Report Homework 4

1 Guide question 1

The training and testing accuracy for spam.cvs:
========== No kernel ==========
Train accuracy: 98.12 %
Test accuracy: 86.25 %
========== RBF kernel ==========
Train accuracy: 95.0 %
Test accuracy: 85.0 %
========== Polynomial kernel ==========
Train accuracy: 52.19 %
Test accuracy: 42.5 %

For the polynomial kernel hyper parameter, we found that when c=1 and d=2, it gives the best result on fakedata set. However, in the spamdataset, the polynomial kernel has poor performance. I tried to adjust the value of c among 0, 1, 2, 2.5 and the results are all not very good. This indicates that polynomial kernel may not be a good solution to separate the spam dataset.

Also, we noticed that the training accuracy is higher than testing accuracy, which indicates that there is overfitting problem with the model. For all three kernel, linear kernel generates the best performance which indicates the spam data is very linearly separable.