Spam Filter

The higher dimensional dataset of the spam filter was difficult to visualize in a way that would make it easy to determine whether or not the data was linearly separable or not, which affected our decision of whether to kernelize the data. To test whether the data was linearly separable, we ran the two class linear SVM without a kernel to see its accuracy. The data was taking a very long time to run on the linear SVM without scaling, thus we scaled X utilizing the standard scaling in sklearn. We separated the training data 80/20 into a training and validation data set to help with our hyperparameter search. To find the best hyperparameters including C, number of iterations, and learning rate, we searched the entire space starting with C values. We varied 9 C values to choose the best validation accuracy with a small number of iterations and step size of 1e-4 to make sure the values converged. The results are shown in Table 1 below.

| C value | Validation Accuracy |
|---|---|
| 0.01 | 0.94625 |
| 0.03 | 0.94625 |
| 0.1 | 0.94625 |
| 0.3 | 0.95875 |
| 1 | 0.95875 |
| 3 | 0.9625 |
| 10 | 0.97375 |
| 30 | 0.98 |
| 100 | 0.98625 |

Table 1: Hyperparameter C Search

Next, after choosing the best C (C=100), we wanted to optimize the number of iterations and the learning rate. We tested different values of these two parameters and found the following validation accuracies.

| Learning Rate | Number of Iterations | Validation Accuracy |
|---|---|---|
| 1e-2 | 1000 | 0.99 |
| 1e-2 | 5000 | 0.99 |
| 1e-2 | 10000 | 0.99 |
| 1e-2 | 20000 | 0.99 |
| 1e-2 | 50000 | 0.98875 |
| 1e-3 | 1000 | 0.98 |
| 1e-3 | 5000 | 0.98 |
| 1e-3 | 10000 | 0.98 |
| 1e-3 | 20000 | 0.98 |
| 1e-3 | 50000 | 0.98 |
| 1e-4 | 1000 | 0.98625 |
| 1e-4 | 5000 | 0.97875 |
| 1e-4 | 10000 | 0.97875 |
| 1e-4 | 20000 | 0.9775 |
| 1e-4 | 50000 | 0.9775 |
| 1e-5 | 1000 | 0.97535 |
| 1e-5 | 5000 | 0.985 |

| | | |
|---|---|---|
| 1e-5 | 10000 | 0.98625 |
| 1e-5 | 20000 | 0.98 |
| 1e-5 | 50000 | 0.97875 |

Table 2: Hyperparameter Learning Rate and Number of Iterations Search

To save time, we performed the C calculation separately from the learning rate and number of iterations but to do get a complete hyperparameter search, we would have varied C with every number of iterations and learning rate as well. The best learning rate was found to be 0.01 and the best number of iterations was found to be 1000.

After testing these hyperparameters on the testing set, we obtained a test set accuracy of 0.98 and training data accuracy of 98.425. We thought that this accuracy was very high for this dataset and that this meant the data was linearly separable. This made utilizing a kernel unnecessary as it would expend additional computational power for similar accuracies. We did make a simple Gaussian kernel with the best hyperparameters above and sigma = 1 that had a test accuracy of 0.822. This kernel performed less well than the linear SVM and therefore we though the linear SVM was the correct way to go.

We obtained a list of the top 15 spam and ham words and have listed them in the table below.

| Spam | Ham |
|---|---|
| clearly | instant |
| otherwiss | urgent |
| remot | datapow |
| believ | wrong |
| file | numberth |
| franc | issu |
| natur | that |
| dollarac | useless |
| water | predict |
| creativ | submit |
| off | august |
| young | these |
| gt | anyth |
| herba | url |
| reason | new |