Joshua Durana

RCD180001

Kernel and Ensemble Methods

SVM uses a hyperplane that separates data into 2 regions, then there are 2 margins that contain instances within them called support vectors. Classification is trying to find the best size of the margin that clearly separates each class. For regression, we're trying to find the hyperplane to predict values for our target. Each kernel changes the shape of the hyperplane. The linear kernel makes the hyperplane a straight line. The polynomial kernel makes the hyperplane polynomial line. Finally the radial kernel makes the hyperplane a circle.

Classification in SVM works well when the data is well separated and is very versatile working with multiclass classification. But, you need to figure out the cost value each time you make a SVM model. This is slightly solved by using the tune function from a library, but that can take time to compute if you have a big dataset. It's also hard to interpret it, due to the hyperplane being multidimensional and being pretty complex.

Random forest works by making multiple decision trees with the subset of the training data, thus making each tree to be independent from each other. Then, when it's time to predict each tree will vote on the most likely outcome. Adaboosting is similar to random forest, but uses the whole dataset and each tree is built sequentially. XGBoost is similar to Adaboost, in that it uses boosting. But, it's extremely fast. Random forest works well if each factor level is distributed well. Adaboost works well with factor levels that aren't really distributed, while being slower than XGBoost. XGBoost is much faster but extremely hard to work with to make it actually work.