David Favela Corella

Dxf200002

CS 4375.003

Kernel and Ensemble Methods

1. SVM, short for Support Vector Machines, is a popular multi-purpose algorithm. Its most common scenario is binary classification, but it can also perform well in multiclass classification and regression. It works by separating the data into two classes. They are separated by a line called the hyperplane and has a margin of data in between the hyperplane that is sustained by the support vectors. The algorithm can then classify by determining which side of the margin an instance falls on.

SVM kernels are functions used to map the linear function to another form so that the data is linearly separable. This is helpful when a linear model does not fit the data and another type of model is needed. Some of the popular kernels used include polynomial and radial. It is to be noted that different kernels also introduce new parameters to the algorithm. The radial kernel uses gamma; a larger gamma can overfit, having low bias and high variance.

From my point of view, the downsides to the SVM algorithm are few. It is an algorithm that can perform linear and logistic regression, making it really flexible in usage. It can also provide different means on classification of data through the use of kernels. A very noticeable weakness I noticed was the amount of time the algorithm took to run. As the cost increased above 10, a single run could take hours in a dataset of 8k values. Another weakness I encountered was it having a lower accuracy in predicting data, but it was not a wide difference.

1. Random forest is part of the ensemble techniques. This division of algorithms typically combine weak learners, either sequentially or in parallel. The sequential approach involved running the algorithm several times and each time it is more heavily weighted. The parallel one aggregates the results of many weak learners that are run at the same time.

Random forest is an approach where the algorithm trains multiple trees on subsets of the data. It makes each tree independent. For regression, results are averaged, and a majority vote is used for classification. This algorithm uses a different data sample of each independent tree and changes variations in it, significantly reducing variance.

The other two algorithms I used were boosting, from the adabag package, and boosting using the XGBoost package.

The adabag package had an accuracy of 0.704. The XGBoost package had an accuracy of 0.6785. The simple decision tree had an accuracy of 0.70425. Overall, the algorithms performed similarly, although the simple decision tree ended up having a slightly higher accuracy.

The ensemble techniques had several advantages I noticed compared to other algorithms. Coming from the SVM algorithms, the ensemble techniques were noticeably faster, this is due to the algorithm running on all the cores of the computer, at least for XGBoost. XGBoost also prunes the tree automatically and provides flexibility on being able to perform classification or regression. The overall ensemble techniques do well on reducing variance and bias. Some of its weaknesses come from having slightly lower accuracy than other algorithms specific to their function.