Claude Shyaka
ID#: 801326243
ECGR-5105

Homework 4: SVM and SVR Models

GitHub Repo: https://github.com/claudeshyaka/ml

In this exercise, we used the cancer dataset to train a Support Vector Machin (SVM) model to classify the types of breast cancer (Malignant vs. Benign), then compared its performance against the Linear Logistic model developed in previous exercises. In addition, a feature extraction technique called Principal Component Analysis (PCA) was used to extract features that were most relevant to the model training. Furthermore, using the housing dataset we trained a Support Vector Regression model and compared its performance against the Linear Regression model investigated in previous experiments. The results of these experiments are presented below. Note for both the housing and cancer datasets the data was split into 80% for training and 20% for testing.

1. In this section, we used PCA feature extraction on the cancer dataset to derive the most relevant features of the dataset and then trained an SVM model.
    a. Using **linear** kernel and **K=2** value, we were able to obtain an **accuracy of 99.1%.** Figure 1 below shows accuracy results for different K values.

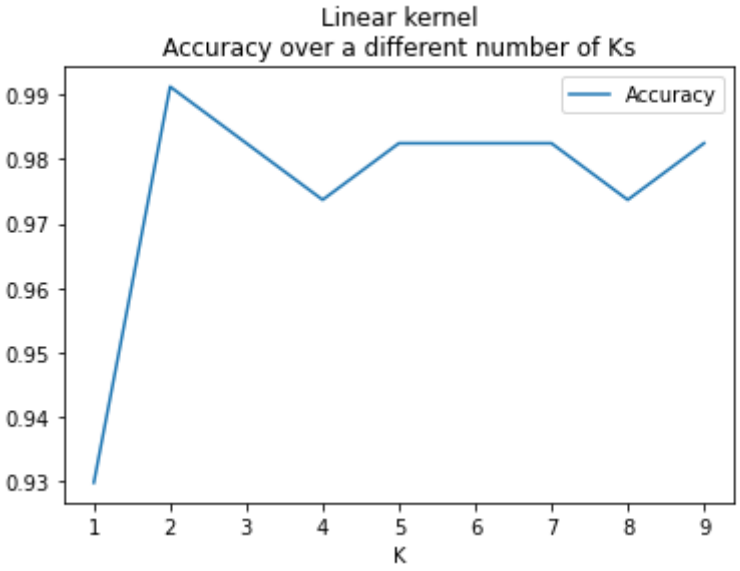

Figure 1: Accuracy results for different K values using a linear kernel

    b. Using the same kernel (linear), we computed the precision and recall results for different K values and obtained results shown in figure 2.

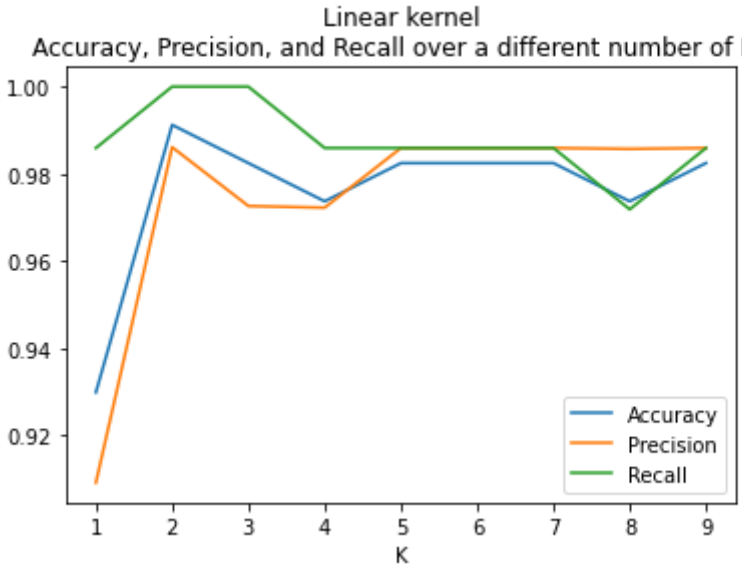

Figure 2 shows the accuracy, precision, and recall results for different K values

c. In addition to the linear kernel, we explored the RBF and sigmoid kernels. The accuracy results for different K values are shown in Figures 3 and 4.

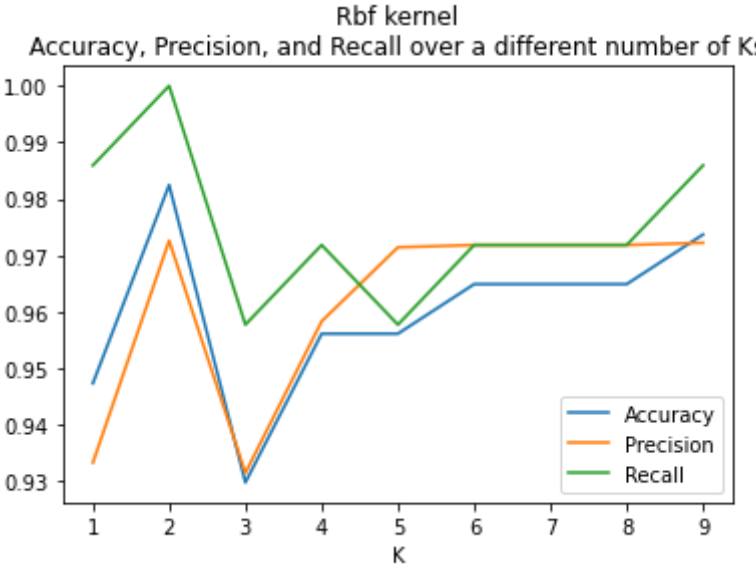

Figure 3 shows the accuracy, precision, and recall results for different K values

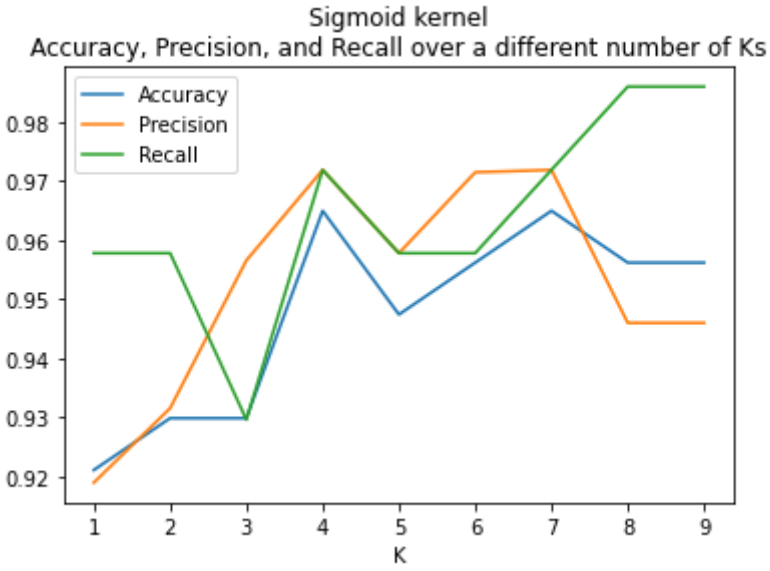

Figure 4 shows the accuracy, precision, and recall results for different K values

d. In homework 3, the logistic regression model trained produced comparable results to the SVM classifier trained shown above. Figure 5 shows the accuracy, precision and recall results for different K values for the Logistic regression classifier from homework 3.

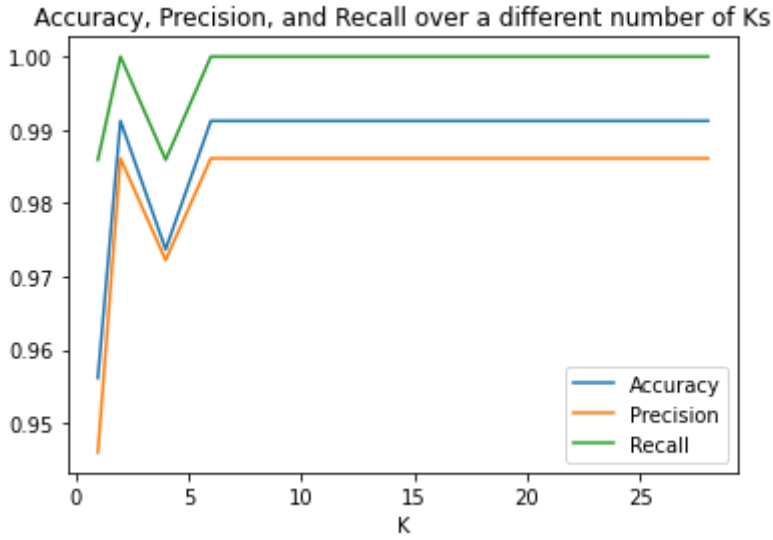

Figure 5

2. In this section, we trained an SVR model using the housing dataset and compared its results to a line regression model with regularization loss. Then, we used PCA feature extraction on the housing dataset to derive the most relevant features of the dataset and then trained an SVR model again.
    a. The SVR model trained using the housing dataset returned a mean squared error of 0.0156, compared to the linear regression model classifier which returned a mean squared error of 0.0171. Thus, the percent difference between the two models is

8.84%. In the end, the SVR model reported better results compared to the Linear regression model.

b. Using **RBF** kernel, we identified the optimal values of K as 10 in PCA, for the housing dataset. Figure 6 shows accuracy or more precisely mean squared error results for different K values.

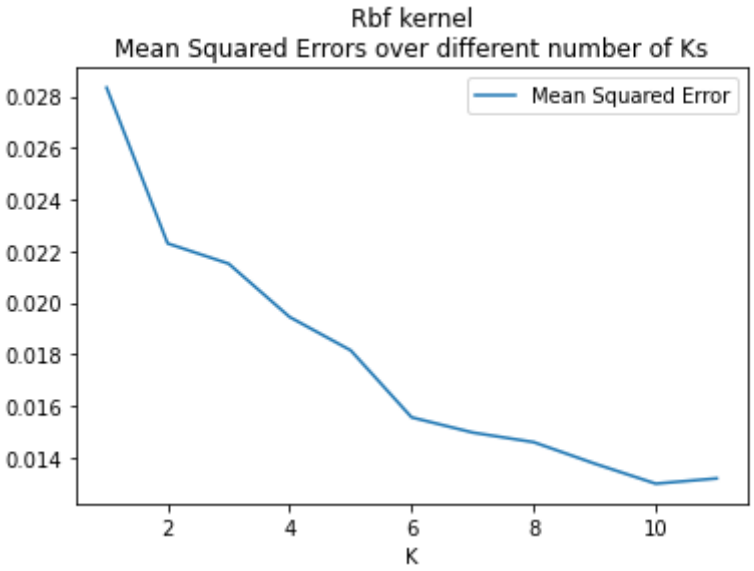

Figure 6 shows mean squared error results for different K values

c. In addition to the RBF kernel, we explored the poly and linear kernels. The mean squared error results for different K values are shown in Figures 7 and 8.

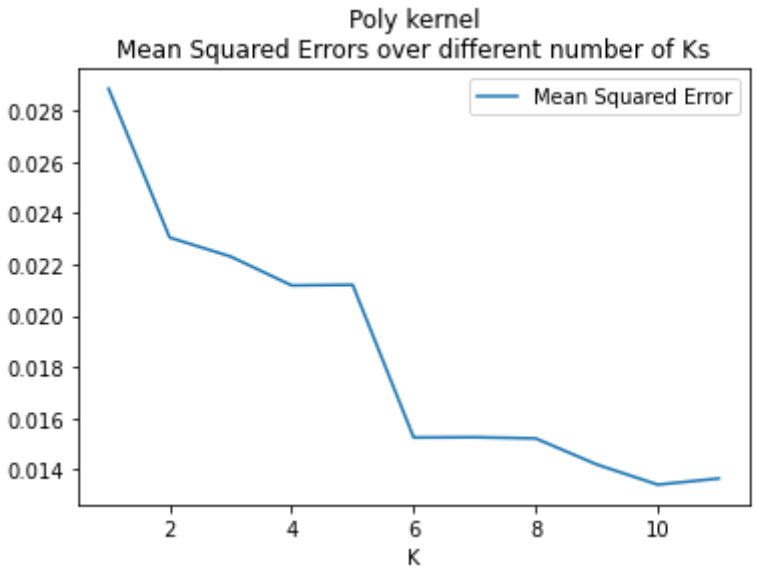

Figure 7 shows mean squared error results for different K values

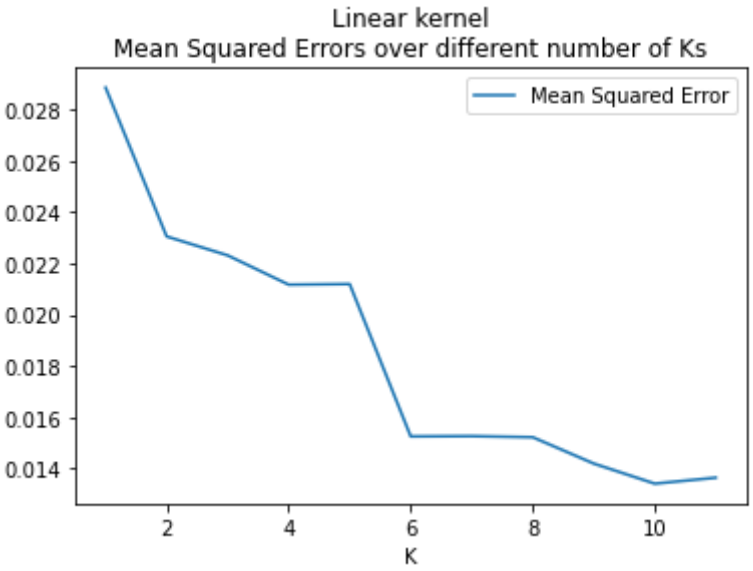

Figure 8 shows mean squared error results for different K values