Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics
Master of Data Science

A Generalized Lagrange Multiplier Method for Support Vector Regression with Imposed Symmetry

THESIS to obtain the **DEGREE** of
MASTER OF DATA SCIENCE

A thesis presented by:
Luis Alfonso Guerrero Montaño

Thesis Advisors:
Dr. Juan Diego Sánchez Torres
Mtra. Sara Eugenia Rodríguez Reyes

Tlaquepaque, Jalisco, Nov., 2022

Instituto Tecnológico y de Estudios Superiores de Occidente
Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics
Master of Data Science Approval Form

Thesis Title: **A Generalized Lagrange Multiplier Method for Support Vector Regression with Imposed Symmetry**
Author: **Luis Alfonso Guerrero Montaño**
Thesis Approved to complete all degree requirements for the Master of Science Degree in Data Science.

Thesis Advisor, **Dr. Juan Diego Sánchez Torres**

Thesis Co-Advisor, **Mtra. Sara Eugenia Rodríguez Reyes**

Thesis Reader, **Dra. Diana Paola Montoya Escobar**

Thesis Reader, **Dra. Alma Nayeli Rodríguez Vásquez**

Academic Advisor, **Dra. Rocío Carrasco Navarro**

Tlaquepaque, Jalisco, Nov., 2022

A Generalized Lagrange Multiplier Method for Support Vector Regression with Imposed Symmetry

Luis Alfonso Guerrero Montaño

Abstract

This thesis presents an approach to support vector regression that extends the classic Vapnik's formulation. After recalling that the classic formulation contains a Lasso regularization structure in its dual form, we propose a generalized Lagrangian function with additional terms to include the Ridge regularization in the dual problem for the case with symmetry. By including both regularization methods, the resulting dual problem with the generalized Lagrangian comprises an elastic net regularization structure. Hence, as an immediate consequence, the classical formulation is a particular case of the current proposal. Finally, to demonstrate the capabilities of this approach, the document includes examples of predicting some benchmark problems. keywords: SVM, Symmetry, SVR, GLMM.

Contents

List of Tables

Dedicated to Luz, Alejandra, and Luis, my beloved family, thanks for your patience and time over these two years. To my friends, Alejandra, Elisa, Angel, Jesus, Alex, y Aldo with whom I share long nights of study and work, they have been great support and encourager to achieve success in this adventure. To my teachers, especially Paola, Rocío, Fernando, and Juan Diego who with their efforts and dedication, guided us to the enjoyment and discovery of this new world of data science. And finally, this work has been not possible without the guidance and support of the work of Sara and Gregorio.

1 Introduction

Contents

1.1 Motivation

Nowadays, using different methodologies to predict results is essential in every field of study and business. These tools are used in medical research, finance predictions, natural language processing, and many more. Even sports are gambling are also common users.

Predictive Analytics is a type of data analysis that helps forecast outcomes or identify trends using a computer model based on a set of known variables.

In particular, the SVR methodology is an example of a predictive algorithm that produces accurate and statistically significant results when applied to various domains. The SVR method is a classic machine-learning technique that has been used for decades. In this thesis, I will propose a new approach to the method using a symmetric kernel function and compare the results with the ones obtained using the classic Levenberg-Marquardt algorithm[1]. The Levenberg-Marquardt algorithm is based on the least squares method, and the Gauss-Newton optimization technique was developed in the early 1960s to solve nonlinear least squares problems.

[1] Jorge J. Moré. The Levenberg-Marquardt algorithm: Implementation and theory. In *Lecture Notes in Mathematics, Berlin Springer Verlag*, volume 630, pages 105–116. 1978. DOI: 10.1007/BFb0067700

1.2 Objective

This work aims to produce an SVR with Symmetric conditions and implement it in Python.

The objective of this thesis was to formulate and release the new SVR methodology with the Symmetric kernel for regression.

In specific, create a mathematical model, implement it and test it in two different datasets (Boston house price [2] and Diabetes [3]) to compare the results against other known methodologies thus as:

[2] scikit-learn developers. sklearn.datasets.load_boston, 2020a. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html

[3] scikit-learn developers. sklearn.datasets.load_diabetes, 2020b. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html

- Linear Regression
- Random Forest
- XGBoost Regressor
- Classic SVR

1.3 *Previous works*

This work is mainly based on the following works, listed without any specific order:

- Support Vector Machines for Pattern Classification [4]
- Generalized Lagrange multiplier method for solving problems of optimum allocation of resources Support Vector Machines for Pattern Classification [5]
- Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. [6]
- An Extended Lagrangian Approach to Support Vector Regression Based on the MAPE Loss [7]
- Imposing Symmetry in Least Squares Support Vector Machines Regression [8]

[4] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7

[5] Hugh Everett III. Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations research*, 11(3):399–417, 1963

[6] Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66 (2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085

[7] Sara Eugenia Rodríguez-Reyes, Pablo Benavides-Herrera, Gregorio Alberto Álvarez-Álvarez, Riemann Ruiz-Cruz, and Juan Diego Sánchez-Torres. An extended Lagrangian approach to support vector regression based on the MAPE loss. In *20th Mexican International Conference on Artificial Intelligence (MICAI 2021)*, oct 2021

[8] M. Espinoza, J.A.K. Suykens, and B. De Moor. Imposing symmetry in least squares support vector machines regression. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 5716–5721, 2005a. DOI: 10.1109/CDC.2005.1583074

1.4 *Document Outline*

The Chapter 2 "Preliminaries and previous results" presents the foundations of the SVR with the mathematical representation, including a Generalized Lagrangian Multiplier Method using an elastic net regularization.

The Chapter 3 "Main Results" included the development of the formulation of the Symmetric kernel approach based on the work of chapter two. This is the main chapter of the thesis and the core of the work.

The Chapter 4 "Applications to real datasets" uses the results 3to apply the method for the first time in two datasets, measuring the performance against other classic methodologies.

Finally, in the Chapter 5 "Conclusions and future work," I expose the conclusion of the work done and make some suggestions for future work to keep improving the development of the proposed method and formulation.

2 Preliminaries and previous results

Contents

2.1 Effectiveness of a regression model

To measure the effectiveness of the SVR with Symmetric conditions model, I will be using different measures that I briefly introduce in this section.

Effectiveness of a regression model

There are many different ways to measure the effectiveness of a regression model. One way is to use the R^2 statistic or coefficient of determination. R^2 measures how well the regression model fits the data. A high R^2 indicates a strong correlation between the datasets' input and output variables. An R^2 close to 1 indicates that the model provides a good representation of the data. Also, a high value of R^2 means that you have used the right predictor variables. The R^2 value tells you about the goodness of fit of a statistical model, i.e., how the statistical relationship between your independent variable(s) and your target variable looks after model building. Another measure is the "adjusted R^2". Adjusted R^2 can be between 0 and 1 and measures the amount of variance in the response variable that the regression model explains. Values of adjusted R^2 closer to 1 indicate a high correlation between the response and the predictors in the model. A value of adjusted R^2 close to 0 indicates a low correlation between the response and the predictors in the model.

MAE is another measure that tells you how close your predictions are to the actual values of your output variable. This measure gives you an idea of the uncertainty associated with the model's predictions. This quantity might be very small for large datasets and, therefore, not worth considering. However, for small sample sizes, this statistic can provide valuable information about how the model behaves under real-world conditions. The term "mean absolute error" (MAE) refers to a measurement of the accuracy of a forecast or prediction. It measures the average of the absolute differences between actual values and values predicted by the model. The higher the MAE value is, the greater the error in predicting the target variable is. Other is the MSE value which is defined as the root-mean-squared difference between the actual value and the value predicted by the model. MSE also takes into account both the mean and standard deviation of the errors. So, you can think of MSE as a measure of the average squared error in your predictions. If the value of the MSE is relatively large, it means that your model is generating a lot of errors, which may be an indication that you need to modify your model in some way. On the other hand, if the value of the MSE is small, then your model may be providing you with accurate predictions, but this does not mean that your model is perfect. A variant of the MSE models is the RMSE which is defined as the average value of the squared difference between the predicted value and the actual value. This statistic is sometimes preferred over the MSE because it is less sensitive to outliers. An outlier refers to an observation that is either very high or very low compared to the rest of the observations in the dataset. The main difference between MSE and RMSE is the location of the average. The MSE statistics is located in the center of the dataset, whereas the RMSE statistics is located at the sample means. When calculating the RMSE, you will multiply each observation by its value and sum the results up. Finally, since you will need to calculate two values for the RMSE, this method will be slower than the MSE method. Finally, the mean absolute percentage error (MAPE) measures the model's accuracy. It is very similar to the MAE in its calculation, but it measures the percentage of error rather than just the absolute amount. This means that a high MAPE value indicates that there is a fairly large amount of error in the data that the model is generating.

2.2 SVR Based on a Generalized Lagrangian

2.2.1 Support vector regression

Support vector regression (SVR) has shown to be a powerful method for proposing empirical models for predicting continuous variables [1]. The interpretability, the formulation as a convex optimization problem[2], the use of kernels[3], and its relationships to other models make the SVR a robust and reliable method for several industrial and research problems.

A well-known fact about the classic formulation of SVR is that it exhibits a Lasso regularization[4] in its dual optimization problem[5]. This event coincides with Lagrange multipliers equal to zero and the appearance of support values and vectors. Besides, the support vector methods and the Lasso regularization present substantial equivalences[6]. On the other hand, the simultaneous use of two different regularization schemes provides desirable models characteristics[7].

A remarkable case of this approach is the elastic net, where the Ridge regularization[8] works together with the Lasso[9]. Moreover, similarly to the previous case, the support vector models with two regularizations present important equivalences to the elastic net regularization[10].

In this chapter, I describe some of the basic concepts like Norms an L_1 and L_2 regularization, which are the base for setting the base to use of a new SVR model.

The chapter also proposes a new SVR by introducing a Ridge regularization term in the dual through the definition of a generalized Lagrangian function.

In this form, the current proposal considers the advantages of the simultaneous use of two different regularization structures while keeping the formality with the generalized Lagrangian approach.

[1] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8; and Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14 (3):199–222, 2004. ISSN 1573-1375. DOI: 10.1023/B:STCO.0000035301.49549.88

[2] S. Boyd and L.Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 978-0-521-83378-3

[3] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001. ISBN 9780262256933; and John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511809682

[4] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. DOI: 10.1111/j.2517-6161.1996.tb02080.x

[5] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7; and Xixuan Han and Line Clemmensen. On weighted support vector regression. *Quality and Reliability Engineering International*, 30(6):891–903, 2014. DOI: https://doi.org/10.1002/qre.1654

[6] Martin Jaggi. An equivalence between the Lasso and support vector machines. *Arxiv*, abs/1303.1152, 2013

[7] L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589–615, 2006; and Julio López, Sebastián Maldonado, and Miguel Carrasco. Double regularization methods for robust feature selection and svm classification via dc programming. *Information Sciences*, 429:377–389, 2018. ISSN 0020-0255

[8] A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151(3):501–504, 1963

[9] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x

[10] Quan Zhou, Wenlin Chen, Shiji Song, Jacob Gardner, Kilian Weinberger, and Yixin Chen. A reduction of the elastic net to support vector machines with an application to GPU computing. 2015

The Generalized Lagrange Multiplier Method (GLMM) [11] helps to connect constrained optimization, and saddle-point problems since saddle points of Lagrangians provide solutions to corresponding constrained optimization problems, as in the case of the SVR[12] based on this saddle-point dynamics.

GLMM was first suggested in the Everett work[13] and then extensively developed in the Gould and Nakayama works [14], primarily to reduce the duality gap between primal and dual issues in non-convex optimization.

Many approaches for constrained optimization have been proposed throughout the years, including penalty function methods and Augmented Lagrangian [15]. However, no comprehensive framework for these strategies has been proposed. With some relaxed conditions, the GLMM could be useful.

Recently, suggested a unique smooth saddle-point dynamics as a fast provable convergent method [16] that assures the constraints and positivity of the Lagrange multipliers without using projections. It has a concept that is very similar to the GLMM.

In recent years, distributed optimization has become one of the most popular study subjects [17]. Consensus protocols, which have also been extensively explored [18], connect centralized and distributed algorithms. Classic Lagrangian [19] is closely related to the linear consensus protocol.

Since convergence performance is affected by the consensus protocols between agents, they are not restricted to the linear type. As a result, it is important to reintroduce the GLMM.

The underlying idea of this thesis is to present an approach to SVR, first developed by Vladimir Vapnik, adding an extended Lagrangian function that includes a weighted elastic net regularization structure, which enables to perform support vector selection and also reduces the influence of correlated support vectors at once.

[11] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004

[12] Diego Feijer and Fernando Paganini. Stability of primal-dual gradient dynamics and applications to network optimization. *Automatica*, 46(12):1974–1981, 2010; and Peng Yi, Yiguang Hong, and Feng Liu. Distributed gradient algorithm for constrained optimization with application to load sharing in power systems. *Systems & Control Letters*, 83:45–52, 2015

[13] Hugh Everett III. Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations research*, 11(3):399–417, 1963

[14] FJ Gould. Extensions of lagrange multipliers in nonlinear programming. *SIAM Journal on Applied Mathematics*, 17 (6):1280–1297, 1969; and H Nakayama, H Sayama, and Y Sawaragi. A generalized lagrangian function and multiplier method. *Journal of Optimization Theory and Apps.*, 17(3):211–227, 1975

[15] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011

[16] Hans-Bernd Dürr, Chen Zeng, and Christian Ebenbauer. Saddle point seeking for convex optimization problems. *IFAC Proceedings Volumes*, 46(23):540–545, 2013

[17] Peng Yi, Yiguang Hong, and Feng Liu. Distributed gradient algorithm for constrained optimization with application to load sharing in power systems. *Systems & Control Letters*, 83:45–52, 2015; Peng Yi, Yiguang Hong, and Feng Liu. Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems. *Automatica*, 74:259–269, 2016; and

[18] Dongkun Han, Graziano Chesi, and Yeung Sam Hung. Robust consensus for a class of uncertain multi-agent dynamical systems. *IEEE Transactions on Industrial Informatics*, 9(1):306–312, 2012; and Yu Zhao, Yongfang Liu, Zhongkui Li, and Zhisheng Duan. Distributed average tracking for multiple signals generated by linear dynamical systems: An edge-based framework. *Automatica*, 75:158–166, 2017

[19] Peng Yi, Yiguang Hong, and Feng Liu. Distributed gradient algorithm for constrained optimization with application to load sharing in power systems. *Systems & Control Letters*, 83:45–52, 2015; and Peng Yi, Yiguang Hong, and Feng Liu. Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems. *Automatica*, 74:259–269, 2016

2.3 Norms

A norm is a function from a real to a complex vector space to the non-negative real numbers such that for every vector in the space, there exists a unique real number called the norm of that vector[20].

Given a vector space $\mathcal{V}$ over a subfield $\mathcal{J}$ of the complex numbers $\mathbb{C}$, a norm on $\mathcal{V}$ is a real-valued function $p : \mathcal{V} \to \mathbb{R}$ with the following properties, where $|s|$ denotes the usual absolute value of a scalar s :

[20] E. Prugovecki. *Quantum Mechanics in Hilbert Space*. ISSN. Elsevier Science, 1982. ISBN 9780080874081. URL https://books.google.com.mx/books?id=GxmQxn2PF3IC

1. Subadditivity/Triangle inequality: $p(x+y) \leq p(x) + p(y)$ for all $x, y \in \mathcal{V}$
2. Absolute homogeneity: $p(sx) = |s|p(x)$ for all $x \in \mathcal{V}$ and all scalars s.
3. Positive definiteness/Point-separating: for all $x \in \mathcal{V}$, if $p(x) = 0$ then $x = 0$.

Due to property 2. implying $p(0) = 0$, some authors replace property 3 . with the equivalent condition: for all $x \in \mathcal{V}$, $p(x) = 0$ if only if $x = 0$. Considering $p \in \mathbb{N}, p \geq 1$, the p the root of the sum (or integral) of the p the-powers of the absolute values of the vector components gives the p-norm on suitable real vector spaces, defined as follows.

$$\|\mathbf{x}\|_p := \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} \tag{2.1}$$

For $p = 1$, the $p - norm$ is the Absolute-value norm, which is a norm on the one-dimensional vector spaces formed by the real or complex numbers.

$$\|\mathbf{x}\|_1 := \sum_{k=1}^{n} |x_k| \tag{2.2}$$

This norm 1 is also known as the $L_1 norm$.

For $p = 2$, the $p - norm$ is the standard Euclidean norm, which gives the ordinary distance from the origin to the point x.

$$\|\mathbf{x}\|_2 := \left(\sum_{k=1}^{n} |x_k|^2 \right)^{1/2} \tag{2.3}$$

This norm 2 is also known as the $L_2 norm$.

2.4 L_1 Regularization

LASSO regularization follows the representation:

$$\sum_{k=1}^{N}\left(y_k - w^T\varphi(x_k) - b\right)^2 - \lambda\sum_{k=1}^{M}|w_k| \tag{2.4}$$

Or in terms of the norm

$$\sum_{k=1}^{N}\left(y_k - w^T\varphi(x_k) - b\right)^2 - \lambda||w_k||_1 \tag{2.5}$$

LASSO is a regularization that only penalizes the positions far away from the training data points, which are the high coefficients. The original LASSO was proposed by Rubin and Scheinberg (1988) as a supervised learning algorithm. It only uses the $|w|$ (modulus) and $|b|$ (bias) to determine the optimal coefficients w and b, which minimize the regularized objective function given, instead of squares of w, as its penalty, LASSO is known as the L_1 norm. It has the effect of forcing the coefficients of the predictors to tend to zero. This means when the independent variables have a linear relationship with the response variable, and then more variables can be used to predict the response variable better

2.5 L_2 Regularization

RIDGE regularization follows the representation:

$$\min_{w}\sum_{k=1}^{N}\left(y_k - w^T\varphi(x_k) - b\right)^2 - \frac{\lambda}{2}\sum_{k=1}^{M}w_k^2 \tag{2.6}$$

Or in terms of the Euclidean norm:

$$\min_{w}\sum_{k=1}^{N}\left(y_k - w^T\varphi(x_k) - b\right)^2 - \frac{\lambda}{2}||w||_2^2 \tag{2.7}$$

RIDGE is a regularization where points are moved to a neighboring grid point if it is closer or added if it is further away. The coefficients are estimated by minimizing the Euclidean distance between each point and its regularized grid. RIDGE was proposed by Jean-Marie Hullot in 1981 and is used to solve both elliptic PDEs and practical problems involving large linear systems, e.g., finding the point with the largest absolute residual in an undetermined system.

RIDGE is known as the L_2 method because it is an $L_2 - norm$ regularizer.

This method has the effect of moving the points to points that are closer to the original data.

2.6 Classical Support Vector Regression

For the case let the set $D = (x_1, y_1), \ldots, (x_N, y_N)$, where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}$. Let $\varphi : X \to \mathcal{F}$ be the function that makes each input point x correspond to a point in the feature space $\mathcal{F}$, where $\mathcal{F}$ is a Hilbert space. This feature space can be of high dimension or even infinite. However, is common to define $X = \mathbb{R}^n$ and $\mathcal{F} = \mathbb{R}^m$. In this form, the approximating function, namely the model, has the form $\hat{y}_k = f(x_k) = w^T \varphi(x_k) + b$ with $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$.

The following problem statement considers such a regression problem as a convex optimization problem.

$$
\begin{aligned}
\min_{w,b,\xi,\xi^*} \mathcal{P}_\epsilon\left(w, b, \xi, \xi^*\right) &= \frac{1}{2} w^T w + C \sum_{k=1}^{N} \left(\xi_k^p + \xi_k^{*p}\right) \\
\text{s.t. } y_k - w^T \varphi\left(x_k\right) - b &\leq \epsilon + \xi_k, \quad k = 1, \ldots, N \\
w^T \varphi\left(x_k\right) + b - y_k &\leq \epsilon + \xi_k^*, \quad k = 1, \ldots, N \\
\xi_k, \xi_k^* &\geq 0, \quad k = 1, \ldots, N
\end{aligned}
\tag{2.8}
$$

where $\varphi(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ and the regularization parameter $C > 0$ determines the balance between the regularity of f and the quantity up to which we tolerate deviations more significant than ϵ. Consider ξ_k and ξ_k^* as slack variables that control the error between the prediction $\hat{y}_k$ and the k-th sample y_k. The number p is either 1 or 2. If $p = 1$, the support vector regressor is called L_1 *soft-margin support vector regressor (L_1 SVR)* and $p = 2$, the L_2 *soft-margin support vector regressor (L_2 SVR)* [21].

[21] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7

Remark 1 *For the present work, only the case L_1 will be considered since it can be easily proven that for the aim of this paper, the L_2 provides an equivalent result.*

Theorem 1 *The primal problem (2.8) with the Lagrangian $\mathcal{L}\left(w, b, \xi, \xi^*; \alpha, \alpha^*, \eta, \eta^*, \mu\right) = \frac{1}{2} w^T w + C \sum_{k=1}^N \left(\xi_k + \xi_k^*\right) - \sum_{k=1}^N \alpha_k \left(\epsilon + \xi_k - y_k + w^T\ \varphi(x_k) + b\right) - \sum_{i=k}^N \alpha_k^* \left(\epsilon + \xi_k^* + y_k - w^T \varphi(x_k) - b\right) - \sum_{k=1}^N \eta_k \xi_k - \sum_{i=k}^N \eta_k^* \xi_k^* - \sum_{i=k}^N \mu_k\,()$, with $\alpha_k, \alpha_k^*, \eta_k, \eta_k^* \geq 0$ results in the following dual problem:*

$$
\begin{aligned}
\max_{\alpha_k, \alpha_k^*} \mathcal{D}\left(\alpha, \alpha^*\right) = &-\frac{1}{2} \sum_{k,l=1}^{N} (\alpha_k - \alpha_k^*)(\alpha_l - \alpha_l^*) \varphi^T(x_k) \varphi(x_l) \\
&+ \sum_{k=1}^{N} (\alpha_k - \alpha_k^*) y_k - \epsilon \sum_{k=1}^{N} \left(\alpha_k + \alpha_k^*\right) \\
s.t.\ &\sum_{k=1}^{N} \left(\alpha_k - \alpha_k^*\right) = 0 \\
&\alpha_k, \alpha_k^* \in \left[0, C\right], k = 1, \ldots, N
\end{aligned}
\tag{2.9}
$$

Proof 1 *See Suykens et. al.*[22] *and Abe*[23].

[22] Johan A K Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002. ISBN 9789812381514. URL https://www.worldscientific.com/worldscibooks/10.1142/5089

[23] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7

Defining $\beta_k = \alpha_k - \alpha_k^*$. Then, $\beta_k \in [-C, C]$ Similarly, defining $|\beta_k| = \alpha_k + \alpha_k^*$, where $|\beta_k| \in [0, C]$. Reformulating the dual problem in terms of β_k in a matrix form:

$$
\begin{aligned}
\max_{\beta} \mathcal{D}(\beta) = -\frac{1}{2}\beta^T K \beta + y^T \beta - \epsilon \|\beta\|_1 \\
\text{s.t. } \beta^T 1_v = 0 \\
|\beta| \preceq C
\end{aligned}
\tag{2.10}
$$

Remark 2 *The equation (2.10) shows the connection between the LASSO and the SVR due to the appearance of a term with the L_1 norm* [24].

[24] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7; and Xixuan Han and Line Clemmensen. On weighted support vector regression. *Quality and Reliability Engineering International*, 30(6):891–903, 2014. DOI: https://doi.org/10.1002/qre.1654

2.7 *A GLMM for the L_1^ϵ-SVR*

To propose a new type of ϵ-SVR, consider the primal problem (2.8) with the following Lagrangian based on the generalized Lagrange multiplier method (GLMM) [25]:

[25] Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66 (2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085

$$
\begin{aligned}
\mathcal{L}(w, b, \xi_k, \xi_k^*; \alpha_k, \alpha_k^*, \eta_k, \eta_k^*) = \frac{1}{2} w^T w + C \sum_{k=1}^{N} (\xi_k + \xi_k^*) \\
- \sum_{k=1}^{N} \alpha_k \left(\xi_k - y_k + w^T \varphi(x_k) + b \right) \\
- \sum_{i=k}^{N} \alpha_k^* \left(\xi_k^* + y_k - w^T \varphi(x_k) - b \right) \\
- \sum_{k=1}^{N} \eta_k \xi_k - \sum_{i=k}^{N} \eta_k^* \xi_k^* \\
- \lambda \left[(1-\epsilon) \sum_{k=1}^{N} (\alpha_k + \alpha_k^*) + \frac{\epsilon}{2} \sum_{k=1}^{N} (\alpha_k + \alpha_k^*)^2 \right]
\end{aligned}
\tag{2.11}
$$

Proposition 1 *The function (2.11) fulfills all the conditions of the GLMM* [26].

[26] Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66 (2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085

Proof 2 *The proof follows directly from the definition; see* [27].

Theorem 2 *The primal problem (2.8) with the Lagrangian (2.11) leads to the*

[27] Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66 (2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085

following dual problem:

$$
\begin{aligned}
\max_{\alpha_k,\alpha_k^*} & \mathcal{D}\left(\alpha,\alpha^*\right) = \\
& -\frac{1}{2}\sum_{k,l=1}^{N}(\alpha_k-\alpha_k^*)(\alpha_l-\alpha_l^*)\varphi^T(x_k)\varphi(x_l) \\
& +\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)y_k \\
& -\lambda\left[(1-\epsilon)\sum_{k=1}^{N}(\alpha_k+\alpha_k^*)+\frac{\epsilon}{2}\sum_{k=1}^{N}(\alpha_k+\alpha_k^*)^2\right] \\
s.t. & \sum_{k=1}^{N}(\alpha_k-\alpha_k^*)=0 \\
& \alpha_k,\alpha_k^*\in[0,C]\,,k=1,\ldots,N.
\end{aligned}
\tag{2.12}
$$

Proof 3 *The proof follows from the stationary conditions:*

- *The first order condition on the parameter w, $\nabla_w\mathcal{L}(w,b,\xi_k,\xi_k^*;\alpha_k,\alpha_k^*,\eta_k,\eta_k^*)=0$, implies $w=\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)\varphi(x_k)$.*
- *The first order condition on the parameter b, $\frac{\partial}{\partial b}\mathcal{L}(w,b,\xi_k,\xi_k^*;\alpha_k,\alpha_k^*,\eta_k,\eta_k^*)=0$, implies $\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)=0$.*
- *The first order condition on the parameter ξ_k, $\frac{\partial}{\partial\xi_k}\mathcal{L}(w,b,\xi_k,\xi_k^*;\alpha_k,\alpha_k^*,\eta_k,\eta_k^*)=0$, implies $\alpha_k+\eta_k=C$*
- *The first order condition on the parameter ξ_k^*, $\frac{\partial}{\partial\xi_k^*}\mathcal{L}(w,b,\xi_k,\xi_k^*;\alpha_k,\alpha_k^*,\eta_k,\eta_k^*)=0$, implies $\alpha_k^*+\eta_k^*=C$*

Then, replacing these critical points in the Lagrangian (2.11).

Besides, the optimal solution must satisfy the Karush Kuhn Tucker (KKT) complementary slackness conditions:

$$\alpha_k\left(\epsilon+\xi_k-y_k+w^T\varphi(x_k)+b\right)=0 \tag{2.13}$$

$$\alpha_k^*\left(\epsilon+\xi_k^*+y_k-w^T\varphi(x_k)-b\right)=0 \tag{2.14}$$

$$\eta_k\xi_k=(C-\alpha_k)\xi_k=0 \tag{2.15}$$

$$\eta_k^*\xi_k^*=(C-\alpha_k^*)\xi_k^*=0. \tag{2.16}$$

Hence, using the complementary slackness conditions, it follows the calculation of b:

$$
\begin{aligned}
b =& y_k-w^T\varphi(x_k)-\epsilon,\ \text{such that} \\
& \alpha_k\in(0,C)
\end{aligned}
\tag{2.17}
$$

Finally, defining $\beta_k=\alpha_k-\alpha_k^*$. Then, $\beta_k\in[-C,C]$ Similarly, defining $|\beta_k|=\alpha_k+\alpha_k^*$, where $|\beta_k|\in[0,C]$. Reformulating the dual problem in terms of β_k in a matrix form:

$$\begin{aligned}\max_{\beta} \mathcal{D}(\beta) = &-\frac{1}{2}\beta^T K \beta + y^T \beta \\ &- \lambda \left[(1-\epsilon)\|\beta\|_1 + \frac{\epsilon}{2}\|\beta\|_2^2\right] \\ \text{s.t. } &\beta^T 1_v = 0 \\ &|\beta| \preceq C\end{aligned} \tag{2.18}$$

Remark 3 *It is shown in (2.18) the connection between the LASSO, the Ridge, and the L_1^ϵ-SVR due to the appearance of a term with the L_1 norm and a squared term with the L_2 norm. This is enough to show that the L_1^ϵ-SVR is in nature a LASSO problem. This new proposal of ϵ-SVR based on the L_1^ϵ-SVR offers a new structure that proposes an Elastic net regularization keeping the box constraints where $0 \leq \alpha_k, \alpha_k^* \leq C$ which makes easier to calculate the b parameter.*[28]

Remark 4 *In the dual problem (2.18), if $\epsilon = 0$ and $\lambda > 0$, the original formulation (2.9) is recovered. This implies that the solution of (2.9) is a lower bound of the solution of (2.18) i.e., when tuning the hyper-parameters, the worst case scenario for (2.18) is (2.9).*

[28] Sara Eugenia Rodríguez-Reyes, Pablo Benavides-Herrera, Gregorio Alberto Álvarez-Álvarez, Riemann Ruiz-Cruz, and Juan Diego Sánchez-Torres. An extended Lagrangian approach to support vector regression based on the MAPE loss. In *20th Mexican International Conference on Artificial Intelligence (MICAI 2021)*, oct 2021

3 Main Results

Contents

3.1 SVR with Symmetric conditions

The main proposal of this thesis is to introduce the SVR with Symmetric conditions model.

Let the set $D = (x_1, y_1), \ldots, (x_N, y_N)$, where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}$. Let $\varphi : X \to \mathcal{F}$ be the function that makes each input point x correspond to a point in the feature space $\mathcal{F}$, where $\mathcal{F}$ is a Hilbert space [1]. This feature space can be of high dimension or even infinite. However, is common to define $X = \mathbb{R}^n$ and $\mathcal{F} = \mathbb{R}^m$. In this form, the approximating function, namely the model, has the form $\hat{y}_k = f(x_k) = w^T \varphi(x_k) + b$ with $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$.

[1] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7

Consider the following optimization problem:

$$\min_{w,b,\xi,\xi^*} \mathcal{P}_\epsilon \left(w, b, \xi, \xi^*\right) = \frac{1}{2} w^T w + C \sum_{k=1}^{N} \left(\xi_k + \xi_k^*\right) \tag{3.1.1}$$

s.t.

$$\begin{aligned} y_k - w^T \varphi\left(x_k\right) - b &\leq \epsilon + \xi_k, \quad k = 1, \ldots, N \\ w^T \varphi\left(x_k\right) + b - y_k &\leq \epsilon + \xi_k^*, \quad k = 1, \ldots, N \\ w^T \varphi\left(x_k\right) &= a w^T \varphi\left(-x_k\right), \quad a \in \{-1, 1\} \\ \xi_k, \xi_k^* &\geq 0, \quad k = 1, \ldots, N \end{aligned} \tag{3.1.2}$$

where $\varphi(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ and the regularization parameter $C > 0$ determines the balance between the regularity of f and the quantity up to which we tolerate deviations more significant than ϵ. Consider ξ_k and ξ_k^* as slack variables that control the error between the prediction $\hat{y}_k$ and the k-th sample y_k.

The constraint $w^T \varphi\left(x_k\right) = a w^T \varphi\left(-x_k\right)$ helps imposing symmetry features in the function f.

For the primal problem (3.1.1), consider the Lagrangian

$$\begin{aligned}
\mathcal{L}\left(w, b, \xi, \xi^{*}; \alpha, \alpha^{*}, \eta, \eta^{*}, \mu\right) = \frac{1}{2} w^{T} w + C \sum_{k=1}^{N}\left(\xi_{k}+\xi_{k}^{*}\right) \\
- \sum_{k=1}^{N} \alpha_{k}\left(\epsilon+\xi_{k}-y_{k}+w^{T}\ \varphi(x_{k})+b\right) \\
- \sum_{i=k}^{N} \alpha_{k}^{*}\left(\epsilon+\xi_{k}^{*}+y_{k}-w^{T} \varphi(x_{k})-b\right) \\
- \sum_{k=1}^{N} \eta_{k} \xi_{k} - \sum_{i=k}^{N} \eta_{k}^{*} \xi_{k}^{*} \\
- \sum_{k=1}^{N} \mu_{k}\left(w^{T} \varphi\left(x_{k}\right) - a w^{T} \varphi\left(-x_{k}\right)\right)
\end{aligned} \tag{3.1.3}$$

with $\alpha_k, \alpha_k^*, \eta_k, \eta_k^* \geq 0$ and $\mu_k \in \mathbb{R}$, Lagrange multipliers.

Remark 5 *The primal problem (3.1.1) with the Lagrangian (3.1.3) results in a dual problem that contains inner products of the form $\varphi^T(x_k)\varphi(x_l)$. The kernel trick allows writing those products as kernel functions $K(x_k, x_l) = \varphi^T(x_k)\varphi(x_l)$.*

Assumption 1 *To impose the constraint $w^T\varphi(x_k) = aw^T\varphi(-x_k)$, it will be assumed the use of kernels which fulfill the following symmetry conditions:*

1. $K(-x_k, x_l) = K(x_k, -x_l)$
2. $K(-x_k, -x_l) = K(x_k, x_l)$

Having established the necessary elements, the following theorem provides the dual optimization problem that relates the primal problem (3.1.1) with the Lagrangian (3.1.3).

Theorem 3 *Under the Assumption 1, the primal problem (3.1.1) with the Lagrangian (3.1.3) results in the following dual problem:*

$$\begin{aligned}
\max_{\alpha_k, \alpha_k^*} \mathcal{D}\left(\alpha, \alpha^{*}\right) = \\
-\frac{1}{2} \sum_{k,l=1}^{N} (\alpha_{k}-\alpha_{k}^{*})(\alpha_{l}-\alpha_{l}^{*})\left(K\left(x_{k}, x_{l}\right)+a K\left(x_{k},-x_{l}\right)\right) \\
+\sum_{k=1}^{N}(\alpha_{k}-\alpha_{k}^{*}) y_{k} - \epsilon \sum_{k=1}^{N}\left(\alpha_{k}+\alpha_{k}^{*}\right) \\
s.t. \sum_{k=1}^{N}\left(\alpha_{k}-\alpha_{k}^{*}\right)=0 \\
\alpha_{k}, \alpha_{k}^{*} \in [0, C], k=1, \ldots, N
\end{aligned} \tag{3.1.4}$$

Proof 4 *The proof follows from the stationary conditions:*

- *The first order condition on the parameter* w, $\nabla_w \mathcal{L}(w,b,\xi,\xi^*;\alpha,\alpha^*,\eta,\eta^*,\mu) = 0$, *implies*

$$w = \sum_{k=1}^{N} (\alpha_k - \alpha_k^*)\varphi(x_k) + \sum_{k=1}^{N} \mu_k \left(\varphi(x_k) - a\varphi(-x_k)\right). \tag{3.1.5}$$

- *The first order condition on the parameter* b, $\frac{\partial}{\partial b}\mathcal{L}(w,b,\xi,\xi^*;\alpha,\alpha^*,\eta,\eta^*,\mu) = 0$, *implies*

$$\sum_{k=1}^{N} (\alpha_k - \alpha_k^*) = 0. \tag{3.1.6}$$

- *The first order condition on the parameter* ξ_k, $\frac{\partial}{\partial \xi_k}\mathcal{L}(w,b,\xi,\xi^*;\alpha,\alpha^*,\eta,\eta^*,\mu) = 0$, *implies*

$$\alpha_k + \eta_k = C \tag{3.1.7}$$

  for all k.

- *The first order condition on the parameter* ξ_k^*, $\frac{\partial}{\partial \xi_k^*}\mathcal{L}(w,b,\xi,\xi^*;\alpha,\alpha^*,\eta,\eta^*,\mu) = 0$, *implies*

$$\alpha_k^* + \eta_k^* = C \tag{3.1.8}$$

  for all k.

- *The first order condition on the parameter* μ_k, $\frac{\partial}{\partial \mu}\mathcal{L}(w,b,\xi,\xi^*;\alpha,\alpha^*,\eta,\eta^*,\mu) = 0$, *implies*

$$w^T\varphi(x_k) = aw^T\varphi(-x_k) \tag{3.1.9}$$

  for all k.

  From (3.1.9)

$$\varphi(x_k) = \frac{1}{2}w^T\left(\varphi(x_k) + \varphi(-x_k)\right) \tag{3.1.10}$$

Then, replacing the critical points (3.1.5)*,* (3.1.6)*,* (3.1.7)*,* (3.1.8)*,* (3.1.9)*, and the identity* (3.1.10) *in the Lagrangian* (3.1.3)*, the dual optimization problem* (3.1.4) *follows.*

Defining $\beta_k = \alpha_k - \alpha_k^*$. Then, $\beta_k \in [-C, C]$ and $|\beta_k| = \alpha_k + \alpha_k^*$, where $|\beta_k| \in [0, C]$. Besides, let $\mathcal{K}(x_k, x_l) = \frac{1}{2}\left(K(x_k, x_l) + aK(x_k, -x_l)\right)$. Those previous definitions permits formulating the dual problem (3.1.4) in terms of β_k in a matrix form:

$$\begin{aligned} \max_{\beta} \mathcal{D}(\beta) &= -\frac{1}{2}\beta^T\mathcal{K}\beta + y^T\beta - \epsilon\|\beta\|_1 \\ \text{s.t. } \beta^T 1_v &= 0 \\ |\beta| &\preceq C \end{aligned} \tag{3.1.11}$$

Remark 6 *The equation* (3.1.11) *shows the connection between the LASSO and the SVR due to the appearance of a term with the* L_1 *norm* [2].

[2] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7; and Xixuan Han and Line Clemmensen. On weighted support vector regression. *Quality and Reliability Engineering International*, 30(6):891–903, 2014. DOI: https://doi.org/10.1002/qre.1654

3.2 *A Generalized Lagrangian Formulation of SVR with Symmetric conditions*

To propose a new type of ϵ-SVR, consider the primal problem (3.1.1) with the following Lagrangian based on the generalized Lagrange multiplier method (GLMM) [3], adding an Elastic net regularization term to the SVR with Symmetric conditions formulation[4]:

[3] Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66 (2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085

[4] M. Espinoza, J.A.K. Suykens, and B. De Moor. Imposing symmetry in least squares support vector machines regression. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 5716–5721, 2005a. DOI: 10.1109/CDC.2005.1583074; and M. Espinoza, J.A.K. Suykens, and B. De Moor. Short term chaotic time series prediction using symmetric ls-svm regression. In *International Symposium on Nonlinear Theory and its Applications (NOLTA2005)*, pages 606–609, October 2005b. DOI: 10.34385/proc.40.3-4-3-1

$$\begin{aligned}
\mathcal{L}\left(w,b,\xi,\xi^*;\alpha,\alpha^*,\eta,\eta^*,\mu\right) = {} & \frac{1}{2}w^T w + C\sum_{k=1}^{N}\left(\xi_k+\xi_k^*\right) \\
& -\sum_{k=1}^{N}\alpha_k\left(\epsilon+\xi_k-y_k+w^T\ \varphi(x_k)+b\right) \\
& -\sum_{i=k}^{N}\alpha_k^*\left(\epsilon+\xi_k^*+y_k-w^T\varphi(x_k)-b\right) \\
& -\sum_{k=1}^{N}\eta_k\xi_k-\sum_{i=k}^{N}\eta_k^*\xi_k^* \\
& -\lambda\left[(1-\epsilon)\sum_{k=1}^{N}\left(\alpha_k+\alpha_k^*\right)+\frac{\epsilon}{2}\sum_{k=1}^{N}\left(\alpha_k+\alpha_k^*\right)^2\right] \\
& -\sum_{k=1}^{N}\mu_k\left(w^T\varphi\left(x_k\right)-aw^T\varphi\left(-x_k\right)\right)
\end{aligned} \tag{3.2.1}$$

Proposition 2 *The function (3.2.1) fulfills all the conditions of the GLMM* [5].

[5] Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66 (2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085

Proof 5 *The proof follows directly from the definition; see* [6].

[6] Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66 (2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085

Theorem 4 *The primal problem (3.1.1) with the Lagrangian (3.2.1) leads to the following dual problem:*

$$\begin{aligned}
& \max_{\alpha_k,\alpha_k^*}\mathcal{D}\left(\alpha,\alpha^*\right) = \\
& \quad -\frac{1}{4}\sum_{k,l=1}^{N}(\alpha_k-\alpha_k^*)(\alpha_l-\alpha_l^*)\left(K\left(x_k,x_l\right)+aK\left(x_k,-x_l\right)\right) \\
& \quad +\sum_{k=1}^{N}(\alpha_k-\alpha_k^*)y_k-\epsilon\sum_{k=1}^{N}\left(\alpha_k+\alpha_k^*\right) \\
& \quad -\lambda\left[(1-\epsilon)\sum_{k=1}^{N}\left(\alpha_k+\alpha_k^*\right)+\frac{\epsilon}{2}\sum_{k=1}^{N}\left(\alpha_k+\alpha_k^*\right)^2\right] \\
& \quad s.t.\ \sum_{k=1}^{N}\left(\alpha_k-\alpha_k^*\right)=0 \\
& \quad \alpha_k,\alpha_k^*\in[0,C],k=1,\ldots,N
\end{aligned} \tag{3.2.2}$$

Proof 6 *The proof follows from the stationary conditions:*

- *The first order condition on the parameter w, $\nabla_w\mathcal{L}\left(w,b,\xi,\xi^*;\alpha,\alpha^*,\eta,\eta^*,\mu\right) =$*

0, *implies*

$$w = \sum_{k=1}^{N} (\alpha_k - \alpha_k^*) \varphi(x_k) + \sum_{k=1}^{N} \mu_k \left(\varphi(x_k) - a\varphi(-x_k) \right). \tag{3.2.3}$$

- *The first order condition on the parameter* b, $\frac{\partial}{\partial b} \mathcal{L} \left(w, b, \xi, \xi^*; \alpha, \alpha^*, \eta, \eta^*, \mu \right) =$ 0, *implies*

$$\sum_{k=1}^{N} (\alpha_k - \alpha_k^*) = 0. \tag{3.2.4}$$

- *The first order condition on the parameter* ξ_k, $\frac{\partial}{\partial \xi_k} \mathcal{L} \left(w, b, \xi, \xi^*; \alpha, \alpha^*, \eta, \eta^*, \mu \right) =$ 0, *implies*

$$\alpha_k + \eta_k = C \tag{3.2.5}$$

for all k.

- *The first order condition on the parameter* ξ_k^*, $\frac{\partial}{\partial \xi_k^*} \mathcal{L} \left(w, b, \xi, \xi^*; \alpha, \alpha^*, \eta, \eta^*, \mu \right) =$ 0, *implies*

$$\alpha_k^* + \eta_k^* = C \tag{3.2.6}$$

for all k.

- *The first order condition on the parameter* μ_k, $\frac{\partial}{\partial \mu} \mathcal{L} \left(w, b, \xi, \xi^*; \alpha, \alpha^*, \eta, \eta^*, \mu \right) =$ 0, *implies*

$$w^T \varphi \left(x_k \right) = a w^T \varphi \left(-x_k \right) \tag{3.2.7}$$

for all k.

From (3.2.7)

$$\varphi \left(x_k \right) = \frac{1}{2} w^T \left(\varphi \left(x_k \right) + \varphi \left(-x_k \right) \right) \tag{3.2.8}$$

Then, replacing the critical points (3.1.5), (3.2.4), (3.2.5), (3.2.6), (3.2.7), *and the identity* (3.2.8) *in the Lagrangian* (3.2.1), *the dual optimization problem* (3.2.2) *follows.*

Defining $\beta_k = \alpha_k - \alpha_k^*$. Then, $\beta_k \in [-C, C]$ and $|\beta_k| = \alpha_k + \alpha_k^*$, where $|\beta_k| \in [0, C]$. Besides, let $\mathcal{K} \left(x_k, x_l \right) = \frac{1}{2} \left(K \left(x_k, x_l \right) + a K \left(x_k, -x_l \right) \right)$. Those previous definitions permits formulating the dual problem (3.2.2) in terms of β_k in a matrix form:

$$\begin{aligned} \max_{\beta} \mathcal{D} \left(\beta \right) = & -\frac{1}{2} \beta^T \mathcal{K} \beta + y^T \beta - \epsilon \|\beta\|_1 \\ & - \lambda \left[(1 - \epsilon) \|\beta\|_1 + \frac{\epsilon}{2} \|\beta\|_2^2 \right] \\ \text{s.t. } & \beta^T 1_v = 0 \\ & |\beta| \preceq C \end{aligned} \tag{3.2.9}$$

Where the kernel K:

$$w^T \varphi(x_k) = a w^T \varphi(-x_k) \tag{3.2.10}$$

Which can be represented by the following equation:

$$\bar{K}(x_k, x_j) = \frac{K(x_k, x_j) + aK(x_k, -x_j)}{2} \tag{3.2.11}$$

Remark 7 *The equation (3.2.9) shows the connection between the LASSO and the SVR due to the appearance of a term with the L_1 norm and RIDGE due to the appearance of a term with the L_2 norm [7].*

The equation (3.2.9) is the main result for the symmetric implementation. It will be the base to implement the symmetric kernel for the applications to real datasets in the next chapter.

[7] Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7; and Xixuan Han and Line Clemmensen. On weighted support vector regression. *Quality and Reliability Engineering International*, 30(6):891–903, 2014. DOI: https://doi.org/10.1002/qre.1654

4 Applications to real datasets

Contents

4.1 Symmetric kernel Implementation

The key element to implement is the symmetric kernel matrix from the equation (3.2.7)

$$w^T \varphi(x_k) = a w^T \varphi(-x_k) \tag{4.1.1}$$

Which can be represented by the following equation:

$$\bar{K}\left(x_k, x_j\right) = \frac{K\left(x_k, x_j\right) + aK\left(x_k, -x_j\right)}{2} \tag{4.1.2}$$

The kernel implementation is based in the RBF function this is:

$$\begin{aligned} K\left(x_k, x_j\right) &= e^{-\frac{||x_k - x_j||^2}{\sigma}} \\ K\left(x_k, -x_j\right) &= e^{-\frac{||x_k + x_j||^2}{\sigma}} \end{aligned} \tag{4.1.3}$$

with $\sigma > 0$.

The result of the implementation in Python is:

```
def kernel_sym(self, X, X1, sigma=0.1, a=1):
    xt = X1  # .T.copy()
    n = X.shape[0]
    nt = xt.shape[0]
    K_1 = np.zeros((n, nt))
    for i in range(n):
        for j in range(nt):
            K_1[i, j] = np.exp(-((np.linalg.norm(X[i, :] -
```

```
                                                  xt[j, :])) /
                                    (2*sigma**2)))

        K_2 = np.zeros((n, nt))
        for i in range(n):
            for j in range(nt):
                K_2[i, j] = np.exp(-((np.linalg.norm(X[i, :] +
                                                  xt[j, :])) /
                                    (2*sigma**2)))
        K = K_1 + (a*K_2)
        return (K)
```

Listing 4.1: Symmetric kernel implementation

To test the new model in different data sets, the first step is to implement it in python. The key element for it is to implement the equation (3.2.9):

```
        min_fun = (1/2)*cp.quad_form(beta, K) - y.T @ beta + Ev @ cp.abs(
     beta) +\
            lamda*(((1-Ev) @ cp.abs(beta)) + ((Ev/2) @ beta**2))
        objective = cp.Minimize(min_fun)
        constraints = [A @ beta == b, G @ beta <= h]
```

Listing 4.2: Equation (3.2.9) implementation

The implementation uses the library "CVXPY", to calculations on the matrices.

To test the implementation, I will use the Boston house-price dataset [1], and the Diabetes dataset [2], both obtained from the "sci-kit learn" libraries.

[1] scikit-learn developers. sklearn.datasets.load_boston, 2020a. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html

[2] scikit-learn developers. sklearn.datasets.load_diabetes, 2020b. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html

4.2 SVR with Symmetric conditions using the Boston house-price dataset

The objective is to predict the price of the houses based on different characteristics of the houses in the Boston area.

The dataset includes fourteen variables:

Variable	Description
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built before 1940
DIS	weighted distances to five Boston employment centrsr
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

Table 4.1: Boston variables

To optimize the parameters of the SVR with Symmetric conditions, a Bayesian Optimization was applied, optimizing the mean absolute error (MAE). The parameters used for $a = -1$ (where a is one of the newly introduced hyper-parameter of the model (4.1.1)):

Variable	Value
C	287.7345
ϵ	0.04
γ	0.04
λ	0.01
σ	3
a	-1

Table 4.2: Kernel hyper-parameters with $a = -1$

And for $a = 1$:

Variable	Value
C	272.2192
ϵ	0.04
γ	0.04
λ	0.01
σ	3
a	1

Table 4.3: Kernel hyper-parameters with $a = 1$

Results comparison using different methodologies:

Method	R^2	Adj R^2	MAE	MSE	RMSE	MAPE
Linear Regression	0.8765	0.8648	2.2637	12.8934	3.5907	10.8676
Random Forest	0.8414	0.8265	2.4284	16.5543	4.0687	11.6503
XGBoost Regressor	0.8765	0.8648	2.2637	12.8934	3.5907	10.8676
SVR	0.8936	0.8836	2.0167	11.1064	3.3326	10.0733
SVR with SC a=-1	0.8694	0.8571	2.4035	13.6329	3.6922	11.9569
SVR with SC a=1	0.8375	0.8222	2.6198	16.9625	4.1185	12.6598

Table 4.4: Boston results

The results of the proposed model are the last two of the above table 4.4. As expected, due to the optimization-based in MAE, this metric is where better results were obtained in comparison with the other metrics with $a = -1$.

The R^2 result was 0.8694, better than a random forest but a little worse than the other three, being 1 the best possible result. On the adjusted R^2 the result was 0.8571 also; this one is better than the random forest but below the other models. Similar performance can be seen in the MAE, MSE, and RMSE; on these metrics, the lower result, the better. On the three, the SVR with Symmetric conditions performs better than the random forest. The exception is MAPE, where the performance was the worst of the models.

In conclusion, the proposed model is consistent with the results obtained from traditional models.

4.3 SVR with Symmetric conditions using the Diabetes dataset

The Diabetes dataset includes the following variables:

Variable	Description
age	age in years
sex	sex
bmi	body mass index
bp	average blood pressure
s1	tc, total serum cholesterol
s2	ldl, low-density lipoproteins
s3	hdl, high-density lipoproteins
s4	tch, total cholesterol / HDL
s5	ltg, possibly log of serum triglycerides level
s6	glu, blood sugar level

Table 4.5: Diabetes Variables

To optimize the parameters of the SVR with Symmetric conditions, a Bayesian Optimization was applied, optimizing the mean absolute error (MAE). The parameters used for this comparison were (with $a = -1$):

Variable	Value
C	660.9515
ϵ	1.0
γ	0.15
λ	0.1
σ	3
a	-1

Table 4.6: Kernel hyper-parameters with $a = -1$

And with $a = 1$

Variable	Value
C	671.0136
ϵ	0.0422
γ	.0630
λ	0.0965
σ	1.7711
a	1

Table 4.7: Kernel hyper-parameters with $a = 1$

Results comparison using different methodologies:

Method	R^2	Adj R^2	MAE	MSE	RMSE	MAPE
Linear Regression	0.4464	0.4010	43.6895	2963.3195	54.4363	38.7632
Random Forest	0.4423	0.3966	44.0939	2985.0625	54.6357	39.7712
XGBoost Regressor	0.4668	0.4231	43.2759	2853.8927	53.4218	37.3279
SVR	0.4359	0.3896	43.5409	3019.5069	54.9500	38.2519
SVR with SC a=-1	0.4105	0.3622	43.8599	3155.0965	56.1702	38.3934
SVR with SC a=1	0.0197	-0.0605	58.8974	5247.0945	72.4368	55.0800

Table 4.8: Diabetes results

The results of the proposed model are the last two of the above table 4.8. As expected due to the optimization-based in MAE, this metric is where better results were obtained compared to the other metrics with $a = -1$.

The R^2 result was 0.4105, worse than the other models, being 1 the best possible result. On the adjusted R^2 the result was 0.3622. Also, this result is the worst of the models. Similar performance can be seen in the MSE and RMSE. However, in the MAE and MAPE results the SVR with Symmetric conditions performs better than the random forest.

In conclusion, the proposed model is consistent with the results obtained from traditional models.

5 Conclusions and future work

Contents

5.1 Conclusions of the first trials

The implementation of a new model, as the SVR with Symmetric conditions, is always the start of a learning path. With the given results on the first trials included in this thesis, is clear that there is some consistency in the performance of the SVR with Symmetric conditions.

The results obtained in this work as the mathematical model and its implementation can be used to keep exploring the SVR model, which is still a powerful tool for data science.

The objective of this thesis was to formulate and release the new SVR methodology with the Symmetric kernel, which was accomplished. More work is needed to test different datasets and verify their efficiency.

5.2 Future work

Future work to develop and study the efficiency of the SVR with Symmetric conditions should include more testing with different datasets.

The SVR with Symmetric conditions models had performed at the same level that other classic models; however, more testing is needed to validate if can improve its result; for example, the optimization could be done based on other metrics and see if the performance improves.

Also, the release of a paper on a specialized forum can help the development to face a bigger forum.

This, as mentioned, is just the start for the SVR with Symmetric conditions development.

Bibliography

Shigeo Abe. *Support Vector Machines for Pattern Classification*. Springer, second edition, 2004. ISBN 978-1-84996-097-7.

S. Boyd and L.Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 978-0-521-83378-3.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

Hans-Bernd Dürr, Chen Zeng, and Christian Ebenbauer. Saddle point seeking for convex optimization problems. *IFAC Proceedings Volumes*, 46(23):540–545, 2013.

M. Espinoza, J.A.K. Suykens, and B. De Moor. Imposing symmetry in least squares support vector machines regression. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 5716–5721, 2005a. DOI: 10.1109/CDC.2005.1583074.

M. Espinoza, J.A.K. Suykens, and B. De Moor. Short term chaotic time series prediction using symmetric ls-svm regression. In *International Symposium on Nonlinear Theory and its Applications (NOLTA2005)*, pages 606–609, October 2005b. DOI: 10.34385/proc.40.3-4-3-1.

Hugh Everett III. Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations research*, 11 (3):399–417, 1963.

Diego Feijer and Fernando Paganini. Stability of primal-dual gradient dynamics and applications to network optimization. *Automatica*, 46 (12):1974–1981, 2010.

FJ Gould. Extensions of lagrange multipliers in nonlinear programming. *SIAM Journal on Applied Mathematics*, 17(6):1280–1297, 1969.

Dongkun Han, Graziano Chesi, and Yeung Sam Hung. Robust consensus for a class of uncertain multi-agent dynamical systems. *IEEE Transactions on Industrial Informatics*, 9(1):306–312, 2012.

Xixuan Han and Line Clemmensen. On weighted support vector regression. *Quality and Reliability Engineering International*, 30(6):891–903, 2014. DOI: https://doi.org/10.1002/qre.1654.

Martin Jaggi. An equivalence between the Lasso and support vector machines. *Arxiv*, abs/1303.1152, 2013.

Mengmou Li. Generalized Lagrange multiplier method and KKT conditions with an application to distributed optimization. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66(2):252–256, 2019. DOI: 10.1109/TCSII.2018.2842085.

Julio López, Sebastián Maldonado, and Miguel Carrasco. Double regularization methods for robust feature selection and svm classification via dc programming. *Information Sciences*, 429:377–389, 2018. ISSN 0020-0255.

Jorge J. Moré. The Levenberg-Marquardt algorithm: Implementation and theory. In *Lecture Notes in Mathematics, Berlin Springer Verlag*, volume 630, pages 105–116. 1978. DOI: 10.1007/BFb0067700.

H Nakayama, H Sayama, and Y Sawaragi. A generalized lagrangian function and multiplier method. *Journal of Optimization Theory and Apps.*, 17(3):211–227, 1975.

E. Prugovecki. *Quantum Mechanics in Hilbert Space*. ISSN. Elsevier Science, 1982. ISBN 9780080874081. URL https://books.google.com.mx/books?id=GxmQxn2PF3IC.

Sara Eugenia Rodríguez-Reyes, Pablo Benavides-Herrera, Gregorio Alberto Álvarez-Álvarez, Riemann Ruiz-Cruz, and Juan Diego Sánchez-Torres. An extended Lagrangian approach to support vector regression based on the MAPE loss. In *20th Mexican International Conference on Artificial Intelligence (MICAI 2021)*, oct 2021.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001. ISBN 9780262256933.

scikit-learn developers. sklearn.datasets.load_boston, 2020a. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html.

scikit-learn developers. sklearn.datasets.load_diabetes, 2020b. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511809682.

Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. ISSN 1573-1375. DOI: 10.1023/B:STCO.0000035301.49549.88.

Johan A K Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002. ISBN 9789812381514. URL https://www.worldscientific.com/worldscibooks/10.1142/5089.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288, 1996. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151(3):501–504, 1963.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589–615, 2006.

Peng Yi, Yiguang Hong, and Feng Liu. Distributed gradient algorithm for constrained optimization with application to load sharing in power systems. *Systems & Control Letters*, 83:45–52, 2015.

Peng Yi, Yiguang Hong, and Feng Liu. Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems. *Automatica*, 74:259–269, 2016.

Yu Zhao, Yongfang Liu, Zhongkui Li, and Zhisheng Duan. Distributed average tracking for multiple signals generated by linear dynamical systems: An edge-based framework. *Automatica*, 75:158–166, 2017.

Quan Zhou, Wenlin Chen, Shiji Song, Jacob Gardner, Kilian Weinberger, and Yixin Chen. A reduction of the elastic net to support vector machines with an application to GPU computing. 2015.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x.