International Conference on Mathematical and Statistical Modeling in Honor of Enrique Castillo. June 28-30, 2006

Selective Naive Bayes Predictor with Mixtures of Truncated Exponentials

María Morales, Carmelo Rodríguez and Antonio Salmerón*
Department of Statistics and Applied Mathematics, University of Almería.

Abstract

Naive Bayes models have been successfully used in classification problems where the class variable is discrete. Naive Bayes models have been applied to regression or prediction problems, i.e. classification problems with continuous class, but usually under the assumption that the joint distribution of the feature variables and the class is multivariate Gaussian. In this paper we are interested in regression problems where some of the feature variables are discrete while the others are continuous. We propose a Naive Bayes predictor based on the approximation of the joint distribution by a Mixture of Truncated Exponentials (MTE). We have designed a procedure for selecting the variables that should be used in the construction of the model. This scheme is based on the mutual information between each of the candidate variables and the class. Since the mutual information can not be computed exactly for the MTE distribution, we introduce an unbiased estimator of it, based on Monte Carlo methods. We test the performance of the proposed model in three real life problems, related to higher education management.

Key Words: Bayesian networks, mixtures of truncated exponentials, naive Bayes models, probabilistic prediction.

1 Introduction

The problem of *classification* consists of determining the class to which an individual belongs given that some features about that individual are known. In other words, classification means to predict the value of a *class* variable given the value of some other *feature* variales. Naive Bayes models have been successfully employed in classification problems where the class variable is discrete (Friedman et al., 1997). A naive Bayes model is a particular class of Bayesian network, which is a decomposition of a joint distribution as a product of conditionals, according to the independencies induced by the structure of a directed acyclic graph in which each vertex

*Correspondence to: Antonio Salmerón. Department of Statistics and Applied Mathematics. University of Almería. Spain.

corresponds to one of the variables in the distribution (Pearl, 1988), and attached to each node there is a conditional distribution for it given its parents. The naive Bayes structure is obtained as a graph with the class variable as root and whose only arcs are those that aim from the class variable to each one of the features.

When the class variable is continuous, the problem of determining the value of the class for a given configuration of values of the feature variables is called *regression* or *prediction* rather than classification. Naive Bayes models have been applied to regression problems but only under the assumption that the joint distribution of the feature variables and the class is multivariate Gaussian (Gámez and Salmerón, 2005). If the normality assumption is not fulfilled, the problem of refression with naive Bayes models has been approached using kernel densities to model the conditional distibution in the Bayesian network (Frank et al., 2000), but the obtained results are poor. Furthermore, the use of kernels introduce a high complexity in the model, which can be problematic specially because standard algorithms for carrying out the computations in Bayesian networks are not valid for kernels. A common restriction of Gaussian models and kernel-based models is that they only apply to scenarios in which all the variables are continuous.

In this paper we are interested in regression problems where some of the feature variables are discrete while the others are continuous. Therefore, the joint distribution is not multivariate Gaussian in any case, due to the presence of discrete variables. We propose a Naive Bayes predictor based on the approximation of the joint distribution by a Mixture of Truncated Exponentials (MTE). The MTE model (Moral et al., 2001) has been proposed in the context of Bayesian networks as a solution to the presence of discrete and continuous variables simultaneously, showing good features as an exact model as well as an approximation of other probability distributions (Cobb, Rumí, and Salmerón, 2005; Cobb, Shenoy, and Rumí, 2006).

The rest of the paper is organised as follows. In section 2 we review the necessary concepts of Bayesian networks and explain how they can be used for regression. The MTE model is introduced in section 3. Afterwards, we propose the naive Bayes predictor based on MTEs in section 4, and a variable selection scheme for it in section 5. The application of the proposed models to three real world problems is described in section 6. The paper

ends with conclusions in section 7.

2 Bayesian networks and regression

Consider a problem defined by a set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. A *Bayesian network* (Jensen, 2001; Pearl, 1988) is a directed acyclic graph where each variable is assigned to one node, which has associated a conditional distribution given its parents. An arc linking two variables indicates the existence of probabilistic dependence between both of them. An important feature of Bayesian networks is that the joint distribution over $\mathbf{X}$ factorises according to the *d*-separation criterion as follows (Pearl, 1988):

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|pa(x_i)) \ , \tag{2.1}$$

where $Pa(X_i)$ denotes the set of parents of variable X_i and $pa(x_i)$ is a configuration of values of them. Figure 1 shows a Bayesian network which encode the distribution

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_5|x_3)p(x_4|x_2, x_3) \ .$$

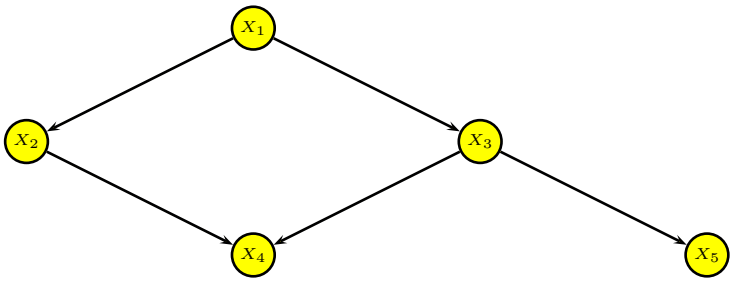

Figure 1: *A sample Bayesian network*

A Bayesian network can be used for classification purposes if it consists of a *class* variable, C, and a set of *feature* variables $X_1, \ldots, X_n$, so that an

individual with observed features $x_1, \ldots, x_n$ will be classified as a member of class c^* obtained as

$$c^* = \arg\max_{c \in \Omega_C} p(c|x_1, \ldots, x_n) \ , \tag{2.2}$$

where Ω_C denotes the support of variable C. Similarly, a Bayesian network can be used for regression, i.e, when C is continuous. However, in this case the goal is to compute the posterior distribution of the class variable given the observed features $x_1, \ldots, x_n$, and once this distribution is computed, a numerical prediction can be given using the mean, the median or the mode.

Note that $p(c|x_1, \ldots, x_n)$ is proportional to $p(c) \times p(x_1, \ldots, x_n|c)$, and therefore solving the regression problem would require to specify an n dimensional distribution for $X_1, \ldots, X_n$ given the class. Using the factorisation determined by the Bayesian network, this problem is simplified. The extreme case is the so-called *Naive Bayes* structure (Friedman, Geiger, and Goldszmidt, 1997; Duda, Hart, and Stork, 2001), where all the feature variables are considered independent given the class. An example of Naive Bayes structure can be seen in figure 2.

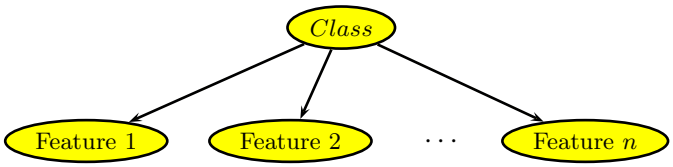

Figure 2: Structure of a Naive Bayes classifier/predictor

The independence assumption behind naive Bayes models is somehow compensated by the reduction on the number of parameters to be estimated from data, since in this case, it holds that

$$p(c|x_1, \ldots, x_n) = p(c) \prod_{i=1}^{n} p(x_i|c) \ , \tag{2.3}$$

which means that, instead of one n-dimensional conditional distribution, n one-dimensional conditional distributions are estimated.

3 The MTE model

Throughout this paper, random variables will be denoted by capital letters, and their values by lowercase letters. In the multi-dimensional case, bold-faced characters will be used. The domain of the variable $\mathbf{X}$ is denoted by $\Omega_{\mathbf{X}}$. The MTE model is defined by its corresponding potential and density as follows (Moral et al., 2001):

Definition 3.1. (MTE potential) *Let $\mathbf{X}$ be a mixed n-dimensional random vector. Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of $\mathbf{X}$, respectively, with $c + d = n$. We say that a function $f : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ is a* Mixture of Truncated Exponentials potential (MTE potential) *if one of the next conditions holds:*

i. $\mathbf{Y} = \emptyset$ *and f can be written as*

$$f(\mathbf{x}) = f(\mathbf{z}) = a_0 + \sum_{i=1}^{m} a_i \exp\left\{\sum_{j=1}^{c} b_i^{(j)} z_j\right\} \tag{3.1}$$

for all $\mathbf{z} \in \Omega_{\mathbf{Z}}$, where a_i, $i = 0, \ldots, m$ and $b_i^{(j)}$, $i = 1, \ldots, m$, $j = 1, \ldots, c$ are real numbers.

ii. $\mathbf{Y} = \emptyset$ *and there is a partition $D_1, \ldots, D_k$ of $\Omega_{\mathbf{Z}}$ into hypercubes such that f is defined as*

$$f(\mathbf{x}) = f(\mathbf{z}) = f_i(\mathbf{z}) \quad \text{if} \quad \mathbf{z} \in D_i \ ,$$

where each f_i, $i = 1, \ldots, k$ can be written in the form of equation (3.1).

iii. $\mathbf{Y} \neq \emptyset$ *and for each fixed value $\mathbf{y} \in \Omega_{\mathbf{Y}}$, $f_{\mathbf{y}}(\mathbf{z}) = f(\mathbf{y}, \mathbf{z})$ can be defined as in* ii.

Example 3.1. *The function ϕ defined as*

$$\phi(z_1, z_2) = \begin{cases} 2 + e^{3z_1+z_2} + e^{z_1+z_2} & \text{if} \quad 0 < z_1 \leq 1, \ 0 < z_2 < 2 \\ 1 + e^{z_1+z_2} & \text{if} \quad 0 < z_1 \leq 1, \ 2 \leq z_2 < 3 \\ \dfrac{1}{4} + e^{2z_1+z_2} & \text{if} \quad 1 < z_1 < 2, \ 0 < z_2 < 2 \\ \dfrac{1}{2} + 5e^{z_1+2z_2} & \text{if} \quad 1 < z_1 < 2, \ 2 \leq z_2 < 3 \end{cases}$$

is an MTE potential since all of its parts are MTE potentials.

Definition 3.2. (MTE density) *An MTE potential f is an* MTE density *if*

$$\sum_{\mathbf{y}\in\Omega_{\mathbf{Y}}} \int_{\Omega_{\mathbf{Z}}} f(\mathbf{y},\mathbf{z})d\mathbf{z} = 1 \ .$$

A *conditional MTE density* can be specified by dividing the domain of the conditioning variables and specifying an MTE density for the conditioned variable for each configuration of splits of the conditioning variables. In (Moral et al., 2001) a data structure was proposed to represent MTE potentials, which is specially appropriate for this kind of conditional densities: The so-called *mixed probability trees* or mixed trees for short. The formal definition is as follows:

Definition 3.3. (Mixed tree) *We say that a tree $\mathcal{T}$ is a* mixed tree *if it meets the following conditions:*

i. *Every internal node represents a random variable (either discrete or continuous).*

ii. *Every arc outgoing from a continuous variable Z is labeled with an interval of values of Z, so that the domain of Z is the union of the intervals corresponding to the arcs Z-outgoing.*

iii. *Every discrete variable has a number of outgoing arcs equal to its number of states.*

iv. *Each leaf node contains an MTE potential defined on variables in the path from the root to that leaf.*

Mixed trees can represent MTE potentials defined by parts. Each entire branch in the tree determines one sub-region of the space where the potential is defined, and the function stored in the leaf of a branch is the definition of the potential in the corresponding sub-region.

Example 3.2. *Consider the following MTE potential, defined for a discrete variable (Y_1) and two continuous variables (Z_1 and Z_2).*

$$\phi(y_1, z_1, z_2) = \begin{cases} 2 + e^{3z_1+z_2} & \text{if } \ y_1 = 0,\ 0 < z_1 \leq 1,\ 0 < z_2 < 2 \\ 1 + e^{z_1+z_2} & \text{if } \ y_1 = 0,\ 0 < z_1 \leq 1,\ 2 \leq z_2 < 3 \\ \frac{1}{4} + e^{2z_1+z_2} & \text{if } \ y_1 = 0,\ 1 < z_1 < 2,\ 0 < z_2 < 2 \\ \frac{1}{2} + 5e^{z_1+2z_2} & \text{if } \ y_1 = 0,\ 1 < z_1 < 2,\ 2 \leq z_2 < 3 \\ 1 + 2e^{2z_1+z_2} & \text{if } \ y_1 = 1,\ 0 < z_1 \leq 1,\ 0 < z_2 < 2 \\ 1 + 2e^{z_1+z_2} & \text{if } \ y_1 = 1,\ 0 < z_1 \leq 1,\ 2 \leq z_2 < 3 \\ \frac{1}{3} + e^{z_1+z_2} & \text{if } \ y_1 = 1,\ 1 < z_1 < 2,\ 0 < z_2 < 2 \\ \frac{1}{2} + e^{z_1-z_2} & \text{if } \ y_1 = 1,\ 1 < z_1 < 2,\ 2 \leq z_2 < 3 \end{cases}$$

A possible representation of this potential by means of a mixed probability tree is displayed in figure 3.

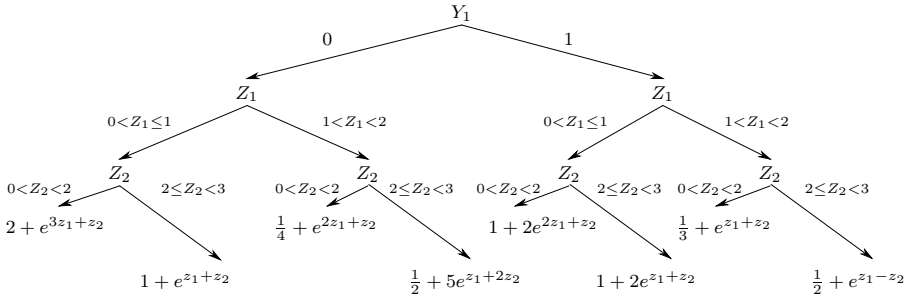

Figure 3: *A mixed probability tree representing the potential ϕ in example 3.2.*

4 The naive Bayes predictor based on MTEs

Our proposal consists of solving the regression problem in which some feature variables are discrete and some other continuous using a predictor with

naive Bayes structure, and modelling the corresponding conditional distributions as MTEs. More precisely, we will use a 5-parameter MTE for each split of the support of the variable, which means that in each split there will be 5 parameters to be estimated from data:

$$f(x) = a_0 + a_1 e^{a_2 x} + a_3 e^{a_4 x} \ . \tag{4.1}$$

We follow the estimation procedure developed by Romero et al. (2006), which has these main steps:

1. A Gaussian kernel density is fitted to the data.
2. The domain of the variable is split according to changes in concavity/convexity or increase/decrease in the kernel.
3. In each split, a 5-parameter MTE is fitted to the kernel by least squares estimation.

Once the model is constructed, it can be used to predict the value of the class variable given that the value of the feature variables is observed. The prediction is carried out by computing the posterior distribution of the class given the observed values for the features. A numerical prediction for the class value can be obtained from the posterior distribution. In the experiments we have carried out, we have observed that the best results are obtained using the expected value rather and the median rather than the mode. In this paper we have computed the posterior distribution using the Shenoy-Shafer algorithm for probability updating in Bayesian networks, but adapted to the MTE case as in (Rumí and Salmerón, 2005).

5 Selecting the feature variables

An important issue to address in any classification or regression problem is to choose the feature variables to be included in the model. In general, it is not true that including more variables increases the accuracy of the model. It can happen that some variables are not informative for the class and therefore including them in the model provides noise to the predictor, besides the increase in the number of parameters to be learnt from data.

There are different approaches to the problem of selecting variables in prediction and classification problems:

- The *filter* approach, which consists of establishing a ranking of the variables according to some measure of relevance respect to the class variable, usually called *filter measure*. Then a threshold for the ranking is selected and variables below that threshold are discarded.
- The *wrapper* approach, however, proceeds by constructing several models with different sets of feature variables, and finally the model with higher accuracy is selected.
- The *filter-wrapper* approach is a mixture of the former ones. First of all, the variables are ordered using a filter measure and then they are incrementally included or excluded from the model according to that order, so that a variable is included whenever it increases the accuracy of the model.

The accuracy of a model is measured in this way:

1. The database containing the information for the feature variables and the class is divided into two parts, D_l and D_t.
2. The model is estimated using database D_l. Usually, D_l contains the 70% of the records in D, chosen at random, while the remaining 30% is assigned to D_t. This is the choice we have adopted in the case studies reported in this paper.
3. The accuracy of the model is measured using database D_t, by measuring the rooted mean squared error between the actual values of the class and those ones predicted by the model for the records in database D_t. If we call $c_1, \ldots, c_m$ the values of the class for the registers in database D_t and $\hat{c}_1, \ldots, \hat{c}_m$ the corresponding estimates provided by the model, the rooted mean squared error is obtained as

$$rmse = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(c_i - \hat{c}_i)^2} \ . \tag{5.1}$$

In this paper we have followed a filter-wrapper approach, using as filter measure the *mutual information* between each variable and the class. The mutual information has been successfully applied as filter measure in classification problems for instance by Pérez et al. (2006). The mutual information between two random variables X and Y is defined as

$$I(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2 \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}dydx \ , \qquad (5.2)$$

where f_{XY} is the joint density for X and Y, f_X is the marginal density for X and f_Y is the marginal for Y.

In the case of using MTEs, each density is expressed as in equation (4.1), what troubles the computation of the integral in equation (5.2). Therefore, we have chosen to estimate the value of the mutual information. The estimation procedure that we have designed is based on the next theorem.

Theorem 5.1. *Let X and Y be two continuous random variables with densities f_X and f_Y respectively, and joint density f_{XY}. Let $f_{X|Y}$ denote the conditional density of X given Y. Let $Y_1,\ldots,Y_n$ be a sample drawn independently from distribution $f_Y(y)$. Let $X_1,\ldots,X_n$ be a sample such that each X_i, $i=1,\ldots,n$ is drawn from distribution $f_{X|Y}(x|Y_i)$. Then,*

$$\hat{I}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)\right) \qquad (5.3)$$

is an unbiased estimator of $I(X,Y)$.

Proof. According to the way in which samples $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$ are obtained, it follows that the joint sample of bivariate points (X_1,Y_1),..., (X_n,Y_n) is actually drawn from the distribution $f_{XY}(x,y)$. Therefore, if we denote by $E_{f_{XY}}$ the expected value with respect to density f_{XY}, we have that

$$\begin{aligned} E[\hat{I}(X,Y)] &= E_{f_{XY}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E_{f_{XY}}\left[\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} E_{f_{XY}}\left[\log_2 \frac{f_{X|Y}(X_i|Y_i)}{f_X(X_i)}\right] \\
&= E_{f_{XY}}\left[\log_2 \frac{f_{X|Y}(X|Y)}{f_X(X)}\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2 \frac{f_{X|Y}(x|y)}{f_X(x)}dydx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2 \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}dydx \\
&= I(X,Y)\ .
\end{aligned}
$$

□

The consistency of estimator $\hat{I}(X,Y)$ is guaranteed by the next proposition.

Proposition 5.1. *Let $\hat{I}_n(X,Y)$ denote estimator $\hat{I}(X,Y)$ when it is computed from a sample of size n. The succession $\{\hat{I}_n(X,Y)\}_{n=1}^{\infty}$ is consistent.*

Proof. It is enough to show that

(i) $\lim_{n\to\infty} E[\hat{I}_n(X,Y)] = I(X,Y)$ and

(ii) $\lim_{n\to\infty} \mathrm{Var}(\hat{I}_n(X,Y)) = 0$.

The proof of (i) is trivial, since according to theorem 5.1, $E[\hat{I}_n(X,Y)] = I(X,Y)$ for all $n > 0$ and therefore the limit is equal to $I(X,Y)$ as well.

In order to prove (ii), we need the expression of the variance of the estimator.

$$
\begin{aligned}
\mathrm{Var}(\hat{I}_n(X,Y)) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)\right)\right) \\
&= \frac{1}{n}\mathrm{Var}\left(\log_2 f_{X|Y}(X|Y) - \log_2 f_X(X)\right)
\end{aligned}
$$

where $\mathrm{Var}\left(\log_2 f_{X|Y}(X|Y) - \log_2 f_X(X)\right)$ does not depend on n and is finite whenever distributions $f_{X|Y}$ and f_X are positive. Therefore, we can conclude that

$$\lim_{n\to\infty} \mathrm{Var}(\hat{I}_n(X,Y)) = 0 \ .$$

□

Now we have the necessary tools for giving a detailed algorithm for constructing the selective naive Bayes predictor.

Algorithm 5.1 (Selective naive Bayes predictor).

INPUT:

- *The class variable C.*
- *The feature variables $X_1, \ldots, X_n$.*
- *A database D for variables $X_1, \ldots, X_n, C$.*

OUTPUT:

- *The selective naive Bayes predictor for variable C.*

1. *For $i := 1$ to n*
   - *Compute $\hat{I}(X_i, C)$.*
2. *Let $X_{(1)}, \ldots, X_{(n)}$ be a decreasing order of the feature variables according to $\hat{I}(X_{(i)}, C)$.*
3. *Divide the database into two sets, one for learning the model (D_l) and the other for testing the accuracy of the learnt model (D_t).*
4. *Construct a naive Bayes predictor, M, for variables C and $X_{(1)}$:*
   (a) *Estimate a marginal MTE density for C, f_C, from database D_l.*
   (b) *Estimate a conditional MTE density for $X_{(1)}$ given C, $f_{X_{(1)}|C}$, from database D_l.*
   (c) *Let $rmse(M)$ be the estimated accuracy of model M using database D_t, according to formula* (5.1).
5. *For $i := 2$ to n*

(a) Let M_1 be the naive Bayes predictor obtained from M by adding variable $X_{(i)}$, i.e., by estimating a conditional density for $X_{(i)}$ given C, $f_{X_{(i)}|C}$, from database D_l.

(b) Let $rmse(M_1)$ be the estimated accuracy of model M_1 using database D_t, according to formula (5.1)*.*

(c) If ($rmse(M_1) \leq rmse(M)$)

- $M := M_1$.

6. Return (M)

6 Case studies

We have tested the performance of the selective naive Bayes predictor in three practical problems related to higher education management. More precisely, we will use data regarding the University of Almería.

Problems 1 and 2 consists of predicting the *performance rate* and *success rate*, respectively, for a given degree program. The study is restricted to a database with information about all the degree programs in the university of Almería in years 2001 to 2004.

The *performance rate* is defined as

$$pr = \frac{n_o}{n_s} \tag{6.1}$$

where n_s is the number of credits[1] of all the subjects selected by the students in a given year, and n_o is the number of credits actually obtained by the students at the end of the same year.

The *success rate* is defined as

$$sr = \frac{n_o}{n_e} \tag{6.2}$$

where n_o is as defined above and n_e is the number of credits for which the students have actually went to the final exam.

In problems 1 and 2, we have the following feature variables:

[1] Spanish university subjects are measured in credits. One credit corresponds to ten hours of lectures.

- **Degree**: The degree program.
- **OptRate**: Rate of credits that can be obtained with subjects freely chosen by the student.
- **OptOffer**: Number of free-election subjects offered (in credits) divided by the number of free-election credits that the student has to obtain to complete the degree.
- **Prt**: Ratio between the amount of practical credits and total amount of credits required in a degree program.
- **SmallGroups**: Ratio between the number of classes with no more than 20 students and the global number of classes in each subject.
- **BigGroups**: Ratio between the number of classes with not less than 80 students and the global number of classes in each subject.
- **Dedication**: Number of credits coursed by the students divided by the number of students.
- **Give-upRate**: Rate of students that leave the university without having finished the degree program.
- **Rate_S-L**: Number of students per lecturer.
- **PhD**: Fraction of credits in the degree taught by lecturers owning the PhD degree.

Problem 3 consists of predicting the number of students in a given subject. In this case, we have a database containing information about all the subjects offered at the University of Almería from years 2001 to 2004. The variables considered in problem 3 are:

- **Degree**: The degree program.
- **Period**: Part of the degree (first or second half) in which the subject is located.
- **Subject**.
- **Course**: The course (year) in which the subject is located.

Table 1: *Estimated mutual information between each variable and the class for problem 1 (predicting the performance rate)*

Variable	Mutual Information
OptOffer	0.1998
Prt	0.1791
SmallGroups	0.1361
OptRate	0.1357
Dedication	0.1119
Give-upRate	0.1033
Rate_S-L	0.0692
Degree	0.0428
BigGroups	0.0235
PhD	0.0169
AccessMark	0

Table 2: *Estimated mutual information between each variable and the class for problem 2 (predicting the success rate)*

Variable	Mutual Information
Prt	0.1400
SmallGroups	0.1026
Dedication	0.0939
OptRate	0.0583
OptOffer	0.0544
Rate_S-L	0.0497
Degree	0.0262
Give-upRate	0.0253
BigGroups	0.0126
PhD	0.0098
AccessMark	0

- **AXX**: Number of students in the given subject in year XX, ranging from 01 to 04.
- **prXX**: Performance rate for the given subject in year XX, ranging from 01 to 04.

Table 3: *Estimated mutual information between each variable and the class for problem 3 (predicting the number of students)*

Variable	Mutual Information
S04	0.8338
S03	0.6853
S02	0.5795
S01	0.4953
Degree	0.0865
Period	0.0630
pr01	0.0526
pr04	0.0520
Subject	0.0516
pr03	0.0470
pr02	0.0450
Course	0.0257

The goal of problem 3 is, therefore, to predict the number of students in year 2005.

In the three problems, we have tested the following models:

- **NB(mean)**: Naive Bayes predictor including all the feature variables and predicting with the mean of the posterior distribution.
- **NB(median)**: Naive Bayes predictor including all the feature variables and predicting with the median of the posterior distribution.
- **SNB(mean)**: Selective naive Bayes predictor obtained by algorithm 5.1 and predicting with the mean of the posterior distribution.
- **SNB(median)**: Selective naive Bayes predictor obtained by algorithm 5.1 and predicting with the median of the posterior distribution.
- **Linear model**: A linear regression model including all the variables, and considering the discrete variables as continuous.

The estimated mutual information for the considered variables in problems 1, 2 and 3 can be seen in tables 1, 2 and 3 respectively. The results obtained in the three problems are summarised in tables 4, 5 and 6.

Table 4: *Results for problem 1 (predicting the performance rate)*

Method	rmse	No. variables
NB(mean)	0.0884	11
NB(median)	0.0921	11
SNB(mean)	0.0818	9
SNB(median)	**0.0814**	**6**
Linear model	0.1154	11

Table 5: *Results for problem 2 (predicting the success rate)*

Method	rmse	No. variables
NB(mean)	0.0462	11
NB(median)	0.0471	11
SNB(mean)	0.0383	**8**
SNB(median)	**0.0381**	**8**
Linear model	0.0476	11

Table 6: *Results for problem 3 (predicting the number of students)*

Method	rmse	No. variables
NB(mean)	31.1731	12
NB(median)	31.7893	12
SNB(mean)	23.6530	**7**
SNB(median)	23.6138	**7**
Linear model	**16.4054**	12

6.1 Results discussion

In two out of the three problems, the selective naive Bayes predictor provides the best results. Only in problem 3 the linear model is better. It is not surprising, since it is known that Bayesian networks are of special interest when representing nonlinear systems.

The variable selection procedure always provides significant improvements, which are specially remarkable in problem 3. It can also be noticed that using the median instead of the mean for the numerical prediction results in more accurate estimations, probably due to the robustness of

the median. However, the differences between both are minor. We have not reported results of the mode, since they were far away from the ones provided by the mean and the median.

An added value of NB and SNB with respect to the linear model is that they do not only provide numerical prediction, but they also give the posterior distribution of the class variable, which allows to make other types of inferences like answering queries as *what is the probability of the number of students being between 100 and 150*.

7 Conclusions

In this paper we have introduced a framework for approaching regression problems where some of the feature variables are discrete, based on the Bayesian network methodology and using mixtures of truncated exponentials as underlying probabilistic model. We have also proposed a variable selection scheme according to the mutual information, and adopting a filter-wrapper strategy.

We have applied the developed models to three practical problems, showing reasonably good behaviour, except for the case of predicting the number of students in which the linear model outperformed the predictors proposed here.

In future works, we plan to test the models in more real-world and synthetic datasets, and compare the performance of the selective naive Bayes predictor versus the Gaussian model developed by Gámez and Salmerón (2005). Furthermore, more sophisticated variable selection strategies can be considered.

References

Cobb, B., Rumí, R., and Salmerón, A. (2005). Modeling conditional distributions of continuous variables in Bayesian networks. *IDA'05. Lecture Notes in Computer Science*, 3646:36–45.

Cobb, B., Shenoy, P., and Rumí, R. (2006). Approximating probability density functions with mixtures of truncated exponentials. *Statistics and Coputing*, In press.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification*. Wiley Interscience.

Frank, E., Trigg, L., Holmes, G., and Witten, I. (2000). Technical note: Naive Bayes for regression. *Machine Learning*, 41:5–25.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29:131–163.

Gámez, J. and Salmerón, A. (2005). Predicción del valor genético en ovejas de raza manchega usando técnicas de aprendizaje automático. In *Actas de las VI Jornadas de Transferencia de Tecnología en Inteligencia Artificial*, pp. 71–80. Paraninfo.

Jensen, F. V. (2001). *Bayesian networks and decision graphs*. Springer.

Moral, S., Rumí, R., and Salmerón, A. (2001). Mixtures of truncated exponentials in hybrid Bayesian networks. In *Lecture Notes in Artificial Intelligence*, vol. 2143, pp. 135–143.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan-Kaufmann (San Mateo).

Pérez, A., Larrañaga, P., and Inza, I. (2006). Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning*, In press.

Romero, V., Rumí, R., and Salmerón, A. (2006). Learning hybrid Bayesian networks using mixtures of truncated exponentials. *Inernational Journal of Approximate Reasoning*, 42:54–68.

Rumí, R. and Salmerón, A. (2005). Penniless propagation with mixtures of truncated exponentials. *ECSQARU'05. Lecture Notes in Computer Science*, 3571:39–50.