2020 4th International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2020)
ISBN: 978-1-60595-674-9

A Regression Method Based on Data Set Classification

Yan ZHANG

School of Computer Science and Technology, East China Normal University, China

Keywords: SVM linear classification, Decision function, Slack variables, SVR regression.

Abstract. A regression method based on data set classification is proposed in this paper, which can be used in the artificial intelligence NLP classification algorithm from SVM to SVR.

Introduction

In theory, SVM(support vector machine) algorithm involves many concepts: margin, support vector, kernel, duality, convex optimization and so on. SVM not only can be used in supervised classification and regression model, but also in unsupervised clustering and anomaly detection. Compared with the popular deep learning (suitable for solving large-scale nonlinear problems), SVM is very good at solving complex nonlinear problems with small or medium-sized training sets. In this paper, a regression analysis method using SVM is introduced. This method based on dataset applies linear separability and non separability of SVM.

SVM's Predecessor: Perceptron

A hypothesis, the training sample x∈Rn, the label y ∈{-1,1}. For linear classifiers:

Parameters: w∈Rn and b∈R

Decision boundary: $w \cdot x+b=0$

When classifying a new point x, the prediction label is sign(wx+b).

When the classification is correct, that is $y(w \cdot x+b)>0$ and loss=0.

When the classification is incorrect, that is $y(w \cdot x+b) \leq 0$ and $loss=y(w \cdot x+b)$.

In case of the classification is correct, y is the label of a point x. If y=1, the prediction value $w \cdot x+b>0$. The classification is 1. If y=1, the prediction value $w \cdot x+b<0$. The classification is -1. $y(w \cdot x+b)>0$ is used to indicate the correct classification. $y(w \cdot x+b)<0$ is used to indicate the incorrect classification

Linear Separable SVM

When two kinds of samples are linearly separable, the perceptron can guarantee to find a solution. These two kinds of samples are completely and correctly distinguished. The solution is not unique. The best solution of linear separable SVM can be found by reducing the number of Solutions through decision functions and constraints.

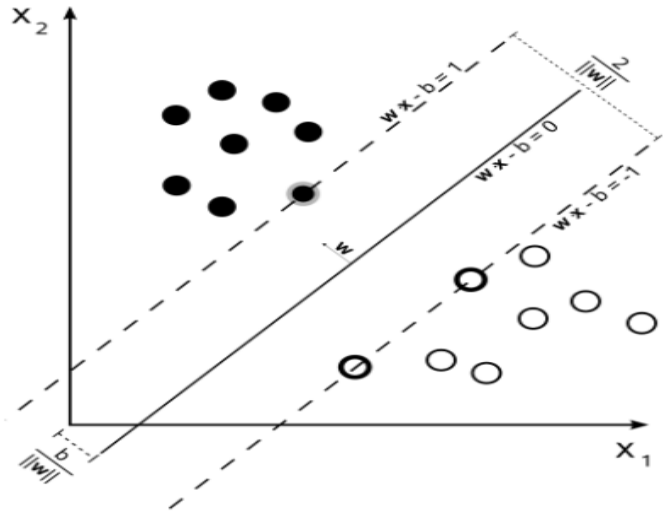

Figure 1. Linear separable SVM.

Decision Boundary and Interval

The decision boundary is defined as w·x+b=0. The boundary (two dashed lines) on both sides are w·x+b=1 and w·x+b= -1. At this point, only the symbol of b is different. The other characters are the same. Among them, w and b are the parameters to be optimized in model training. The following information can be obtained from the above diagram:

The distance between the two dashed lines is $\frac{2}{\|w\|}$.

The direction of the parameter w for optimizing is the normal vector direction of the decision boundary (w is perpendicular to the decision boundary).

At this point, there are three points on the boundary. The three points are the support vector.

In SVM, the goal of optimization is to maximize the width of margin γ. Because $\gamma=\frac{1}{\|w\|}$. ||w|| is the module length of parameter w to be optimized. Therefore, the optimization objective is equivalent to the minimization of||w||. This can be expressed as:

For $(x^{(1)},y^{(1)})..(x^{(m)},y^{(m)}) \in R^d\times\{-1,1\}$, $\min_{w\in R^d,b\in R}\|w\|^2$,

s.t.$y^{(i)}(w\cdot x^{(i)}+b)\geq 1$ is true for all i=1,2,…,m..

Using perceptron and SVM to classify setosa and non setosa in dataset, the margin around the decision boundary determined by SVM is larger. So when classifying more unknown samples, more accurate classification results are got.

SVM Linear Indivisible

Setosa is linear and divisible. There are some overlaps between virginica and its adjacent versicolor. They are linear and indivisible. Still using SVM for classification, the principle is to add a relaxation variable (slack) ξ to the decision function.

For $(x^{(1)},y^{(1)}) .. (x^{(m)},y^{(m)}) \in R^d\times\{-1,1\}$, $\min_{w\in R^d,b\in R}\|w\|^2+C\cdot\sum_i^m\cdot\xi^i$,

s.t.$y^{(i)}$ $(w\cdot x^{(i)}+b)\geq 1-\xi_i$ is true for all i=1,2,...,m.

After adding relaxation variables, Breaking the boundary of both sides (controlled by C in the above formula) to a certain extent is allowed. A certain amount of misclassification is allowed. The two types of data that are originally linear and inseparable are separated.

SVR Uses SVM for Regression Analysis

Support vector regression model (SVR) uses SVM to fit the curve and do regression analysis. In the SVM model, the points on the boundary and inside the two boundaries that violate the margin are regarded as support vectors. The points play a role in the subsequent prediction. According to the dual

form, the final model is a linear combination of all training samples. The weight of other points that are not support vectors is 0.

Here is the graph of soft margin (decision function or relaxation variable) of SVR model:

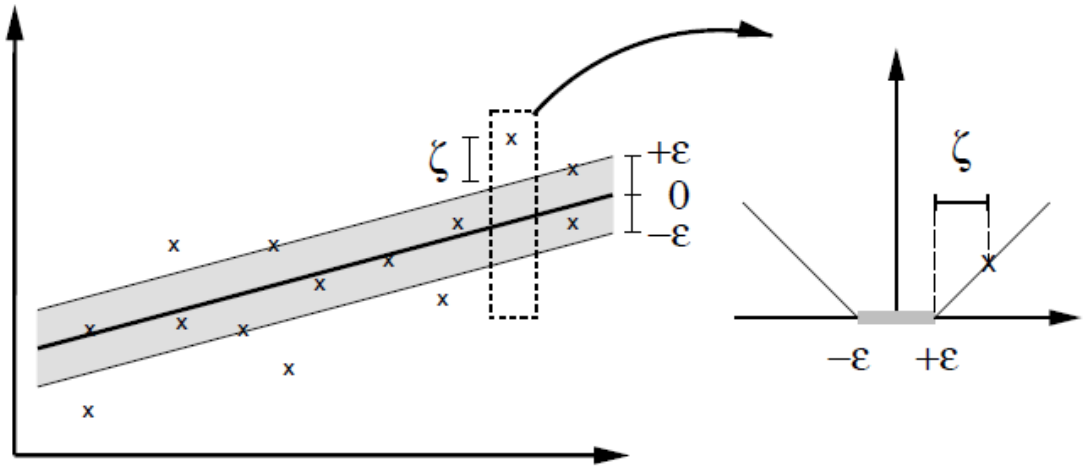

Figure 2. Soft margin of SVR model.

As you can see from the figure, the errors of these points inside the margin are all 0. Only those points beyond the margin will be calculated as errors. Therefore, the task of SVR is to cover as many sample points as possible with a fixed width band (the width is controlled by the parameter ϵ), so that the total error is as small as possible.

Conclusion

In this paper, the perceptron, linear separable and non-separable were analyzed comprehensively. The decision function and relaxation variable were constructed to realize linear classification. The method of regression analysis using SVM was proposed. The method broadened the thinking of regression based on data set classification.

Reference

[1] Fisher, R.A. The uses of multiple measurements in taxonomic problems. Annals of Human Genetics, 7(7),179-188. 1936.

[2] Zhao Chunhui. Algorithm comparison and minimum error interval SVR based on LS-TSVR[D]. Liaocheng University. 2018.

[3] Guo Jian and Liu Quanjing. A location error correction algorithm based on SVR[J]. Computer application research.2017.

[4] Liu Xia and Lu Wei. Application of SVM in text classification[J]. Computer education.2017.1.

[5] Zeng Shuiling and Xu Weihong. Handwritten numeral recognition based on supporting vector machine[J]. Computer and digital engineering.2006

[6] Xiao Jianhua, Wu jinpei and Yang Shuzi. Research on comprehensive evaluation method based on SVM. Computer Engineering[J]. 2002.