Programming Assignment 3 Summary Report

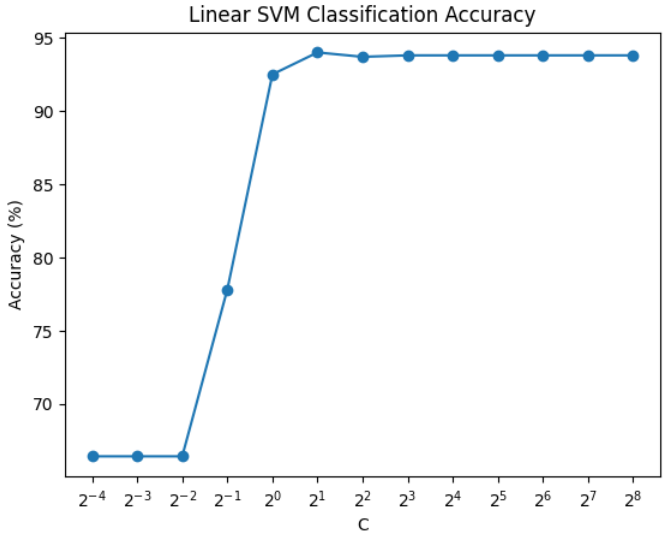
For this assignment, two SVM models were trained to determine if a genomic sequence is ncRNA. The source program can be tested by running “python main.py”, and all charts are attached along with this submission in the “plots” folder.

Training Data

Data was provided with the assignment, and is loaded using the svm_read_problem() function of LIBSVM.

Linear SVM

The first model was trained using linear SVMs. After trying the 13 different C values, the classification accuracy for each C are plotted below. I found that the best accuracy occurred when C = 2:



RBF Kernel SVM

To train the SVM with RBF kernels, 5-fold cross validation was used to find the best C and α value. First, the training dataset was randomly halved. Then, one half was split into 5 subsets and a model was trained for each validation subset, averaging the results to determine the performance of the hyperparameters. With a seed of 0 for randomization, the resulting matrix is printed to the console and shown below. The best model resulted from C = 64 and $\alpha = 2^{-4}$:

```
[[68.3 68.3 68.3 68.3 68.3 68.3 68.3 68.3 68.3 68.3 68.3 68.3 68.3]
 [68.3 68.3 68.3 68.3 68.3 68.3 68.3 68.3 69.  70.4 68.9 68.3 68.3]
 [68.3 68.3 68.3 68.3 68.3 68.3 68.6 74.8 77.  76.1 74.2 70.5 68.3]
 [68.3 68.3 68.3 68.3 68.3 73.3 80.7 83.3 83.  81.2 78.3 74.9 69.9]
 [68.3 68.3 68.3 68.4 81.4 90.4 91.3 89.7 88.2 84.7 81.3 78.5 75. ]
 [68.3 68.3 69.5 88.8 93.7 93.7 92.9 92.1 89.3 86.9 82.8 78.4 76. ]
 [68.3 70.2 90.8 94.5 94.6 94.2 93.5 92.7 90.1 88.  82.3 78.7 75.9]
 [70.7 92.1 94.8 95.  95.4 94.3 94.2 92.8 90.3 87.6 81.7 78.5 76. ]
 [92.5 94.7 95.2 95.  95.3 94.9 94.6 92.4 89.8 87.  81.4 78.5 76. ]
 [95.1 95.5 95.1 95.2 95.2 94.5 94.8 92.  89.8 87.2 81.4 78.5 76. ]
 [95.6 95.  95.3 95.1 94.9 94.  93.  91.7 90.4 86.6 81.4 78.5 76. ]
 [95.1 94.9 95.1 94.9 94.4 93.7 93.6 91.2 89.2 86.4 81.4 78.5 76. ]
 [95.  95.2 95.  94.9 93.7 92.8 92.5 91.3 88.  86.4 81.5 78.5 75.9]]
Best Model:     c=64    a=0.0625
```

Finally, the whole training set was used to train the best C and α parameters found from cross validation. The best model from cross validation had an accuracy of 95.6%, and using the full training set resulted in an overall accuracy of 94.0%. This decrease in accuracy may be due to slight overfitting to the training and validation sets.

```
Validation Dataset Accuracy: 95.6
Full Dataset Accuracy: 94.00599400599401
```