SVM Classification Challenge 5

Shirish Pandagare

I) INTRODUCTION

Support Vector Machine (SVM) is a supervised learning algorithm that can be used for both regression and classification. SVM is based on the idea of finding a hyperplane that best divides the data into two classes. Unlike the other classifiers, SVM has a margin which separates these two classes. The main aim is to search for the hyperplane such that it maximizes the margin or gap between the classes. The support vectors are the points in the data set which lies on the margin of the hyperplane. Thus, if we change the position of the support vector, the hyperplane will also change. The classes are not always separated by a linear boundary, instead, they can be non-linear. For this, SVM comes with kernel function can be used to determine the non-linear boundaries of the classes.

II) APPROACH

SVC (Support Vector Classification) is used to determine the decision boundary for the given dataset. SVC comes with four different kernel options to be used in the algorithm which are 'linear', 'Polynomial function', 'Radial Basis function' & 'Sigmoid function'. The assessment of the classification can be divided into two different types, i.e. External validation measures and internal validation measure. External validation can be done using the information available with us, i.e. the classes of each points from the training dataset. Jaccard Coefficient and F-measures are used to validate the classification. On the other hand, Internal validation is done using the information derived from the data, which is the predicted class values for each point in the testing dataset. Silhouette Coefficient is used to see how good the testing dataset is clustered.

The four kernels are used to obtain the decision boundary and the assessment is done using aforementioned validation methods. The best results out of the four models was selected as the model for the classification.

III) RESULTS

The three different statistics were considered for each of the model. The value of Jaccard Coefficient and F-measure (normalized) varies from 0 to 1, where 1 being the perfect clustering. As the ground truth of the classes is required to compute these parameters, training dataset is used. Silhouette coefficient helps to measure the separation distance between the obtained cluster. It ranges from -1 to 1, where value close to 1 indicates that the cluster are placed far from each other. As true purpose for this test to evaluate how the unknown dataset are classified, testing dataset was used to evaluate the Silhouette coefficient. These parameters are used to accessed the model and select the best among the four.

The following are the results obtained using the validation techniques mentioned above for the models using different kernels.

Kernel	Jaccard Coefficient	F measure	Silhouette Coefficient
Linear	0.7700	0.7666	0.3033
Polynomial	0.6950	0.6946	0.2765
Radial	0.7750	0.7702	0.2699
Sigmoid	0.5300	0.5298	0.1328

The Radial kernel model has the highest Jaccard and F-Measure score, however, the Linear kernel model has very close values to the Radial kernel. Hence, these two models show the best fit for the given training dataset. On the other hand, Linear kernel model has the highest Silhouette coefficient, which indicates that Linear model is able to separate the testing data more accurately than the other models. The following is the scatter plot along with the linear hyperplane and its margin. It can be seen that the Linear hyperplane is able to separate the classes which validates the choice of our kernel selected.

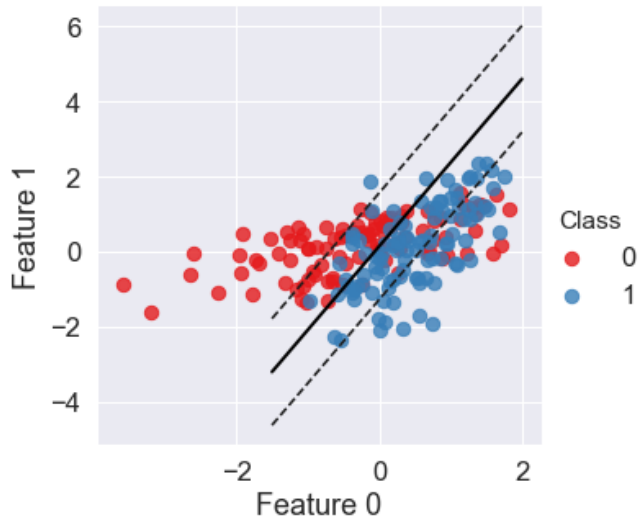

Figure 1: - SVM hyperplane on the training dataset

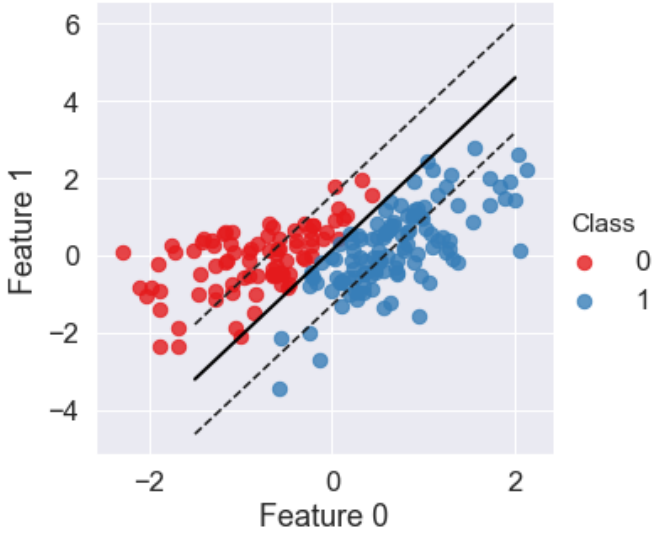

Figure 2:- SVM hyperplane on the testing dataset