Grayson Bass
ID Number: 801074960
Homework 4
Github: https://github.com/gbass2/IntroToML

Problem 1:

For problem 1, the cancer dataset was imported from the sklearn library which can be found in section In[6]. Once the data was imported, a standard scaler was used. There were 3 kernels that were used for problem 1. The first was a linear kernel, the second was a polynomial kernel, and the third was Radial Basis Function. Section In[18] shows the code and plots for the linear SVM kernel. PCA was used with varying number of components ranging from 1 to 30. The accuracy, precision, and recall were plotted with each iteration being an increasing number of components for the PCA extraction. The optimal number of components for the linear kernel with PCA is 14. This is where the accuracy was at its highest and the recall and precision were near their highest values. The model with the polynomial kernel was created the using the same code as with the linear model except the kernel for the model was switched to a polynomial kernel. The code and plots are in section In[16]. The accuracy, precision, and recall were plotted the same way as with the linear kernel. The optimal number of components for the polynomial kernel with PCA are 1 and 2 components. This is where the accuracy, precision, and recall were at their highest. The accuracy and precision were at its highest with 10 to 14 components. The model with the RBF kernel was created the using the same code as with the linear and polynomial model except the kernel for the model was switched to a polynomial kernel. The code and plots are in section In[17]. The accuracy, precision, and recall were plotted the same way as with the linear and polynomial kernels. The optimal number of components for the RBF kernel with PCA is 8. This is where the accuracy was at its highest and the recall and precision were near their highest values. Comparing the results of the three kernels. It appears that the polynomial kernel showed the best results but only by a small fraction compared to the RBF kernel. The linear kernel took about an hour to run on my machine and the poly kernel and RBF took less than a minute. Because of the time consumption required to run the linear kernel, it performed worse than the other two. Comparing the results of the polynomial kernel to the results of homework 3 problem 2, the SVM model performed better than the logistic regression model. The results were more stable and did not fluctuate as much with different number of components.

Problem 2:

For problem 2, the housing dataset was used. The binary inputs were converted to a numerical value from the yes/no that was in the dataset. The data was then scaled using a standard scalar. The code for this is in section In[7]. A plot of the regression model can be found in section In[12]. PCA extraction was used to reduce the 30 features into 1 so that the model could be plotted. The plot shows the regression line with three kernels. The first was a linear kernel, the second was a polynomial kernel, and the third was Radial Basis Function. A scatter plot of the data was also shown.

A for loop was used to calculate the mean squared error for each number of components that was set for the PCA extraction. The mean squared error was then plotted against the iterations which was the number of components used at each iteration. The mean squared error was calculated as the accuracy, precision, and recall can only be calculated for a classification model. The code and plot for the linear model can be found in section In[8]. The optimal number of components for the linear model was 1. The mean squared error increased as more components were introduced to the PCA extraction. The code and plot for the polynomial model can be found in section In[9]. The optimal number of components for the polynomial model was 2. The mean squared error increased as more components were introduced to the PCA extraction. The code and plot for the RBF model can be found in section In[10]. The optimal number of components for the RBF model was 1. The mean squared error increased as more components were introduced to the PCA extraction. Looking at the plots of the mean squared errors, the linear kernel produced the lowest values for the mean squared error ranging from roughly 1.7 to 1.8 but the RBF produced the smoothest increasing plot of the mean squared error. Comparing the results to the linear regression model in homework 1, the SVR models produced better results.