Performance of Support Vector Machines in Classification of Graduates' Salary

Raj Choudhary*
*School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127
Email: raj.choudhary2016@vitstudent.ac.in *

***Abstract*: Support vector machine (SVM) is the another powerful and widely used supervised learning algorithm, which can be considered as an extension of the perceptron. Using the perceptron algorithm, we minimized misclassification errors. However, in SVMs, our optimization objective is to maximize the margin i.e. the distance between the separating hyperplane (decision boundary) and the training samples that are closest to this hyperplane, which are the so-called support vectors.**
***Index Terms*— SVM, Support Vectors, Margins, Kernel, Clustering, Supervised Learning, Slack variable, Non-Linear Classification.**

1. Introduction

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. Additionally, SVM can also be used for Regression analysis.
For some situations where we know that a nonlinear region can separate the groups more efficiently, SVM is very helpful there. It handles this by using a kernel function (nonlinear) to map the data into a different space where a hyperplane (linear) cannot be used to do the separation. It means a non-linear function is learned by a linear learning machine in a high-dimensional feature space while the capacity of the system is controlled by a parameter that does not depend on the dimensionality of the space. This is called ***kernel trick*** which means the kernel function transforms or maps the data into a new higher dimensional feature space to make it possible to perform the linear separation. It then takes the inner product of the new vectors. The image of the inner product of the data is the inner product of the images of the data. Two of the widely used kernel functions are shown below:

Polynomial:

$$k(x_i, x_j) = (x_i . x_j)^d$$

Radial Basis Function (rbf):

$$k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$

2. Methodology

The Aspiring Minds' Employability Outcomes 2015 dataset is collected from the Aspiring Minds official website. The data is cleaned of missing values and anomalous data points. The attributes are made numeric and categorical for the computation of split points by the SVM. Then using the Sklearn's pre-processing library dataset is splitted into train and test dataset. Both sets are created such that they are having only two features viz. 'Domain' and 'agreeableness' on the basis of which the SVM model will learn the categories.
SVM follows 3 simple steps to fit a model. These steps are –

1. Define an optimal hyperplane: by maximizing the width of the margin (w).
2. For non-linearly separable problems: have a penalty term for misclassifications (C).
3. Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.

In certain situations, some points lie across the margin and get misclassified. For this, SVM has a property called ***Slack variable*** which allows some instances to fall of the margin, but penalize them. Thus, it is helpful in better regularisation of the model. The algorithm always tries to maintain the slack variable to zero while maximizing margin. However, it does not minimize the number of misclassifications (NP-complete problem) but the sum of distances from the margin hyperplanes. Other major factors that helps in regularisation are Gamma, tolerance, decision function ('ovo', 'ovr'). Finally using the trained SVM models, the performance is evaluated on the Aspiring Minds' Employability Outcomes 2015 test dataset. It aims to classify the engineering graduate who are fresh out of college based on their salary given the candidates' academic background, standardized test performance and personality test scores.

3. Dataset – AMEO 2015

For every engineer, AMEO dataset provides anonymized bio data information along with their respective skill scores and employment outcome information. Specifically, the following information is available for every engineer: AMEO 2015 has gained traction since its public release. Aspiring Minds annually publishes the National Employability Report, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset.

4. Analysis and Result

For this paper, different SVM models were trained using GridSearchCV technique of sklearn. For this both linear and rbf we set as kernel parameter and for C and gamma the values from the list was taken (0.0001, 0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0). CrossValidation cv parameter was set to 10. This resulted in fitting of 10 folds for each 72 candidates, totaling to 720 tasks. The execution took around 5.4 minutes. From the results, we found that for classification using SVC(), the best results are obtained using 'rbf' kernel and C and gamma values 0.1 and 1.0 respectively. Using these parameters, training accuracy of 37.148% and test accuracy of 35.625 was obtained.
When regression for the salary was performed based on the 'Quant' attribute, SVR() with the same parameter list was used and it took around 40 seconds to complete. From it, we found that the best results are obtained using 'linear' kernel with C = 1000. These parameters resulted in training r2 score of 0.046 and test r2 score of 0.024
The results of the dataset could further be improved by tuning the other hyperparameters of the respective functions.

5. Conclusion

Hence, we have discovered some interesting insights into the usage of SVM model and its parameters. One of the curious observations is the large overlap of data points, as we are still getting a very lesser accuracy of at most 37.148% using such complex models, which will prove to be a challenge during the model development phase. The analytics can also suggest graduate's methods of improving their job prospects.

6. Reference

[1]. V. Aggarwal, S. Srikant, and H. Nisar, "Ameo 2015: A dataset comprising amcat test scores, biodata details and employment outcomes of job seekers," in AMEO 2015: A Dataset Comprising AMCAT Test Scores, Biodata Details and Employment Outcomes of Job Seekers. ACM, 2016.