SVM Kernels and Ensemble Methods

Subash Chandra and Derrick Martin

Support Vector Machine is a supervised learning algorithm that is mainly used for Classification, but can also be used for regression. The key of SVM is that it tries to separate data by finding the Optimal hyperplane that maximizes the separation between the classes. It does this by finding the similarities between the data points, and then construct the hyperplane as to maximize the confidence of the data to be on either side of it. As the data becomes less separable in the current dimension, another dimension can be added, and the vectors can be recalculated and added to the current model.

SVM Kernels allow regular transformation of the dataset to find patterns that would not normally be able to be found in data. There are many possible kernels that will each do varying levels of transformation like Polynomial Kernels, or Gaussian Kernels on one hand, or Radial Basis Kernels which is much simpler. Kernels have to be chosen carefully. The wrong kernel can lead to pretty unusable models. Some kernels may have complexity that grows with the size of the dataset, so using these would only be good in small datasets.

SVM Algorithms also have “Regularization” Parameters which allows the user to specify how much mis-classification is allowed, with a higher parameter meaning that there will be more accurate classification. This parameter essentially allows the optimizing hyperplane to be more complex, but can lead to overfitting under certain circumstances.

Random Forest – This is an Ensemble Technique that uses several decision trees that vote on the most likely choice, and the class with the most votes is chosen as the prediction. Because it uses a large number of relatively un-correlated models, it is usually able to perform better than any single model alone, and relative weaknesses of certain models that it uses will be outweighed by the wisdom of the crowd that is voting alongside it.

XGBoost – is another Ensemble technique that is popular because it is as good as Random Forest but much faster. It achieves this by running a model, and then sequentially combining future models by taking the predecessor models errors into account. It is also able to selectively weight the predictors and models that perform well, and weed out those that don’t. XGBoost is highly popular because of its ease of use, and accuracy, as well as speed.

Adaboost – Is another ensemble Boosting technique which re-assigns weights based on incorrectly classified instances. By targeting these outliers, it is able to accurately find the optimizing hyperplane that separates the data.

XGBoost was by far the best, being even faster than the base case of the Decision Tree, while maintaining accuracy and error rate.