Edward Eastwood

COMP 379

SVM vs. KNN

For homework three, I chose to use the Scikit Learn Support Vector Machine. The results from this classifier were compared with my own implementation of a K-Nearest Neighbors classifier. Both of these classifiers were trained and tested on the same data. The data set used was the wine data set provided by the UCI machine learning repository. The wine data set has three different classes based on the grapes used in making the wine. Each wine is very easily classifiable by a good classification algorithm.

The SVM provided by Scikit learn was extremely successful. Using a C value of 1 and a decision function of “one-vs-rest” for multi-class classification, the SVM is able to achieve 100% accuracy on the test data. When the C value is reduced to a number below 1, the success of the SVM also begins to fall. The SVM is able to obtain these results so easily because of the separability of the wine data.

My K-Nearest Neighbors implementation did not nearly have the success that the SVM did. At best, I could only get a 34% success rate on the test data. The algorithm proceeds as follows. For each testing sample, iterate through each row in the “training” or “comparison” data and find the k nearest samples. Closeness was determined by Euclidean distance (implementation provided by scikit learn). Then, find the mode of the nearest neighbor’s classes and save it to the output. This is a naive implementation of KNN but I did it without using any resources for help or reference. It did better than I had predicted. Originally, instead of Euclidean distance, I determined closeness by summing the difference between each feature of the test and the comparison sample. This provided for extremely poor results (< 20% correct). I

then decided a more standard distance metric would be appropriate. Tweaking the k value from the default value of 5 does not improve the accuracy.

Each algorithm was able to provide some level of success on the wine data set. Each of them can also handle multi-class classification fairly easily, which is very useful for the wine data set which has three different classes. Overall, the SVM with a C value of 1 and a one-vs-rest approach for the decision function performed the best.