HW4 Write-up

Elias Crum

Data Loading and Preprocessing:
I chose to use a wine dataset for HW4 analysis. Loading the wine dataset was done with sci-kit learn using the `datasets.load_wine()` command. The data values were standardized using the sklearn `StandardScalar()` method.

1. The data was split into 80% training and 20% testing data using the `train_test_split()` sklearn method.

2. I trained an SVM classifier using an n-fold cross validation training method made from scratch. Before splitting the dataset into n-folds, the value n was chosen as a random integer between 1 and the length of the dataset/2 (so that every testing set would contain at least two instances). Within `cross_fold_svm()` method itself, there are three methods, the first that splits the input dataset into folds `split()`, a second that trains the default sklearn SVM on input training data `train()`, and a third that that tests an inputted model against an inputted development dataset `test()`. These three methods are all called within a loop over the number of folds, which is determined randomly as stated earlier. Thus, for each iteration of the loop, one fold is taken out of the training data and used as the development data. The results are printed when the method is run, but for a number of folds = 5, the average SVM accuracy for all folds was 95%.

3. I implemented a `grid_search()` method that optimized the accuracy of an SVM model by altering the hyperparameter values of C and gamma. I chose C and gamma because these parameters are the most influential to SVM model tuning. My grid search method utilized 9 values of C and 9 values of gamma that would be searched through (more or other values could be used by I used .00001 to 10000 for simplicity). The method is quite simple and just consists of a nested for loop ensuring that each C value will be used with each gamma value as the hyperparameters for a sklearn `SVC()` model. Each of these models will be trained on training data from the wine dataset. The accuracy of each model is determined by testing on a development data set. The highest accuracy value is recorded along with the highest C and gamma values. When run, the highest accuracy obtained was 0.977 with a C value of 0.01 and a gamma of 0.0001.

4. I tested the testing data on the SVM model with the best hyperparameters determined in the previous grid search optimization step. The optimized SVM performed with 100% accuracy on the testing set. I used the accuracy performance metric because the wine data set has 178 total instances, of these n[0] = 59, n[1] = 71, and n[3] = 48. Since all classifications are fairly well represented within the dataset (i.e. there is not one label with only 5 or 10 instances), the data is fairly balanced and accuracy should be adequate for model evaluation. When tested on the

training data, the accuracy was lower at 98.6%. The training results were lower which was odd. This is likely just because the training data has many more instances and thus presents a larger sample size compared to the test dataset. The accuracy was still very high but there is little that can be determined from the results of testing on training data because there are large amounts of bias involved.