Machine Learning Assignment

Shubbham Gupta
19205033

The assignment is submitted to University College Dublin
in part fulfillment of the requirements for the module
Foundation of Data Science 1

School of Mathematics and Statistics
University College Dublin

Submitted to: Claire Gormley
30/09/2020

Support Vector Classifier

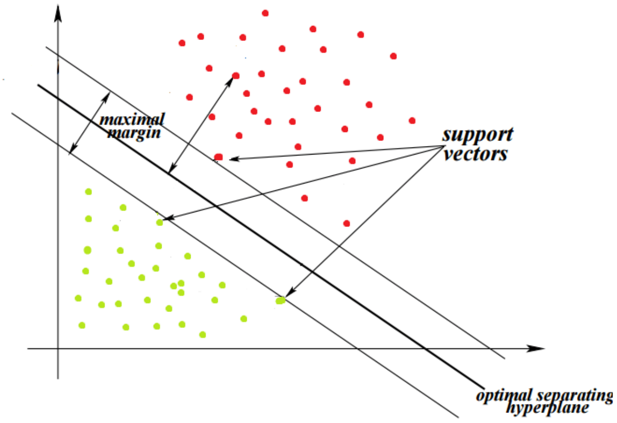
A support vector classifier is a supervised machine learning algorithm that analyzes the data used for classification and regression analysis (Noble. 2006). It learns by assigning labels to objects. For instance, recognizing handwritten digits by analyzing a large collection of scanned images of handwritten numbers. Alternatively, SVCs have been successful in a variety of biological implementations. These applications involve classifying DNA sequence, proteins, microarray gene profiling, mass spectra, and cancer detection.

In essence, an SVC is a mathematical algorithm for maximizing difference among various classes for a given collection of data. SVC method can be explained by following three concepts: (i) the hyperplane, (ii) the kernel function, (iii) the mathematical formulation.

Hyperplane, Support Vector and Margin

A hyperplane is a decision plane that divides observations from different classes or categories in a linearly separable dataset. For a given data of N dimensions, there exists a hyperplane with N-1 dimensions. For instance, if data is an one-dimensional, it will be divided by a single point. A line will act as hyperplane for two-dimensional data.

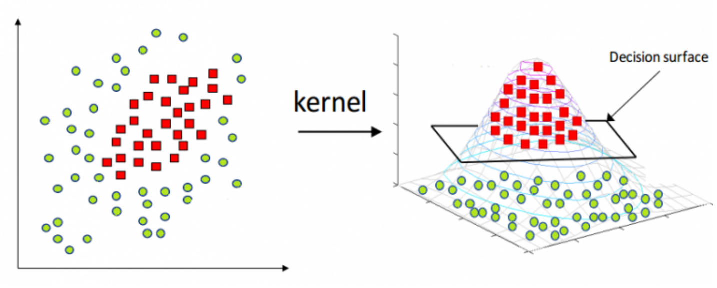
Support vectors are the data points, closest to the hyperplane. These points are necessary for the construction of the classifier as the margin will be calculated using these support vectors. A margin is a perpendicular distance between the hyperplane and the nearest data points. As there may exist infinite hyperplanes between two clusters of classes, calculating margin will provide us with an optimal hyperplane. The optimal hyperplane is the one with the largest margin between classes. The following figure shows the support vectors, margin, and optimal hyperplane.



We have assumed that data can be separated into two clusters easily so far, but it will not be possible when dealing with real-world data. There are some observations that will lie on the wrong side of the hyperplane, it will be marked as an 'error'. If the model is constructed with no error, it will lead to poor classification when tested with previously unseen data. Introducing few anomalous observations to fall on the wrong side of hyperplane will act as a cost function in SVC that will reduce the over-fitting of the data. This cost function will allow us to put some points through hyperplane without affecting the final result. Minimizing the cost function and controlling the number of data points allowed to violate separating hyperplane are some necessities that are needed to keep in mind during the construction of the classifier model.

Kernel Function

Until now, we have assumed that data is linearly separable and there exists at least a single hyperplane separating the different classes. However, the data in study may not always be linearly separable, thus creating difficulty for SVC to work properly. To resolve this issue, kernel functions are introduced in SVC.



The kernel function is a mathematical operation which transforms the data by adding a dimension to the data i.e. it will projects a low dimensional data to a higher-dimensional space. The transformation of two-dimensional to three-dimensional data has been showed in the figure above. The data will become separable in resulting higher dimensional projection if the kernel function is good for the data. The most common kernel functions in SVC are the linear kernel, polynomial kernel, and Radial Basis Function (RBF) kernel.

Mathematical Formulation of Support Vector Classifier

For a training set of N data points $\{y_k, x_k\}_{k=1}^{N}$, where x_k is the k^{th} input and y_k is the k^{th} output, the following equation is used for consturction of SVC (Suykens & Vandewalle 1999).

$$y(x) = sign\left[\sum_{k=1}^{N} \alpha_k y_k \psi(x, x_k) + b\right]$$

where α_k is the positive real constant and b is a real constant. $\psi(x, x_k)$ is the kernel function used for this training set.

Dataset

Data set used here is the Ionosphere dataset taken from UCI Machine Learning Repository. The radar data is collected by a system in Goose Bay, Labrador (Sigillito, V. G. et al. 1989). It measures the efficiency of an ionosphere area by checking which radar signals passes through ionosphere and which signals bounces back by the ionosphere. The signals are divided into two classes i.e. 'good' and 'bad'. The input of the database contains information of 17 pulse numbers for the system which are further described by two attributes corresponding to the complex values returned from the signal in form of a continuous real number.

Comparison of Support Vector Classifier with Random Forest Classifier

To check the implementation of Support Vector Classifier, the ionosphere dataset has been used. Before using the dataset for training, the training subset has been class balanced to improve the model construction. The classifier uses RGB kernel for model construction with regularization parameter of 1 and gamma parameter "scale". Performance of Support Vector Classifier is compared with Random forest Classifier. The model is constructed with the entropy criterion and 200 trees in the forest model. To get an efficient result, both classifiers are validated using 10 fold cross-validation method. Then each model has been tested by running 25 times and the average result has been taken. The following table shows the confusion matrix for both the models.

Prediction outcome of SVC

		good	**bad**	**total**
actual value	**good′**	71	0	71
	bad′	4	41	45
total		75	41	

Prediction outcome of RFC

	good	**bad**	**total**
good′	71	0	71
bad′	6	39	45
	77	39	

From comparing the confusion matrix of both the models, it can be seen that both the models has shown very good result. SVC has correctly classified 112 observations with only 4 misclassifications. While RFC has correctly classified 110 observations out of a total of 116. It can be seen that both the models has correctly classified the 'good' observations and show error in classifying 'bad' observations. It could be the result of class balancing as fabricated observations are created for 'bad' class. Comparing accuracy of the classifiers, it can be seen that SVC has shown an accuracy of 96.5% while RFC has shown an accuracy of 94.8%. The following table shows the skill score matrix for each model.

Skill Score Matrix

Support Vector Classifiers

	precision	**recall**	**f1-score**	**support**
bad	1.00	0.91	0.95	45
good	0.95	1.00	0.97	71

Random Forest Classifiers

	precision	**recall**	**f1-score**	**support**
bad	1.00	0.87	0.93	45
good	0.92	1.00	0.96	71

In the skill score matrix, precision, recall and f1 score are provided for each class in both models. It is developed from the above confusion matrix. Precision is the ratio of correctly classified positive observations to the total number of positive results including wrongly classified. The recall is the ratio of correctly classified positive observations to total positive observations in the dataset. F1-score is a measure of test's accuracy which is calculated by taking a harmonic mean of precision and recall. From the table, it can be seen that SVC has better precision and better recall than RFC in each class. The same has been reflected in the corresponding f1-score of the classifier.

As the dataset used is binary classified i.e. output has two classes, SVC has shown very efficient result. Apart from Random Forest classifier, other classifiers are also used for comparison such as K-nearest neighbour, Decision tree classifier, Naive Bayes Theorem. However, Support Vector classifier has shown better results than all the classifiers mentioned. It can be concluded that support vector classifier is very effective in high dimensional space and situations where several dimensions are greater than the number of samples. Due to its fast execution and efficient result, it can be used in yielding valuable insights into growing quantity and various industries.

Reference

Noble, W.S. (2006), 'What is a support vector machine?', *Nature Biotechnology*, 24(12), 1565-1567, URL: *https://doi.org/10.1038/nbt1206-1565*

Suykens, J. & Vandewalle, J. (1999), 'Least squares support vector machine classifiers', *Neural Processing Letters*, 9(3), 293–300, URL: *https://doi.org/10.1023/A:1018628609742*

Sigillito, V. G., Wing, S. P., Hutton, L. V., & Baker, K. B. (1989), 'Classification of radar returns from the ionosphere using neural networks', *Johns Hopkins APL Technical Digest*, 10(3), 262-266