Meinhard Capucao, Khang Thai

The SVM was invented by Vladimir Vapnik in the 1960s. SVM performs well in a variety of scenarios, but is used for two-group classification scenarios.. SVM creates a hyperplane which separates data into classes. For multiple predictors, the hyperplane is in multidimensional space, which takes into account having multiple predictors. SVM then takes points of data and categorizes them into either side. This is achieved by finding an optimal hyperplane to categorize points of data. The line is a decision boundary, and data points will be classified based on what side of the decision boundary it will fall on.

SVM Kernels are a set of mathematical functions that take in data as an input and transform it into the required form. Kernels are shortcuts that help avoid complicated situations, and help form the hyperplane in higher dimensions. Kernels help to solve nonlinear problems since they have different forms, such as linear, polynomial, radial, and sigmoid. They help with manipulating the shape of the hyperplane and the overall decision boundary. The type of kernel determines the overall decision boundary, for example, if the kernel type is linear then the decision boundary. Since the kernel does the hard work, then all the user has to determine is what type of kernel is optimal based on the decision boundary.

The first strength of SVM is its proficiency when the classes are clearly separable, since the decision boundary will be able to tell. SVM is also able to handle non-linear data, since the kernel trick can be employed to change the shape of the hyperplane and find the best decision boundary optimally. SVM is also able to solve both regression and classification. Some disadvantages of SVM is the liong training time it requires since it needs a lot of calculations, and even more so for large datasets. It is also difficult to choose the correct kernel at times unless one has complete understanding of the data. Lastly, one needs to scale the variables for SVM.

Random Forest is an ensemble of decision trees that are merged together to get an accurate and stable prediction. Random Forest is generally used for multiclass problems. It is different from SVM because SVM is for two-class problems while the Random Forest makes a bunch of binary trees and then combines them all together to predict a certain outcome. Random Forest is better used for bigger dataset but SVM is better if used for fewer data and more sparse data.

The other algorithm we used was XGBoost. XGBoost uses simple decision trees to be able to predict the correlation. XGBoost trees are made in parallel with each other which increases how well it can perform. It scans all the data and uses them to evaluate the quality of the split at every possible split in the dataset. XGBoost is something that is still being developed by many data scientists and is currently one of the leading algorithms that is used to solve all types of challenges with regression or classification. The downside is that because it is still being updated, there may be some datasets that it can not work with or be used to predict correctly.

The third algorithm we used was superlearner. Superlearner is an algorithm that uses cross-validation to find the performance of multiple machine learning models. In our datasets, Superlearner was not very beneficial or even useful. We didn’t post the results in the pdf because it wouldn’t run correctly and didn’t want to have misinformation. Superlearner is not as good as the other examples because it is a loss based learning method and is an asymptotically optimal system.