Dean&Francis
ISSN 2959-6157

Explanations of Theory and Performance of Support Vector Machine

Haowei Wang

Beijing National Day School, Beijing, China

*Corresponding author:bndscarmelo@hotmail.com

Abstract:

Hitherto, data points are extremely popular in many difference branches of natural science. For this purpose, this is a mathematical article mainly focuses on how to separate or categorize data points efficiently based on their characteristics. The methods are the following. For the linear separable data, the most fundamental Support Vector Machine(SVM) model can be used, while for non-linear separable data, the slack variables and kernel tricks are two efficient techniques. To test whether there is a significant difference among kernels, two kernels, the Gaussian and polynomial kernels, are chosen. The results for those two are quite alike (used a group of data of iris to showcase this). Next, the author applies two kernels to categorize data points of breast cancer. The whole point of this is to separate data in the most efficient way. In this manner, researchers can better predict events, such as injuries or diseases, and take quicker actions. This paper highlights the importance of SVM model in dealing with data-driven problems.

Keywords: Support vector machine; Slack variables; Kernel functions.

1. Introduction

When encountering a large amount of irregular or chaotic data, people often try to identify patterns or characteristics to gain a thorough understanding of the circumstances. For instance, in scientific research, data needs to be clearly categorized. In order to meet this goal, some method must be introduced, that is the support vector machine: a supervised learning algorithm that helps grouping data in a rather efficient way, and they are particularly effective in high-dimensional spaces The core concept of Support Vector Machine (SVM) is about finding a decision boundary that best separates data points of different classes [1].

In the simplest case, which is in two-dimensional space, the goal is to find a linear line that optimally divides data points into two categories. The effectiveness of categorization is determined by maximizing the margin between the boundary lines of each class. When facing more complex scenarios that data points cannot be linearly separated, some strategies such as slack variables and the kernel trick are required. Slack variables have a certain degree of flexibility, allowing some data points to be misclassified. This is especially useful when having noisy data. The kernel trick, on the other hand, transforms data into higher dimensions where it becomes linearly separa-

ISSN 2959-6157

ble. Through this means, the limitations of linear SVMs when dealing with non-linear separable problems can be addressed. Visualization plays a critical role in understanding SVMs [2]. Researchers can create detailed visualizations that reveal the characteristics of the plot through computer programming. Different kernel functions, such as the Gaussian and polynomial kernels, can be applied to various datasets to observe their classification accuracy.

The topic of studying SVMs has much significance. First, researchers can enhance their classification accuracy and achieve a better performance, particularly in high dimensional data. Real-world data is often complex and cannot be linearly separated. Therefore, techniques like kernel tricks allow SVMs to handle non-linear relationships by transforming data into higher dimensional spaces. Such capability is crucial, for it can manage problems that traditional model cannot. SVMs provide flexibility through the use of slack variables and different kernel functions, which allows managing noisy data. The ability to manage noise and adapt to different problems enhances the reliability of SVMs, ensuring the use of it in multiple domains. Visualization of SVM models can provide intuitive and clearer understanding, which is crucial for capturing traits of the model. Using visualization tools, such as those charts graphed by Python programs, helps making results of the model more concise and provides a better overview.

2. Theory of Support Vector Machine

2.1 Basic Principles

First, this article introduces the most fundamental one in SVMs, which is points that can be linearly separated. There are some data points in a two-dimensional space, and the goal is to use a linear line to separate these data points into two categories. After categorization, when new data points are added, this line can be used to determine or predict which category the new points belong to. This line is also known as decision boundary. The thing is how to draw the boundary lines optimally. The greater the distance between the two boundary lines, the higher the efficiency and accuracy. The distance is so called: margin. Imagine the distance between the boundary lines is small (close), and when a new data point is introduced and is close to the decision boundary, the probability of the classification error will be higher. Therefore, finding the optimal decision boundary is equivalent to finding the maximum margin [3].

Moving this line up and down with the same amount of R gives the equations for the two boundaries: $Ax+By+C=r$ and $Ax+By+C=-r$. To find the spacing one can use the formula for the distance of parallel lines $\frac{|C_1-C_2|}{\sqrt{A^2+B^2}}$, and substitute C_1 and C_2 with the numerical values. This gives the distance, $\frac{|C_1-C_2|}{\sqrt{A^2+B^2}}$, where the incongruity arises. The reason for the incongruity is that it contradicts the thought model, because according to the model, if the direction of the decision boundary is determined, the two boundary lines and their intervals are also determined, since three lines are parallel to each other, so the factors affecting the distance should only be related to the direction, that is, A and B, and so there should not be r.

Now one should eliminate the effect of r. The solution is as follows: divide the system of equations by the same amount of r. The original system of equations is changed to $Ax+By+C=1$ and $Ax+By+C=-1$ (Since ABC are just symbols, there is no need to replace them with other symbols after dividing by r.), and at the same time, people can optimize the distance equation as $\frac{2}{\sqrt{A^2+B^2}}$. Now let the first class point (p) is on one side of the first class boundary and the second class point (q) is on the other side of the second class boundary, and now the problem is, the author can only come up with Ax_p+By_p+C+1 and Ax_p+By_p+C-1 as a dissimilarity, and it is impossible to determine which one has the positive or negative sign. So now, in order to eliminate this new incongruity, a new symbol needs to be introduced. Now the author labels the data points on the +1 side as category V_p, and the data points on the -1 side as category V_q Because $sign(Ax_i+By_i+C)$ that is the sign of Vi, after all the processes above, one can easily obtain that $V_i(Ax_i+By_i+C)\geqslant 1$. And now the equation is much more

2.2 Kernel Tricks

In some cases, the data points cannot be linearly separated. In these circumstances, using the kernel tricks is the most efficient way [4].

There are four major kernels. First of all, for the linear kernel,

$$K(x,y)=x\cdot y \tag{1}$$

it represents the simplest form of a kernel function. It simply calculates the dot product of two input vectors. This method is can be used when a simple linear model is suffi-

cient to predict the results.

The second is polynomial kernel,

$$K(x,y)=(x\cdot y+c)d \tag{2}$$

It represents a non-linear transformation of the input space. The purpose of these parameters is to control the influence and flexibility of the model. This can be used for data that is not linearly separable but can be separated with polynomial decision boundaries.

The most common one is Gaussian kernel,

$$K(x,y)=exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right). \tag{3}$$

This function measures the similarity between two points based on their Euclidean distance. The parameter sigma controls the width of the Gaussian function, the smaller the values, the tighter it would be. This method is used to tackle with non-linear relationships among data, and it is effective when the decision boundary is in a curved shape.

The last one is Sigmoid kernel, also known as hyperbolic tangent kernel,

$$K(x,y)=tanh(ax\cdot y+c) \tag{4}$$

It is widely used when the data has characteristics similar to those modeled by neutral networks. However, the sigmoid kernel is less commonly used comparing to the Gaussian RBF and polynomial kernels.

Slack variable is a very useful method when categorizing data points that are non-linear separable. A non-negativity constraint on the slack variable is also added. The aim of slack variable is allowing some data points to violate the SVM's hard interval constraints, which gives the model some fault tolerance. By applying this method, some errors that occur when categorizing can be tolerated. Slack variable is particularly useful when coping with noisy data (datasets) that are often encountered in applications.

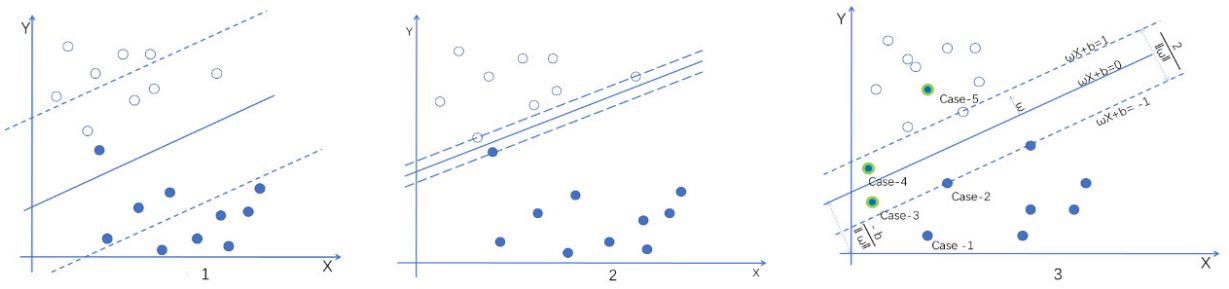

Fig. 1 Different distribution cases of data points in the coordinates.

The different distribution cases of data points in the coordinates are shown in Fig. 1. The margin in the left panel of Fig. 1 is what people want to get, but there is a solid point occurring on the other side of the decision boundary, this is known as the anomaly point. At this time, the method in the middle plot can be used to divide the data points. Moreover, loosen the restriction and use the method of Fig. 1 for categorization is also an alternative option [5]

$$min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i \tag{5}$$

Subject to $y_i(\omega\cdot x_i+b)\geqslant 1-\xi_i$ and $\xi_i\geqslant 0$ ($i=1,2,\cdots,n$).Here, ωmeans the weight vector, b means offset, and C is a regularization parameter that controls the trade-off between the size of the interval and the penalty for violating it. ξ is the slack variable, which represents the degree to which the data point x_i violates the interval constraint.

2.3 Benefits and Visualization

There are several benefits of SVMs. SVM is a very powerful classifier, especially suitable for dealing with high-dimensional data sets and nonlinear problems. Since SVM is optimized to maximize the interval, it has good robustness and can avoid over-fitting problems. Robustness means the ability to maintain good performance). SVM can use different kernel functions to adapt to different problems, for example, linear kernel functions for linear separable problems, Gaussian kernel functions for nonlinear separable problems, etc. SVM is relatively robust to noisy data, and soft intervals can be set to avoid over-reliance on noisy data.

In the realm of computational research, the visualization of Support Vector Machine (SVM) algorithms is an indispensable tool for elucidating the underlying patterns and classifications that are obscured within the algorithm's complex decision-making processes. Visualizations are obviously important in SVMs, as they provide graphical representations of the decision boundaries that the algorithm constructs. These boundaries are critical in understanding how the SVM categorizes data points and can reveal insights into the algorithm's performance. For example, a clear visualization can help identify areas where the SVM model might be over-fitting or under-fitting the data, thereby applying necessary adjustments to the model's parameters [6]. Python can be used to create visualization of SVMs.

ISSN 2959-6157

3. Working with Support Vector Machine

The use of Python for visualization is further buttressed by libraries such as Matplotlib and Seaborn. These libraries not only simplify the process of generating plots but also enable researchers to customize the appearance of these visualizations to suit their specific needs. In this case, the ability to produce publication-ready figures directly from the code shows the importance of Python in the field of data visualization. Take some demand in drawing the graph as examples, 'NumPy' provides support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays, and 'Matplotlib', on the other hand, offers a wide range of visualization options that can be customized to create detailed and informative plots.

The code draws the graph in several steps. First, the script begins by importing the necessary libraries. 'Numpy','matplotlib.pyplot'and 'sklearn' are imported to numerical operations, plotting the data and results and facilitate accessing the dataset. The train_test_split from model_selection is used to divide the dataset into training and testing subsets, and the SVC class from 'SVM' in 'sklearn' is imported to create a support vector classifier. The code then proceeds to load the dataset and to select only the first two features from the dataset for training purposes. The data is split into training and testing sets, with the testing set comprising 30% of the total data. A random state of 42 is set for the train_test_split function, which ensures that the division of data can be reproduced consistently.

Next, an SVM classifier with a radial basis function (RBF) kernel is instantiated. The SVC class is configured with the kernel set to 'RBF', the gamma parameter set to 'scale', and the regularization parameter C set to 1.0. SVM classifier is trained using the training dataset features and labels with the fit method. After training, predictions are made on the test dataset with the predict method. The script then calculates the accuracy of the model predictions by comparing them to the true labels of the test set. The accuracy is the mean of correct predictions and is printed in a formatted string to two decimal places.

In the plotting section, a grid of points is created to visualize the decision boundary of the SVM classifier. The points 'xx', 'yy' form a mesh grid within the range of the features in the dataset. The classifier's predictions for this grid are then reshaped and contour plotted to show the decision boundary. The training data for each class is plotted using different colors for distinction. Finally, the features' names from the dataset are used as label for the x and y axes, and the title of the plot is set to "SVM Classification with RBF Kernel".

The datasets used in the visualizations, such as the data of iris, are well-established benchmarks in the field of machine learning. These datasets are chosen for their diversity and the richness of features they provide, which are essential for demonstrating the versatility of SVM algorithms. In the breast cancer dataset visualization, the color-coded scatter plot allows quicker identification of the decision boundary and the distribution of data points across different classes: some points lie on the wrong side of the boundary, while more points lies on the right side of the boundary. Those points ensure to calculate the accuracy of different plots which has different kernels and different data sets.

When using the Gaussian Kernel to draw a graph of iris data set, the accuracy is 0.8, and for the cancer data set, the accuracy changes to 0.9. The other kernels accuracy also changes: the accuracy of Polynomial Kernel's plot of iris data set is 0.73, and the cancer's data set's one is 0.91. These data content reveal that both Polynomial and Gaussian Kernels achieve high accuracy rates when applied to their respective datasets, while there is no one-size-fits-all kernel solution for SVM. The performance, in other words, the accuracy, of each kernel is correlated to the characteristics of the dataset, including the distribution and nature of the data points. This finding emphasizes the need for empirical testing and comparison of different kernels when applying SVM algorithms. It highlights the importance of understanding datasets at hand and how different kernels might interact with its features to produce the best classification results.

4. Conclusion

In the very beginning, the main purpose of using SVM is mentioned. This article first introduces the basic concept of support vector machine, and then outlines the steps for using this method to categorize data points. Afterwards, two types of techniques, slack variables and kernel tricks, which are quite useful for grouping non-linear separable data points, are presented. This article briefly shows about what 'slack variables' is, and later focuses on two Kernel tricks that are widely used—the Gaussian kernel and polynomial kernel. Through programming, lines can be drawn for separating data of breast cancer, and people can use the model to make further prediction. The plot of categorized iris's data points shows no significant differences among two kernels. Among any generations, disease has always been a non-negligible issue, and people are seeking all means to address it. That is why the author uses these two kernels to separate data of breast cancer later on. This can allow prediction of cancer and ensure taking

quicker therapeutic actions. The most important finding of this work is that it underscores the importance of SVM in solving data-related problems. For this purpose, since the author is interested in sports, especially American football and basketball (both are high-risky), the author wants to use SVM models to predict and decrease the risks, or even better prevent injuries from happening.

References

[1] HUANG, SHUJUN, et al. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics, 2017, 15(1):41–51.

[2] S. Keerthi, D. DeCoste. A modified finite Newton method for fast solution of large scale linear SVMs. Journal of Machine Learning Research 2005, 6: 341–361.

[3] Prasanna, Thushara Haridas, et al. Identification of polar liquids using support vector machine based classification model. IAES International Journal of Artificial Intelligence (IJ-AI), 2022, 11(4): 1507.

[4] Li ling and Wu mingdong. “Research on classification extraction and change of land use in Yanting County based on support vector machine model.” Science and Innovation. 2024, 16:148-150.

[5] Cano Lengua, Miguel Angel and Erik Alex Papa Quiroz. A Systematic Literature Review on Support Vector Machines Applied to Classification. 2020 IEEE Engineering International Research Conference (EIRCON), 2020.

[6] Gaye, Babacar, et al. Improvement of Support Vector Machine Algorithm in Big Data Background. Mathematical Problems in Engineering, 2021, 1–9.