Cody Falzone
SID: 860929046
CS170
Project 2 Report

Feature Selection with Nearest Neighbor & Leave-one-out

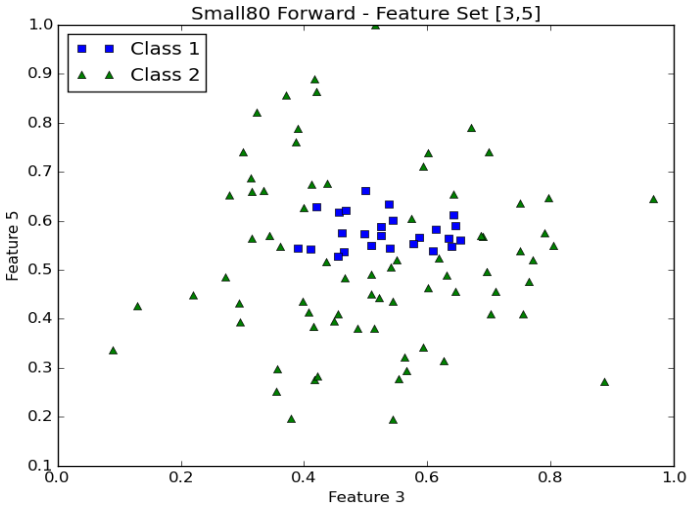
This report considers an implementation of a feature selection search. Feature selection search is used to determine the best set of features given a data set for training. This feature set is then used to classify unknown instances we may see in the future. So it is important that we pick the best feature set that allows us to classify instances with a good accuracy.

In order to implement this feature search selection, multiple components need to work together. Namely, a classier, validator, and selection algorithm. For the classifier, a nearest neighbor classifier is used. A leave-one-out validator is used. For the selection algorithms, two greedy search algorithms are used , forward selection and backward elimination.

The nearest neighbor algorithm classifies an unknown instance by checking the closest known instance based on distance. The nearest neighbor classifier is sensitive to noise. A solution to this is expanding the nearest neighbor algorithm to check multiple closest neighbors known as the KNN (k nearest neighbors) where k is some odd number of neighbors used to classify the unknown instance.

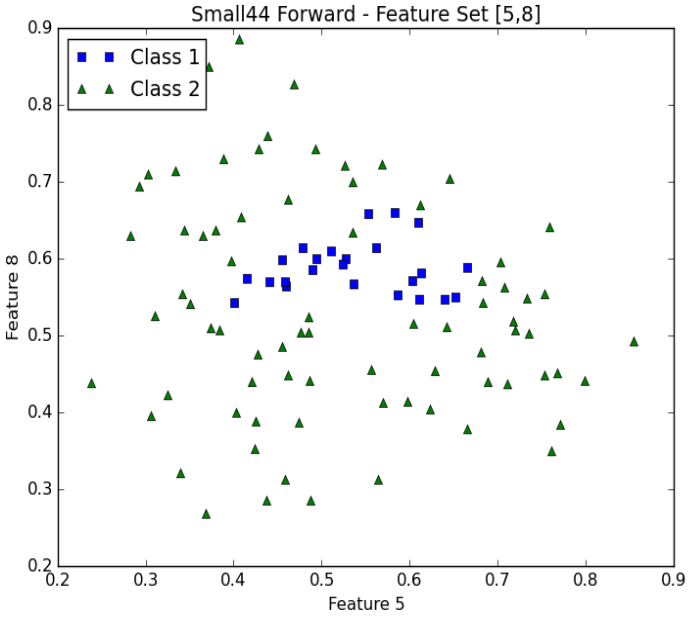
The leave-one-out validator is used to check the accuracy of the classifier. It takes in the feature set we are testing alone with the training data of instances. We then take one of the instances out, hence the name leave-one-out, and then keep the rest of the instances as known instances. Then we pass in the one instance we left out and tell our classifier to classify it. We then check to see if the classification was correct since we already know its classification. We repeat this process for all the instances in the training data set, leaving all the instances out one by one. After this process is complete, we take the number of instances the classifier got correct and divide it by the total number of instances to get an accuracy percentage for the corresponding feature set. This leave-one-out can be expanded by leaving sets of instances out at a time instead of just one at a time. This is known as k-fold cross validation. This tests the algorithm k times, leaving one section out of the training data each time.

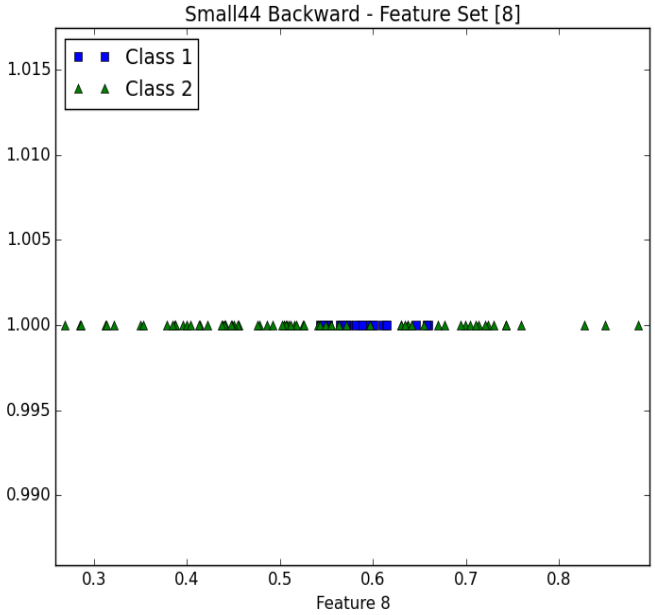
Two feature selection searches are used, namely forward selection search and backward elimination search. These searches use the classifier and validator to select the feature set with the best accuracy. The forward selection search starts with an empty set of features. Then, it adds features sets that all contain a single feature for all the features. For all these feature sets, it tests the accuracy using the validator. It then keeps the set with the best accuracy, and creates new features sets. Each of the new feature sets are created by adding a new single feature. This process is repeated until we test the set with all the features. The highest accuracy feature set of all the sets tested is kept. On the contrary, backward elimination starts with a set of all features. It then tests the accuracy of this set, and moves on to create a new sets by removing just a single feature. All these sets are then tested for accuracy and once again, this process is repeated until we end up with an empty set. The highest accuracy feature set is kept.

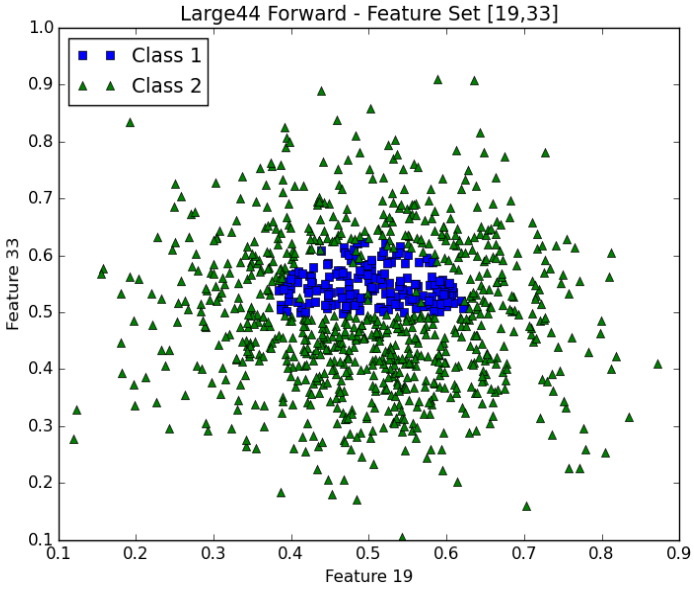


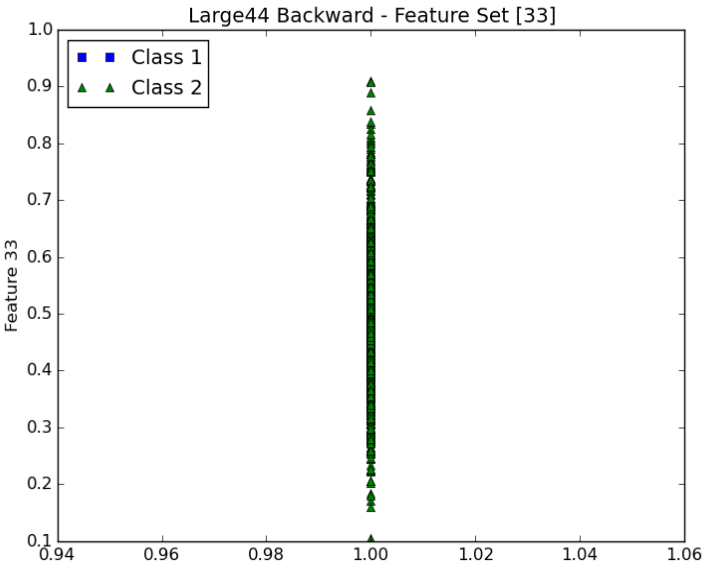
Accuracy Table:

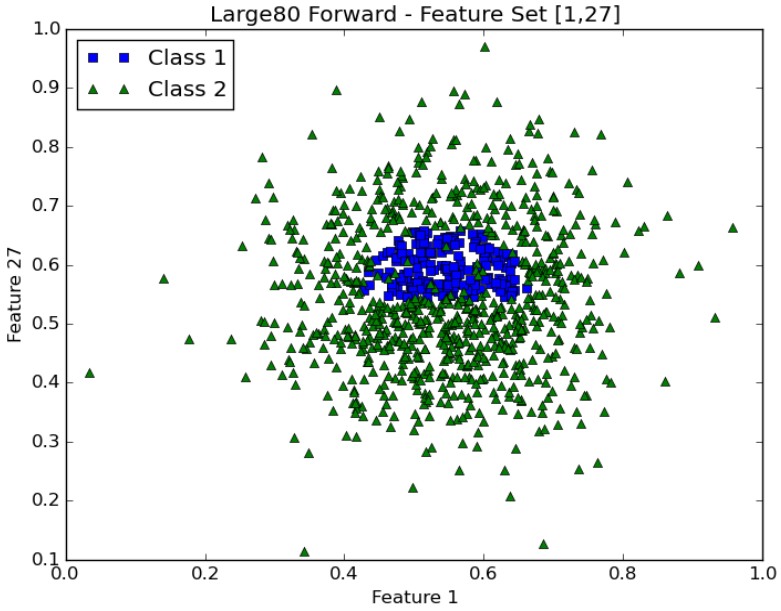
Small80 Forward – 92.0%
Small80 Backward – 83.0%
Large80 Forward – 95.5%
Large80 Backward – 75.4%
Small44 Forward – 90.0%
Small44 Backward – 86.0%
Large44 Forward – 96.2%
Large44 Backward – 84.0%











Plots of the feature sets found by the search algorithm are shown above along with their accuracies. Small80 backward plot is not shown, it finds feature set {2,4,5,7,10} with an accuracy of 83.0%. Large80 backward is also not shown, it finds feature set {2,3,9,12,14,17,18,19,25,28,30,35,40} with an accuracy of 75.4%. The forward and backward algorithms are greedy, meaning they do not do an exhaustive search. This means they may find different results. For these training data sets, the forward search finds better feature sets with higher accuracies. This means that for the training data, there is one or two features that are relevant and give us the best accuracies and the others are irrelevant or do not improve our accuracy. The backward elimination search eliminates or best features early in its search so when it tests the single features, the best one or two features have been eliminated and therefore are not tested.

Comparing the plots of the data, the forward search plots are easier to see the clusters of data. The two classes are easier to separate with class 1 cluster being towards the center and class 2 forming around the center circle of class 1. On the contrary, the backward search plots shows a less accurate feature set and are harder to visually separate the classes. The two classes are overlapping more in the lower accuracy backward search plots.