Leu, Rebecca

Support Vector Machine Classifier

In R

**What is SVM?**

SVM is a classifier machine learning algorithm. It is one of the options for modeling data to make predictions. SVM builds hyperplanes that separate data into different classes. Although there are different ways that the SVM can build these planes, the main idea is to maximize the distance from one datapoint to the next between classes.

**Using SVM to predict heart disease**

For this project we are using a heart disease data set which has 14 variables and 200 observations. The 14th variable is a 0 or 1 indicating whether the person has heart disease or not. The other 13 observations are the predictors of the disease. The dataset does not have column names for the data, so we aren’t sure what these qualifiers are but they are all numeric values.

**Preparing the data and getting started**

For our model we are going to use the carat package in R which is knows as the classification and regression training function. After importing our data we initially look at the data using the str function and the head function. With this we can confirm all of our data is numeric.

Before we can do anything to prepare the data, we need to separate the data into our testing and training dataset. We do this using the createDataPartition function and specify which column of variables we are targeting and what percentage we are going to split the data by. After splitting the data we convert both of the sets into data frames again.

After we split the data we have to correct the data to prepare it for the model. This is done by checking whether we have missing values in the data and standardizing the data. We do need to convert some of the data into categorical variables.

Once we have fixed any errors in the data, and prepared the dataset for our machine learning algorithm, we can begin to use the SVM model options built into the carat package. Here we use the train command and specify which column we are wanting to predict as well as what data we are working with and what method we want to use.

Once we complete this command, we have a model. Now it is time to use it to predict our values in our test dataset. Using the predict command we get our prediction values and then see how our results match up to the model for accuracy. Using the confusion matrix we get a model accuracy of 86.67%. Next we can do some customizations to see if we can get a better accuracy. After selection our C values in Linear classifier of 0.05 we get an accuracy of 87.78, so we didn’t make much of a difference.

We can also try another method of SVM in the carat dataset to see if we get better results. This time we are using the RBF kernel which is a non-linear kernel. Using our train method of svmRadial and selection a C and sigma parameter we can test out model’s accuracy. This accuracy is giving us an 87.78%. when we tune our classifier using different values of C and sigma using a grid search, we get an accuracy of 86.67% which shows us that we could be overfitting our model.