo de la clasificación es indicarnos si un determinado paciente es susceptible de ser considerado como caro por el gasto que podría generar el próximo año.

Sea como fuere, el objetivo de ambas técnicas es el mismo, detectar paciente sobre los que se podría actuar para proporcionales una temprana y adecuada atención primaria provocando, por otra parte, el ahorro de los gastos en tratamientos no productivos. Los resultados de ambos procedimientos están en la sección 6.9.

6.9. Diferentes estrategias seguidas

Volviendo al Capítulo 1, allí se muestra un esquema general del procedimiento seguido (Figura 1.3). Como se aprecia en dicho esquema, la parte de clustering más PCA es común a todas las diferentes estrategias seguidas, lo único diferente es la elección del camino de clasificación o de regresión para conseguir el objetivo final y de qué método en concreto.

Las muestras de test han sido asignadas al cluster cuya distancia a su centroide hace mínima la distancia cuadrática definida como:

$$Distancia = \sum_{i=1}^{9} (x_i - centroide_i)^2 \tag{6.6}$$

donde i es el indicador de cada una de las dimensiones de los datos.

Siendo los clusters definitivos sobre los que se ha evaluado las diferentes estrategias son los descritos en el Cuadro 6.4, sobre los que se ha aplicado a cada uno su PCA correspondiente obtenido en la etapa de entrenamiento.

	1º Cluster	2º Cluster	3º Cluster
Número de muestras	2547	858	139
Coste medio	356.3455 €	967.2534 €	3564.4 €
Desviación standard	1152.9 €	1633 €	6818.1 €
Coste máximo	23877 €	22602 €	74246 €
Coste mínimo	0 €	0 €	33.44 €
Porcentaje de pacientes caros	3.8 %	13.64 %	75.54 %

Cuadro 6.4: Información de los clusters para el conjunto de test

Y siendo los histogramas de los costes de cada cluster los mostrados en la Figura 6.7.

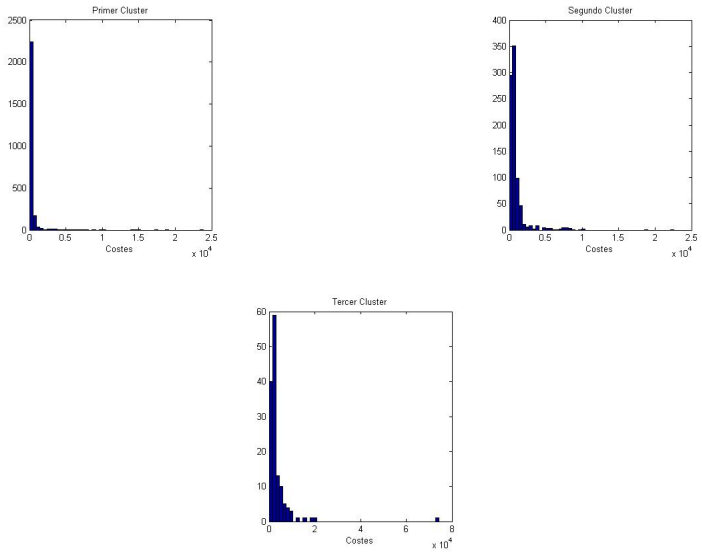


Figura 6.7: Histograma de los costes de los distintos clusters para el conjunto de test

A continuación se procederá a mostrar los resultados obtenidos con las diferentes estrategias.

6.9.1. Clasificación SVM lineal

Usando hiperparámetros según la tasa de acierto

En primer lugar se mostrarán los resultados conseguidos con los hiperparámetros obtenidos en fase de entrenamiento/validación que maximizan la tasa de acierto para cada cluster.

Recordando la función de coste final mostrada en el Cuadro 6.1 (sección 6.6), se aprecia como cada predicción puede afectar de 3 posibles maneras en el aspecto económico: provocando o un gasto innecesario o bien un ahorro, o bien no alterando las arcas monetarias.

Habiendo recordado ésto, la distribución de los porcentajes de cada caso en cada cluster fijándonos solamente en el aspecto monetario utilizando SVM lineal para clasificación es la siguiente:

Figura 6.8: Resultado de la clasificación de los pacientes del cluster "barato" desde un punto de vista económico

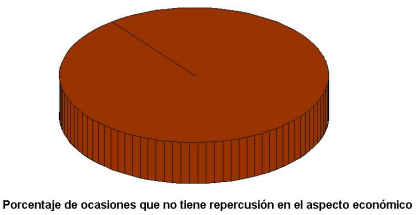


Figura 6.9: Resultado de la clasificación de los pacientes del cluster "intermedio" desde un punto de vista económico

Figura 6.10: Resultado de la clasificación de los pacientes del cluster "caro" desde un punto de vista económico

A continuación se muestran los porcentajes de ocasiones en los que el clasificador ha predicho que se trata de un caso barato o caro en cada cluster.

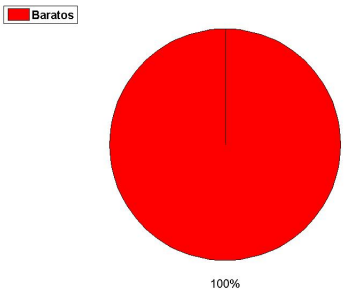


Figura 6.11: Porcentaje de las predicciones del clasificador en el cluster "barato"

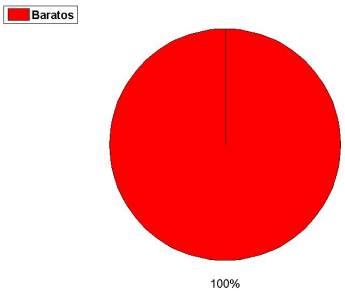


Figura 6.12: Porcentaje de las predicciones del clasificador en el cluster "intermedio"

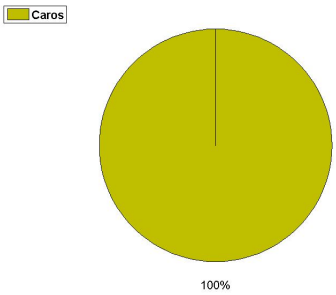


Figura 6.13: Porcentaje de las predicciones del clasificador en el cluster “caro”

Provocando una tasa de acierto en cada cluster igual a:

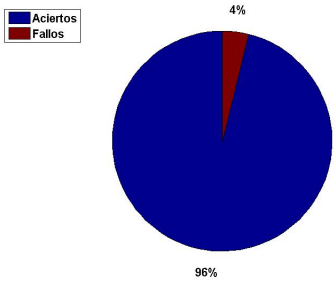


Figura 6.14: Tasa de acierto del clasificador en el cluster “barato”

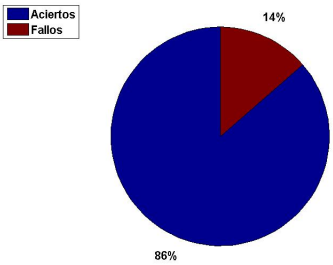


Figura 6.15: Tasa de acierto del clasificador en el cluster “intermedio”

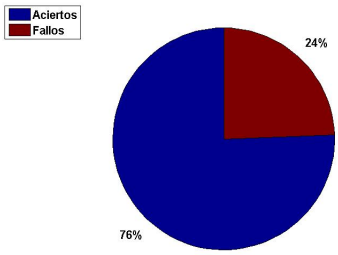


Figura 6.16: Tasa de acierto del clasificador en el cluster “caro”

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	0	0
Segundo cluster	0	0
Tercer cluster	313150.28 €	2252.88 €
Total	313150.28 €	88.36 €

Cuadro 6.5: Tabla de información monetaria conseguida con una SVM lineal de clasificación y usando los hiperparámetros que dan lugar a las mayores tasas de acierto

A la vista de los resultados la conclusión es clara, el ahorro conseguido es exclusivamente debido al algoritmo de clustering.

Usando hiperparámetros según la función de coste 6.1 (sección 6.6)

En este apartado se muestran los resultados obtenidos con los hiperparámetros que en la etapa de entrenamiento/validación daban los mejores resultados usando validación cruzada con la función de coste 6.1 (sección 6.6).

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

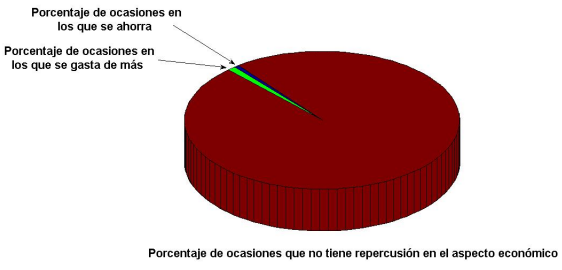


Figura 6.17: Resultado de la clasificación de los pacientes del cluster "barato" desde un punto de vista económico

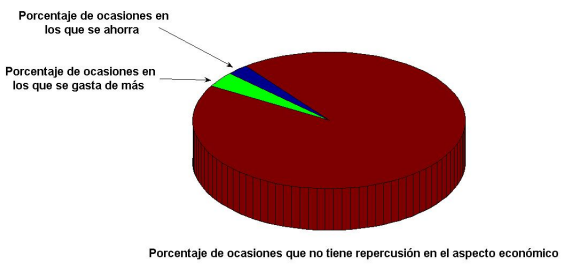


Figura 6.18: Resultado de la clasificación de los pacientes del cluster "intermedio" desde un punto de vista económico

Figura 6.19: Resultado de la clasificación de los pacientes del cluster "caro" desde un punto de vista económico

A continuación se muestran los porcentajes de ocasiones en los que el clasificador ha predicho que se trata de un caso barato o caro en cada cluster.

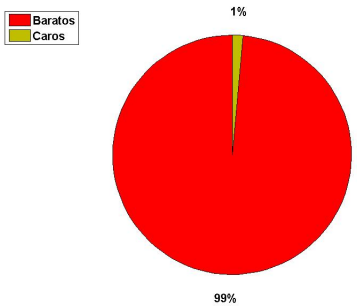


Figura 6.20: Porcentaje de las predicciones del clasificador en el cluster "barato"

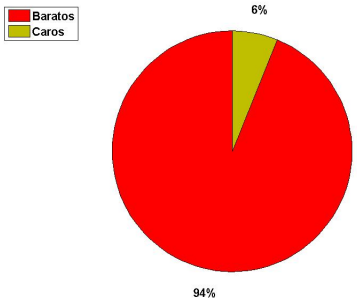


Figura 6.21: Porcentaje de las predicciones del clasificador en el cluster "intermedio"

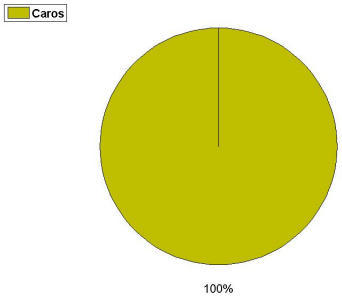


Figura 6.22: Porcentaje de las predicciones del clasificador en el cluster “caro”

Provocando una tasa de acierto en cada cluster igual a:

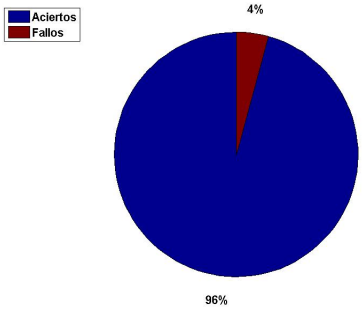


Figura 6.23: Tasa de acierto del clasificador en el cluster “barato”

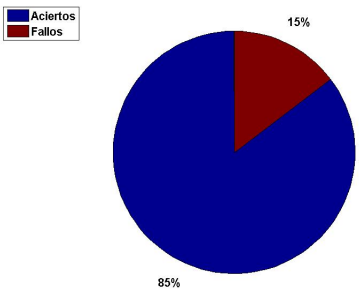


Figura 6.24: Tasa de acierto del clasificador en el cluster “intermedio”

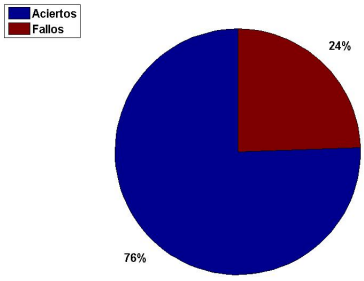


Figura 6.25: Tasa de acierto del clasificador en el cluster “caro”

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	34315.8 €	13.47 €
Segundo cluster	51002.7 €	59.44 €
Tercer cluster	313150.28 €	2252.88 €
Total	398468.78 €	112.43 €

Cuadro 6.6: Tabla de información monetaria conseguida con una SVM lineal de clasificación y usando los hiperparámetros óptimos para la función de coste 6.1 (sección 6.6)

La mejora respecto al caso anterior se produce únicamente en los clusters barato e intermedio. Es en éstos en donde ahora sí el algoritmo de clasificación interviene produciendo una ostensible mejora en el ahorro medio por paciente. En el cluster caro se sigue ahorrando exclusivamente gracias a la correcta detección de pacientes caros por parte del algoritmo de clustering.

6.9.2. Clasificación SVM RBF

Usando hiperparámetros según la tasa de acierto

En este apartado se muestran los resultados de utilizar una SVM RBF para clasificación con los parámetros que maximizan la tasa de acierto en cada cluster en la etapa de entrenamiento/validación.

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

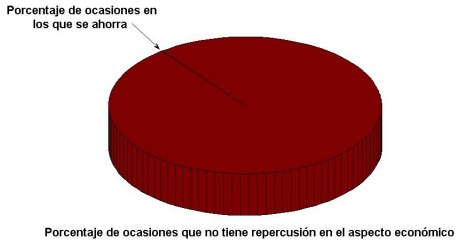


Figura 6.26: Resultado de la clasificación de los pacientes del cluster "barato" desde un punto de vista económico

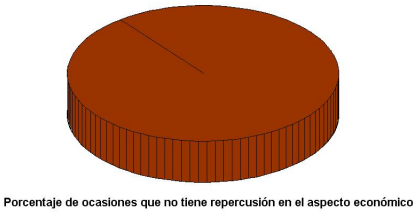


Figura 6.27: Resultado de la clasificación de los pacientes del cluster "intermedio" desde un punto de vista económico

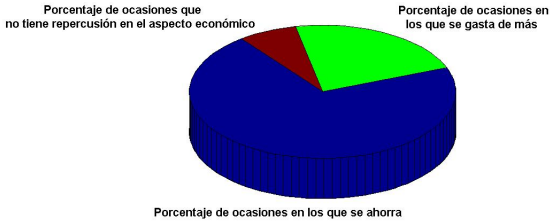


Figura 6.28: Resultado de la clasificación de los pacientes del cluster "caro" desde un punto de vista económico

A continuación se muestran los porcentajes de ocasiones en los que el clasificador ha predicho que se trata de un caso barato o caro en cada cluster.

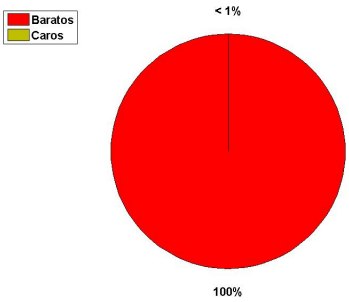


Figura 6.29: Porcentaje de las predicciones del clasificador en el cluster “barato”

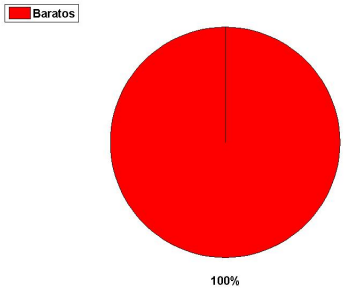


Figura 6.30: Porcentaje de las predicciones del clasificador en el cluster “intermedio”

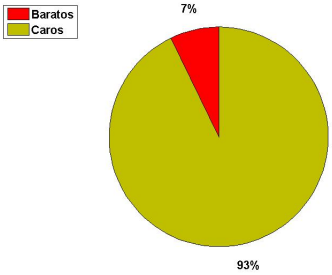


Figura 6.31: Porcentaje de las predicciones del clasificador en el cluster “caro”

Provocando una tasa de acierto en cada cluster igual a:

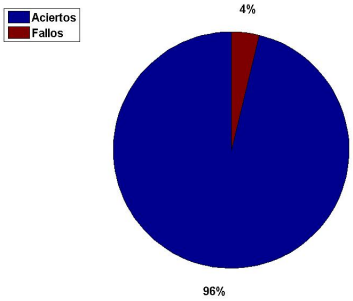


Figura 6.32: Tasa de acierto del clasificador en el cluster “barato”

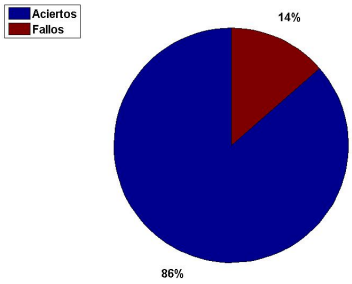


Figura 6.33: Tasa de acierto del clasificador en el cluster “intermedio”

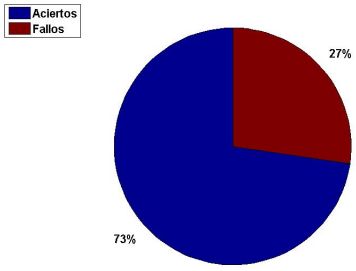


Figura 6.34: Tasa de acierto del clasificador en el cluster “caro”

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	2035.45 €	0.79 €
Segundo cluster	0 €	0 €
Tercer cluster	307536.83 €	2212.49 €
Total	309572.28 €	87.35 €

Cuadro 6.7: Tabla de información monetaria conseguida con una SVM RBF de clasificación y usando los hiperparámetros que dan lugar a las mayores tasas de acierto

En ocasiones como ésta el clasificador establece una frontera compleja que da lugar a mayores tasas de acierto en la etapa de entrenamiento que en el caso de clasificación SVM lineal, pero con peores resultados de acierto en la etapa de test debido a que se trata de un modelo sobreajustado. Además, en este caso la peor tasa de acierto se ha traducido a un menor ahorro medio por paciente que el conseguido con el clasificador SVM lineal con los hiperparámetros que intentan maximizar la tasa de acierto.

Usando hiperparámetros según la función de coste 6.1 (sección 6.6)

En este apartado se muestran los resultados obtenidos con los hiperparámetros que en la etapa de entrenamiento/validación daban los mejores resultados usando validación cruzada con la función de coste 6.1 (sección 6.6).

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

Figura 6.35: Resultado de la clasificación de los pacientes del cluster "barato" desde un punto de vista económico

Figura 6.36: Resultado de la clasificación de los pacientes del cluster "intermedio" desde un punto de vista económico

Figura 6.37: Resultado de la clasificación de los pacientes del cluster "caro" desde un punto de vista económico

A continuación se muestran los porcentajes de ocasiones en los que el clasificador ha predicho que se trata de un caso barato o caro en cada cluster.

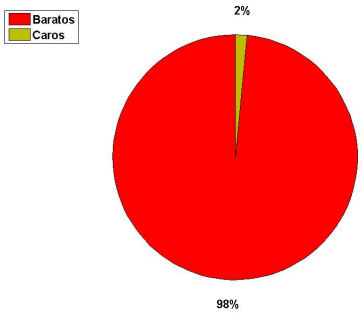


Figura 6.38: Porcentaje de las predicciones del clasificador en el cluster “barato”

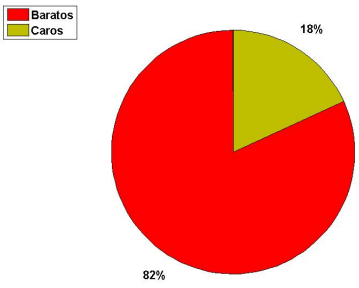


Figura 6.39: Porcentaje de las predicciones del clasificador en el cluster “intermedio”

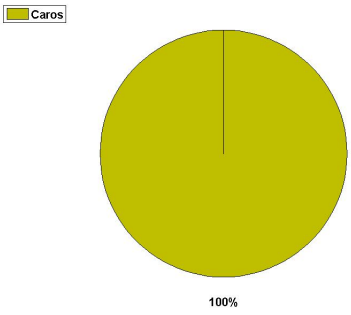


Figura 6.40: Porcentaje de las predicciones del clasificador en el cluster “caro”

Provocando una tasa de acierto en cada cluster igual a:

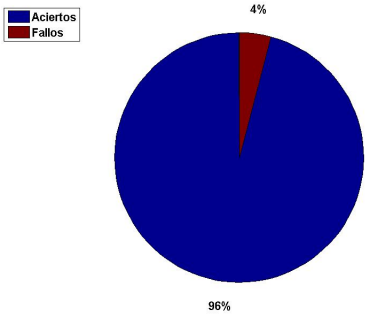


Figura 6.41: Tasa de acierto del clasificador en el cluster "barato"

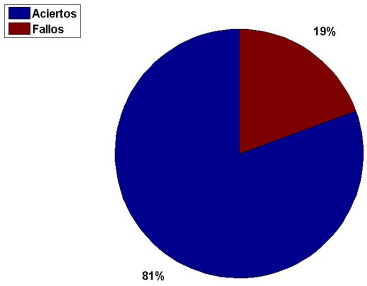


Figura 6.42: Tasa de acierto del clasificador en el cluster "intermedio"

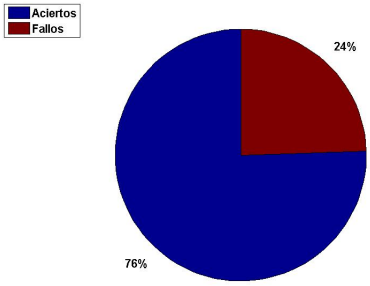


Figura 6.43: Tasa de acierto del clasificador en el cluster "caro"

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	39206.33 €	15.39 €
Segundo cluster	74591.17 €	86.93 €
Tercer cluster	313150.28 €	2252.87 €
Total	426947.78 €	120.47 €

Cuadro 6.8: Tabla de información monetaria conseguida con una SVM RBF de clasificación y usando los hiperparámetros óptimos para la función de coste 6.1 (sección 6.6)

El gran aumento en el ahorro medio por paciente se debe a la mejora conseguida en la detección de pacientes caros en los clusters barato y especialmente en el intermedio. Al

igual que con el clasificador SVM lineal, el ahorro conseguido en el cluster caro se debe exclusivamente a la buena detección por parte del k-means de los pacientes que generan costes grandes.

6.9.3. Regresión lineal

En este apartado se mostrarán los resultados tras aplicar un regresor lineal que minimiza la función de coste cuadrática.

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

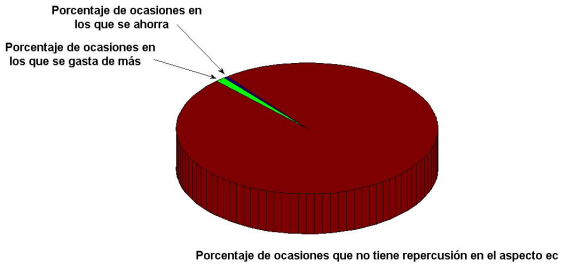


Figura 6.44: Resultado de la regresión lineal sobre los pacientes del cluster "barato" desde un punto de vista económico

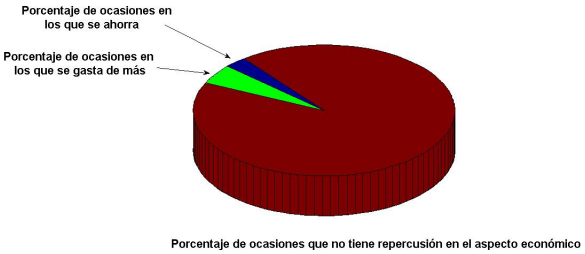


Figura 6.45: Resultado de la regresión lineal sobre los pacientes del cluster "intermedio" desde un punto de vista económico

Figura 6.46: Resultado de la regresión lineal sobre los pacientes del cluster "caro" desde un punto de vista económico

Siendo las gráficas de los valores predichos contra los reales en los diferentes cluster las siguientes:

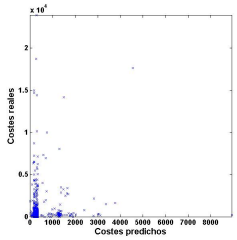


Figura 6.47: Gráfica de valores predichos contra reales en el cluster “barato”

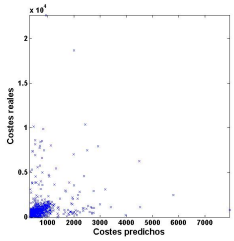


Figura 6.48: Gráfica de valores predichos contra reales en el cluster “intermedio”

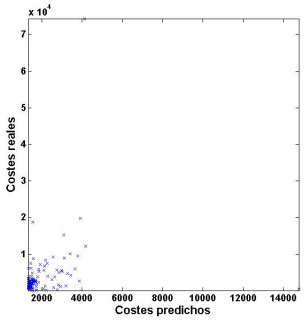


Figura 6.49: Gráfica de valores predichos contra reales en el cluster “caro”

Y los errores absolutos de predicción de cada una de las muestras de cada cluster:

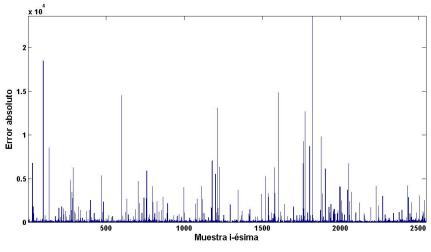


Figura 6.50: Gráfica de errores absolutos de cada muestra del cluster “barato”

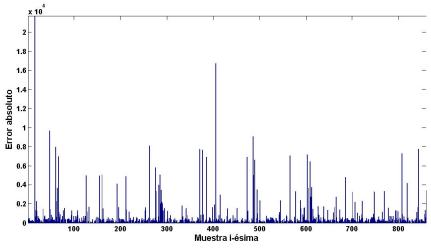


Figura 6.51: Gráfica de errores absolutos de cada muestra del cluster “intermedio”

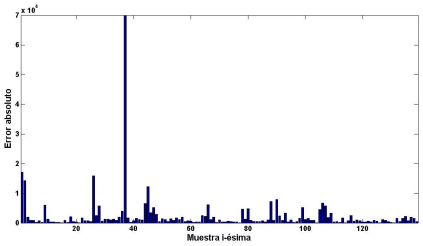


Figura 6.52: Gráfica de errores absolutos de cada muestra del cluster “caro”

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	10467 €	4.1 €
Segundo cluster	48033 €	55.98 €
Tercer cluster	313122 €	2252.67 €
Total	371622 €	104.85 €

Cuadro 6.9: Tabla de información monetaria conseguida con una regresión lineal

Como vemos, una “simple” regresión lineal da unos resultados bastante aceptables en los clusters caro e intermedio.

6.9.4. Regresión SVM lineal

Usando hiperparámetros según el error cuadrático medio

Usándose una máquina de regresión con los hiperparámetros óptimos según el error cuadrático medio en los barridos realizados en la etapa de entrenamiento/validación, los resultados son los mostrados a continuación.

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

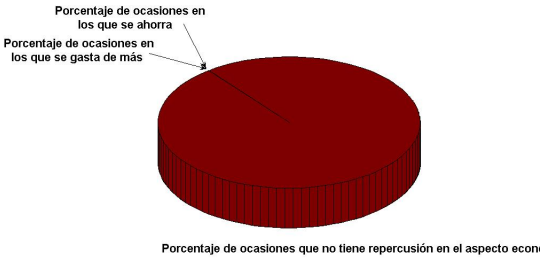


Figura 6.53: Resultado de la regresión SVM lineal sobre los pacientes del cluster "barato" desde un punto de vista económico

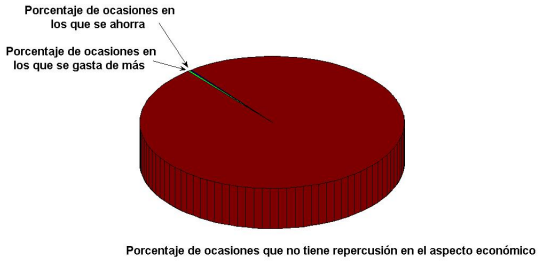


Figura 6.54: Resultado de la regresión SVM lineal sobre los pacientes del cluster "intermedio" desde un punto de vista económico

Figura 6.55: Resultado de la regresión SVM lineal sobre los pacientes del cluster "caro" desde un punto de vista económico

Las gráficas de los valores predichos contra los reales en los diferentes cluster son:

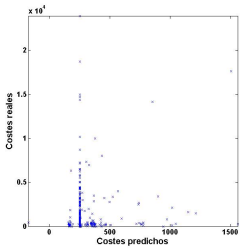


Figura 6.56: Gráfica de valores predichos contra reales en el cluster "barato"

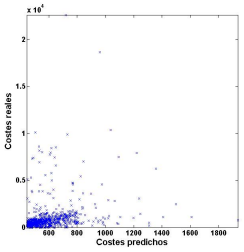


Figura 6.57: Gráfica de valores predichos contra reales en el cluster "intermedio"

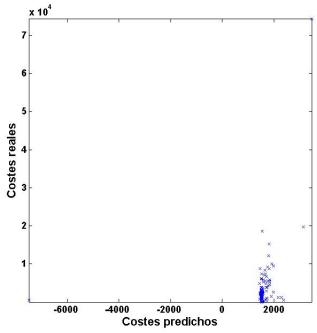


Figura 6.58: Gráfica de valores predichos contra reales en el cluster “caro”

Y los errores absolutos de predicción de cada una de las muestras de cada cluster:

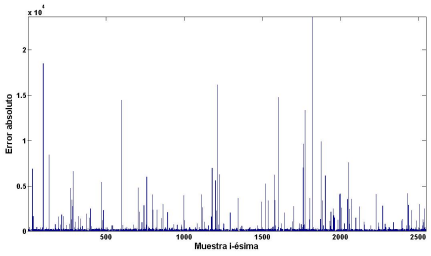


Figura 6.59: Gráfica de errores absolutos de cada muestra del cluster “barato”

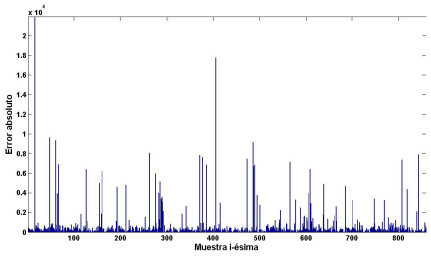


Figura 6.60: Gráfica de errores absolutos de cada muestra del cluster “intermedio”

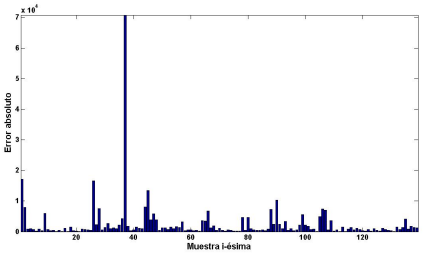


Figura 6.61: Gráfica de errores absolutos de cada muestra del cluster “caro”

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	15261.71 €	5.99 €
Segundo cluster	5121.90 €	5.96 €
Tercer cluster	313966 €	2258.74 €
Total	334349.61 €	94.34 €

Cuadro 6.10: Tabla de información monetaria conseguida con una SVM lineal de regresión y usando los hiperparámetros que minimizan el error cuadrático medio

Resultando unos resultados peores que los obtenidos con la regresión lineal que minimiza la función de coste cuadrática. Hay que percatarse que hay un conjunto de hiperparámetros de la SVM de regresión lineal que da lugar al mismo resultado obtenido con la regresión lineal del apartado anterior. En este caso el resultado no es el mismo porque se han elegido aquellos hiperparámetros que han dado el menor error cuadrático de la batería de barridos realizados, y ha resultado que dicho conjunto de hiperparámetros no equivale a la solución del regresor lineal.

Usando hiperparámetros según la función de coste 6.1 (sección 6.6)

A continuación la máquina de regresión con los hiperparámetros óptimos según la función de coste 6.1 (sección 6.6) en la etapa de entrenamiento/validación.

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

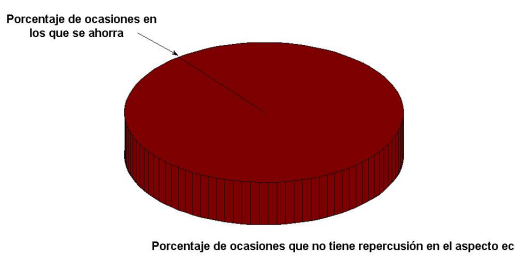


Figura 6.62: Resultado de la regresión SVM lineal sobre los pacientes del cluster "barato" desde un punto de vista económico

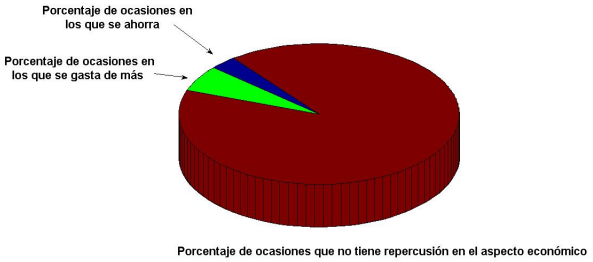


Figura 6.63: Resultado de la regresión SVM lineal sobre los pacientes del cluster "intermedio" desde un punto de vista económico

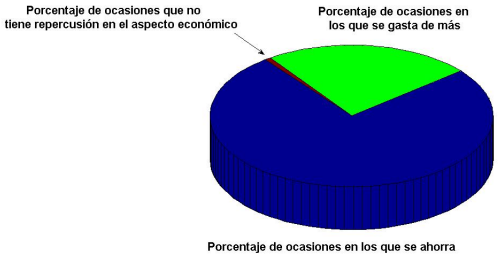


Figura 6.64: Resultado de la regresión SVM lineal sobre los pacientes del cluster "caro" desde un punto de vista económico

Las gráficas de los valores predichos contra los reales en los diferentes cluster son:

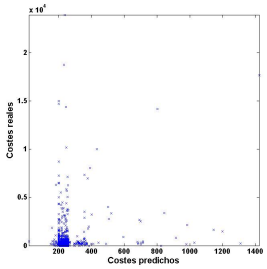


Figura 6.65: Gráfica de valores predichos contra reales en el cluster "barato"

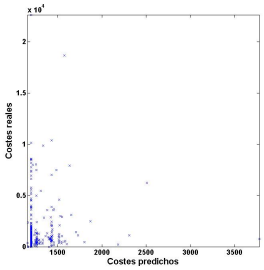


Figura 6.66: Gráfica de valores predichos contra reales en el cluster "intermedio"

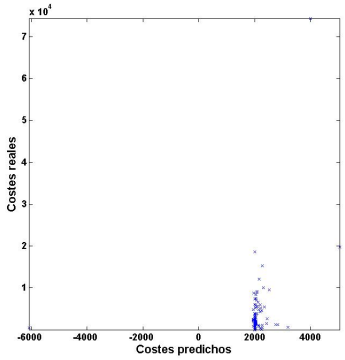


Figura 6.67: Gráfica de valores predichos contra reales en el cluster "caro"

Y los errores absolutos de predicción de cada una de las muestras de cada cluster:

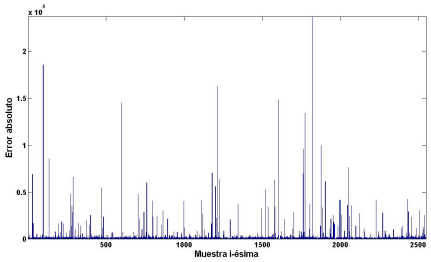


Figura 6.68: Gráfica de errores absolutos de cada muestra del cluster "barato"

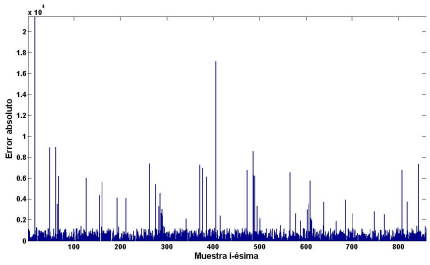


Figura 6.69: Gráfica de errores absolutos de cada muestra del cluster "intermedio"

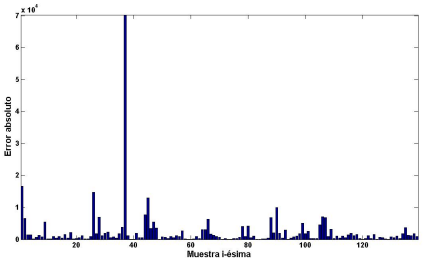


Figura 6.70: Gráfica de errores absolutos de cada muestra del cluster "caro"

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	16339.32 €	6.41 €
Segundo cluster	52034.44 €	60.64 €
Tercer cluster	313966 €	2258.74 €
Total	382339.76 €	107.88 €

Cuadro 6.11: Tabla de información monetaria conseguida con una SVM lineal de regresión y usando los hiperparámetros óptimo para la fundión de coste 6.1 (sección 6.6)

Consiguiendo un ahorro medio por paciente y año que ahora sí mejora el resultado obtenido con el regresor lineal.

6.9.5. Regresión SVM RBF

Usando hiperparámetros según el error cuadrático medio

Usándose una máquina de regresión con los hiperparámetros óptimos según el error cuadrático medio en los barridos realizados en la etapa de entrenamiento/validación, los resultados son los mostrados a continuación.

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

Figura 6.71: Resultado de la regresión SVM RBF sobre los pacientes del cluster "barato" desde un punto de vista económico

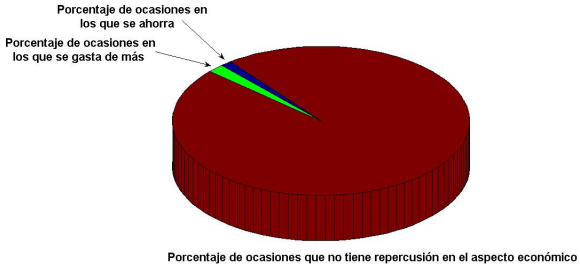


Figura 6.72: Resultado de la regresión SVM RBF sobre los pacientes del cluster "intermedio" desde un punto de vista económico

Figura 6.73: Resultado de la regresión SVM RBF sobre los pacientes del cluster "caro" desde un punto de vista económico

Las gráficas de los valores predichos contra los reales en los diferentes cluster son:

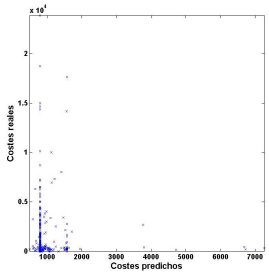


Figura 6.74: Gráfica de valores predichos contra reales en el cluster “barato”

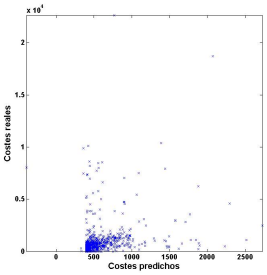


Figura 6.75: Gráfica de valores predichos contra reales en el cluster “intermedio”

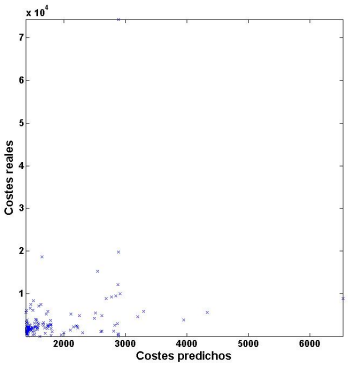


Figura 6.76: Gráfica de valores predichos contra reales en el cluster “caro”

Y los errores absolutos de predicción de cada una de las muestras de cada cluster:

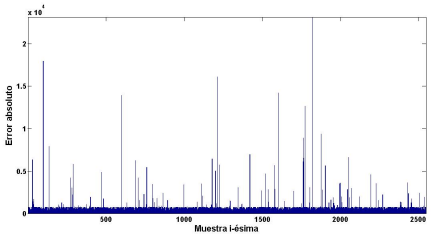


Figura 6.77: Gráfica de errores absolutos de cada muestra del cluster “barato”

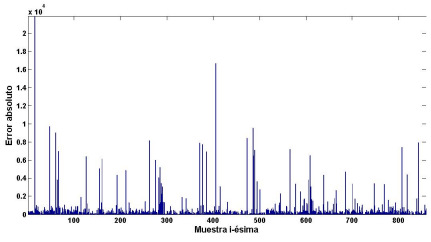


Figura 6.78: Gráfica de errores absolutos de cada muestra del cluster “intermedio”

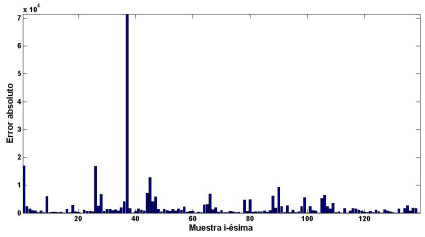


Figura 6.79: Gráfica de errores absolutos de cada muestra del cluster “caro”

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	16040.22 €	6.29 €
Segundo cluster	39979.92 €	46.59 €
Tercer cluster	313122.47 €	2252.67 €
Total	369142.61 €	104.15 €

Cuadro 6.12: Tabla de información monetaria conseguida con una SVM RBF de regresión y usando los hiperparámetros que minimizan el error cuadrático medio

A la vista de los resultados, podemos decir en este caso que el regresor, que en realidad busca minimizar el error cuadrático medio, da unos resultados bastante óptimos en materia de ahorro económico.

Usando hiperparámetros según la función de coste 6.1 (sección 6.6)

A continuación la máquina de regresión con los hiperparámetros óptimos según la función de coste 6.1 (sección 6.6) en la etapa de entrenamiento/validación.

La distribución de porcentajes en cada cluster de los distintos casos referidas al aspecto monetario son los siguientes:

Figura 6.80: Resultado de la regresión SVM RBF sobre los pacientes del cluster "barato" desde un punto de vista económico

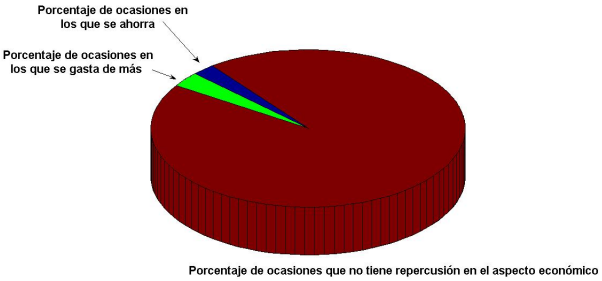


Figura 6.81: Resultado de la regresión SVM RBF sobre los pacientes del cluster "intermedio" desde un punto de vista económico

Figura 6.82: Resultado de la regresión SVM RBF sobre los pacientes del cluster "caro" desde un punto de vista económico

Las gráficas de los valores predichos contra los reales en los diferentes cluster son:

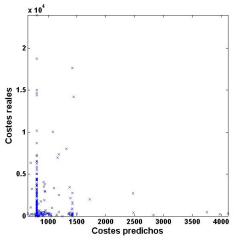


Figura 6.83: Gráfica de valores predichos contra reales en el cluster “barato”

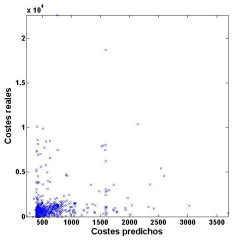


Figura 6.84: Gráfica de valores predichos contra reales en el cluster “intermedio”

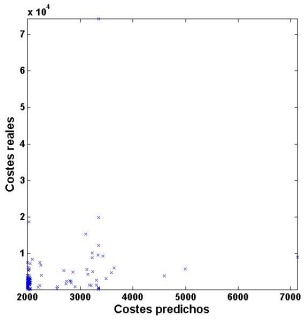


Figura 6.85: Gráfica de valores predichos contra reales en el cluster “caro”

Y los errores absolutos de predicción de cada una de las muestras de cada cluster:

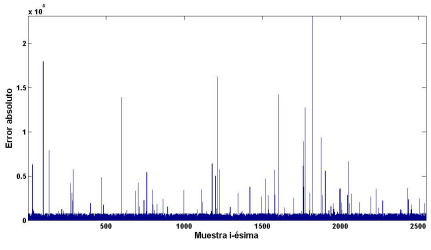


Figura 6.86: Gráfica de errores absolutos de cada muestra del cluster “barato”

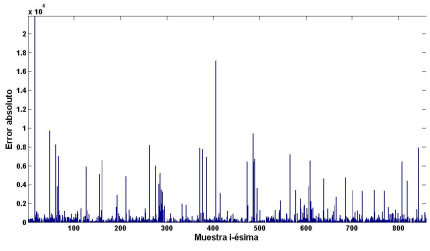


Figura 6.87: Gráfica de errores absolutos de cada muestra del cluster “intermedio”

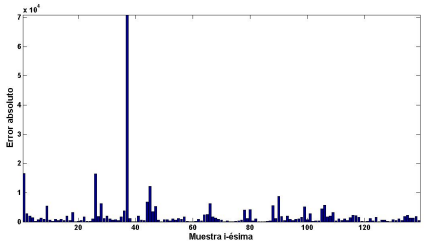


Figura 6.88: Gráfica de errores absolutos de cada muestra del cluster “caro”

Lo que se traduce a:

	Ahorro total	Ahorro por paciente y año
Primer cluster	17351.91 €	6.81 €
Segundo cluster	67455.49 €	78.61 €
Tercer cluster	313122.47 €	2252.67 €
Total	397929.87 €	112.28 €

Cuadro 6.13: Tabla de información monetaria conseguida con una SVM RBF de regresión y usando los hiperparámetros óptimos para la función de coste 6.1 (sección 6.6)

Siendo éste el mejor resultado dentro de los regresores.

Capítulo 7

Conclusiones y futuras líneas de investigación

En este capítulo se comentará, a la vista de los resultados obtenidos en el Capítulo 6, las conclusiones que podemos extraer de los mismos, así como posibles futuras líneas de investigación que permitan reducir en la medida de lo posible las simplificaciones que se han llevado a cabo, generando un modelo más acorde a la realidad.

7.1. Conclusiones

En primer lugar se ha realizado un estudio económico del impacto logrado con nuestro modelo de aprendizaje estadístico en el sistema sanitario debido a una adecuada atención primaria en pacientes psiquiátricos. Cogiendo el mejor resultado obtenido en ahorro medio por paciente y año, que es igual a 120.47 € conseguido con una SVM RBF de clasificación, y teniendo en cuenta que el coste medio de la base de datos, tras la purga inicial, proporcionada por Enrique Baca García es igual a 448 €, estamos consiguiendo un 26.8 % de ahorro, lo que, pese a todas las simplificaciones, es una cifra bastante llamativa. La purga inicial que se menciona es la explicada en el Capítulo 2, y fue realizada en el historial de pacientes para eliminar aquellos pacientes que hemos considerado que tienen un comportamiento psiquiátrico que no sigue ninguna pauta, y por tanto, es inútil tenerlos en cuenta en el estudio, pues podría desvirtuar el modelo.

De todas maneras, y aunque nuestra base de datos solo proporciona los costes debido a la asistencia de los pacientes a las diferentes entidades sanitarias (ambulatorio, hospital y urgencias), es decir, no incluye ni gastos de medicamentos ni otros gastos causados por los pacientes, la cifra de referencia de 1311.69 € usada en esta memoria es una cifra de "prueba" y que fue considerada como razonable por el grupo de psiquiatría como límite entre pacientes caros y baratos, y que además está fundamentada en que representa la media de los costes directos por paciente (libro de la Estrategia en Salud Mental del Sistema Nacional de Salud, [5]).

Otro aspecto a destacar es la aportación en los resultados de las diferentes técnicas usadas. Fijándonos en los resultados totales observamos cómo gran parte del ahorro es conseguido generalmente por la correcta detección de pacientes que han generado un gasto considerado como caro, es decir, la labor realizada por el algoritmo de clustering, en este caso el k-means, es más que óptima, aislando un gran porcentaje de los pacientes causantes de un gasto mayor del coste de referencia (1311.69 €). Por otra parte, la influencia de los algoritmos de clasificación/regresión solo se observa en los clusters barato e intermedio, y es en este último donde radica principalmente las diferencias de ahorro entras las diferencias técnicas, ya que el ahorro conseguido en el cluster caro es común a todas las técnicas y debido, como se ha dicho, al algoritmo de clustering, que hace que, aunque los técnicas de regresión/clasificación no funcionen bien, el ahorro detectando a todos como caros sea muy grande.

Estableciendo un ranking de los resultados obtenidos con las diferentes técnicas las mejores serían las de clasificación y luego las de regresión. Y dentro de ellas han resultado mejores las de tipo RBF por delante de las lineales. Es decir, la clasificación quedaría:

1. Clasificación RBF: 120.47 €.
2. Clasificación lineal: 112.43 €.
3. Regresión RBF: 112.28 €.
4. Regresión SVM: 107.88 €.
5. Regresión lineal: 104.85 €.

El hecho de que dé mejores resultados una SVM RBF que una lineal se debe a que estamos ante problema en el que las relaciones entre las variables independientes y la dependiente se ajustan mejor de una manera no lineal.

Por tanto, las principales conclusiones que podemos extraer son: el buen resultado obtenido pese a las simplificaciones que se han realizado, la buena detección de pacientes caros por el algoritmo de clustering responsable en ocasiones de todo, o de una gran parte, del ahorro conseguido, y que la inclusión de regresión/clasificación solo tiene una incidencia positiva en los clusters barato e intermedio. En resumen, este proyecto es un buen primer paso para construir un esquema que permita implementar un sistema de detección temprana de uso potencialmente caro del sistema sanitario por parte de pacientes psiquiátricos.

7.2. Futuras líneas de investigación

Consideramos que las ampliaciones más interesantes de este trabajo puede estar relacionadas con:

- La consecución de una estimación más realista de todos los costes generados por los pacientes, no solo de los debidos a los distintos tipos de consultas psiquiátricas (ambulatorio, urgencias y hospital), ya que nuestro modelo intenta dar una atención temprana adecuada a aquellos pacientes potencialmente caros, y obviamente habrá que tener en cuenta todos los gastos que éstos ocasionan. Es por este motivo por el cual también se debe incluir toda la información referente a los costes directos e indirectos provocados por los pacientes a la sanidad pública. Para conseguir esto sería recomendable iniciar una línea de investigación por patología, debido a que cada una tendrá un coste medio por paciente y otros costes como pueden ser los farmacéuticos, que pueden diferir de manera notable respecto a otra patología.

- La generación de un modelo de caracterización del comportamiento clínico de pacientes que incorporara datos relacionados con los diagnósticos y otras variables socioeconómicas o familiares de los pacientes. Así obtendríamos un modelo de aprendizaje más completo y que nos permitiría observar posibles relaciones entre tipos de enfermedades y sus comportamientos clínicos.

- La inclusión de la función de coste que mide el ahorro real conseguido en la estructura del funcional de la SVM, de manera que la selección de hiperparámetros que maximizan el ahorro se haga de una manera más ortodoxa, ya que así obtendríamos directamente soluciones que lo que han buscado ha sido maximizar el ahorro tal y como ha sido definido en la memoria, con un determinado conjunto de hiperparámetros.

- El estudio de más algoritmos de clustering, ya que vista la importancia de éstos en el resultado, puede resultar que la utilización de otro algoritmo de clustering basado en algo que vaya más allá de la distancia euclídea, pueda mejorar el ahorro conseguido o, por lo menos, hacer que la posterior tarea de clasificación/regresión sea más fructífera.

- Conseguir que los algoritmos de clasificación/regresión tengan alguna repercusión positiva en el ahorro medio conseguido en el cluster caro, y que éste no se deba solo a la buena detección de un gran porcentaje de pacientes potencialmente caros.

Bibliografía

[1] Ethem Alpaydin. *Introduction to Machine Learning.* The MIT Press, 2004.

[2] Christopher M. Bishop. *Neural networks for pattern recognition.* Clarendon Press, 1995.

[3] Christopher J.C. Burges. A tutorial on support vector machines for pattern. 1998.

[4] Chih-Jen Lin Chich-Wei Hsu, Chih-Chung Chang. A practical guide to support vector classification. 2007.

[5] Ministerio de Sanidad y Consumo. *Estrategia en Salud Mental del Sistema Nacional de Salud.* 2007.

[6] Giovanni Giambene. *Queuing Theory and Telecommunications.* Springer, 2005.

[7] Simon Haykin. *Neural Networks.* Prentice-Hall, 1999.

[8] Ralf Herbrich. *Learning kernel classifiers : theory and algorithms.* MIT Press, 2002.

[9] Robert J. Howlett. *Radial basis function networks.* Physica, 2001.

[10] I.T. Jolliffe. *Principal Component Analysis.* Springer, 2002.

[11] T Kohonen. *Self-Organization and Associative Memory.* Springer-Verlag, 1988.

[12] T Kohonen. *Self-Organizating Maps.* Springer-Verlag, 1997.

[13] Igor Kononenko. *Machine learning and data mining : introduction to principles and algorithms.* Horwood Publishing, 2007.

[14] J. Srinivas M. Ananda Rao. *Neural Networks, Algorithms and Applications.* Alpha Science, 2003.

[15] Minsky Marvin. *Information, Computer and Artificial Intelligence.* MIT Press.

[16] Jahon Shawe-Taylor Nello Cristianini. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[17] Nils J. Nilsson. *Introduction to Machine Learning.* Standford University, 1996.

[18] Daniel Peña. *Estadística: Modelos y Métodos (2^a parte).* Alianza, 1986.

[19] Antonio Artés Rodriguez, José Miguel Leiva Murillo, Mario de Prado Cumplido, Ricardo Santiago Mozos, Fernando Pérez Cruz, Ángel Navia Vázquez, and Aníbal R. Figueiras Vidal. Estado del arte y líneas de investigación abiertas en máquinas de vectores soporte.

[20] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear. *Science*, 2000.

[21] Jaakk0 Peltonen Samuel Kaski, Janne Sinkonnen. Bankruptcy analysis with self-organizing maps in learning metrics. 2001.

[22] Schölkopf and Smola. *Learning with Kernels.* The MIT Press, 2002.

[23] John Shawe-Taylor. *Kernel methods for pattern analysis.* Cambridge Univerity Press, 2004.

[24] Daniel W. Stroock. *An Introduction to Markov Processes.* Springer, 2005.

[25] Richard S. Sutton. *Reinforcement learning.* Kluwer Academic, 1992.

[26] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear. *Science*, 2000.

[27] Vladimir Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1989.

[28] Andrew Webb. *Statiscal Pattern Recognition.* Wiley, 2002.

[29] Hui Xiong Weili Wu and Shashi Shekhar. *Clustering and Infomation Retrieval.* Kluwer Academic Publishers, 2004.

[30] Quintín Martín Martín y Yanira del Rosario de Paz Santana. *Aplicación de las redes neuronales artificiales a la regresión.* La Muralla, 2007.