Machine Learning Lab 7

Support Vector Machines

Decision A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. Basically support vector machine works on the principle of separation of classes. What SVM does is, it finds out a line/hyper-parameter(in multidimensional space) that separates out classes.

The dataset

The dataset used to perform this experiment is the wine quality dataset, it is a combination of data on two types of wine variants, namely red wine and white wine, of the portuguese “Vinho Verde” wine. The dataset contains information on the parameters for fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.

Experiment

In this experiment I used the sklearn’s support vector classifier (SVC) algorithm to predict the quality of a wine.

Using the pandas library in I loaded the red wine and white wine datasets into the memory from their respective csv files and then merged the two datasets into one single pandas dataframe.

Using the pandas.Dataframe.describe() function in pandas I calculated the various statistical measures of each of the columns of the dataset.

For performing the experiment I started with plotting the scatter plot for each of the features in the dataset with every other feature, this helped to find if there were any features which were linearly separable. In the case of my dataset they were not.

Next I divided the dataset into training and testing portions using the train_test_split functionality in sklearn

I setup the grid search for SVC with RBF(Radial Basis function) kernel and the value of gamma ranging from 1e-4 to 0.5, and the value of parameter C ranging from 1 to 1000. I ran the grid search and for the parameters of C=1, and gamma =0.5, I achieve the best performance with a precision score of 0.84.

Next for visualization purposes I selected two most important features from the dataset. In this case the feature importance was decided based on the feature importance values calculated in the previous lab assignments using random forests.

Thus the two features of fixed acidity and volatile acidity were selected and an SVC was trained on this model and the decision boundaries were predicted on the same.

The code and plots can be found in the accompanying jupyter notebook.