Report

Statistical Data Mining I: Homework 2

Sai Lone (MPS)
Person Number: 50139692

---

1. Exercise 9 modified ISL

a) First split the results into a training set and testing set, then fit a linear model using least squares on applications.

```
> linear_mod.mse <- mean((test$Apps - linear_mod.pred)^2)
> linear_mod.mse
[1] 1135758
```

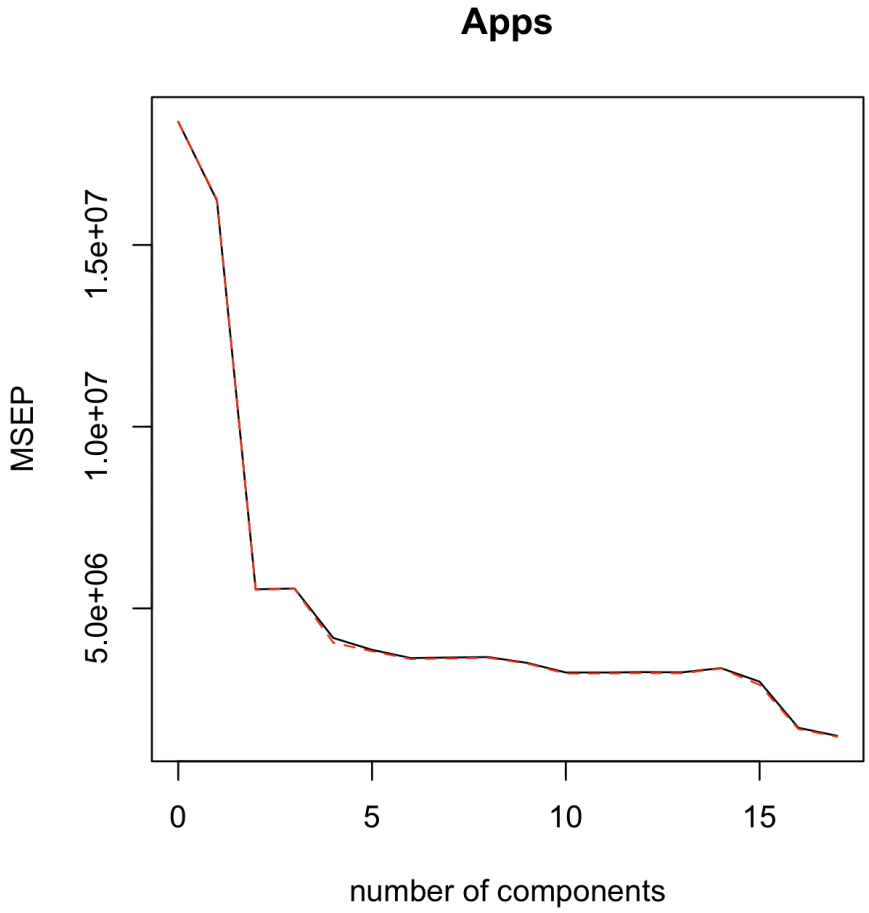
b) Use ridge regression model to obtain the best lambda value then predict the MSE value.

```
> ridge.fit <- cv.glmnet(train.matrix, train$Apps, alpha=0, lambda=grid, thresh=1e-12)
> lambda_best <- ridge.fit$lambda.min
> lambda_best
[1] 0.01
> ridge.pred <- predict(ridge.fit, s=lambda_best, newx=test.matrix)
> ridge.mse <- mean((test$Apps - ridge.pred)^2)
> ridge.mse
[1] 1135714
```

d) Use lasso model to and cross-validation to obtain the best lambda and predict the MSE value.

```
> lambda_best <- lasso.fit$lambda.min
> lambda_best # λ chosen by crossvalidation.
[1] 0.01
> lasso.pred <- predict(lasso.fit, s=lambda_best, newx=test.matrix)
> lasso.mse <- mean((test$Apps - lasso.pred)^2)
> lasso.mse
[1] 1135660
> lasso.pred_non_zero
19 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -7.900363e+02
(Intercept)  .
PrivateYes  -3.070103e+02
Accept       1.779328e+00
Enroll      -1.469508e+00
Top10perc    6.672214e+01
Top25perc   -2.230442e+01
F.Undergrad  9.258974e-02
P.Undergrad  9.408838e-03
Outstate    -1.083495e-01
Room.Board   2.115147e-01
Books        2.912105e-01
Personal     6.120406e-03
PhD         -1.547200e+01
Terminal     6.409503e+00
S.F.Ratio    2.282638e+01
perc.alumni  1.130498e+00
Expend       4.856697e-02
Grad.Rate    7.488081e+00
```

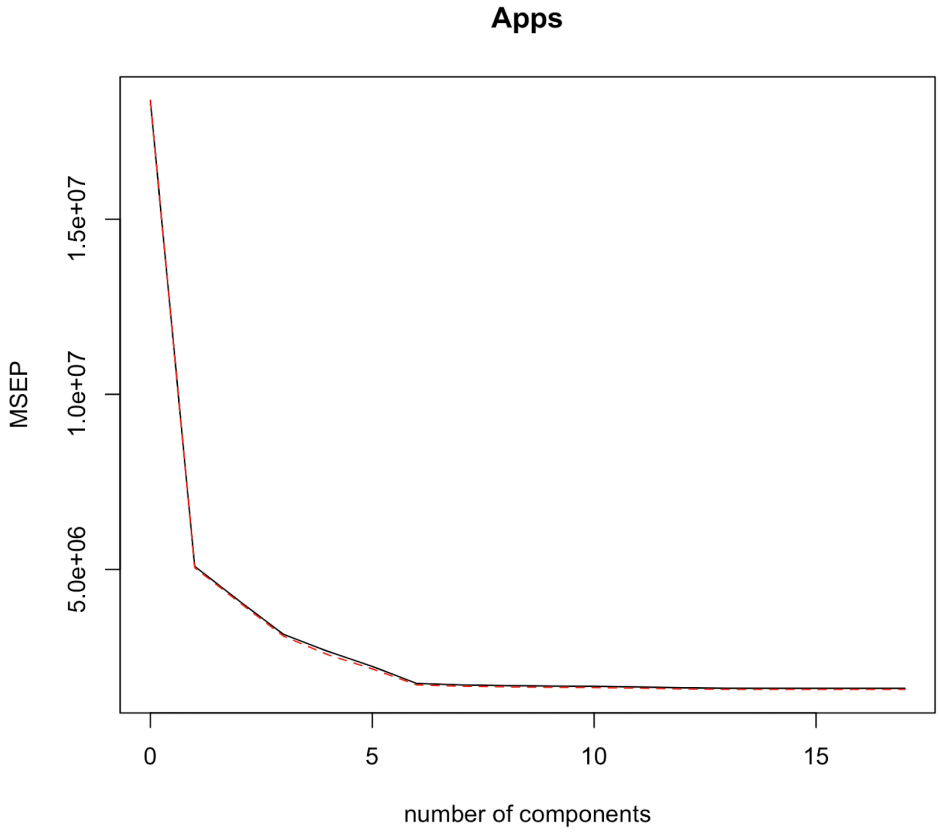
e) Fit a PCR model on the training set, with k chosen by cross-validation. The test error and the value of k selected by cross-validation is as follows:




```
> pcr.pred <- predict(pcr.fit,test, ncomp=10)
> pcr.mse <- mean((test$Apps - pcr.pred)^2)
> pcr.mse
[1] 1723100
```

f) Fit a PLS model on the training set, with k chosen by cross-validation. The test error and the value of k selected by cross-validation is as follows:

```
> psl.mse <- mean((test$Apps - psl.pred)^2)
> psl.mse
[1] 1131661
```



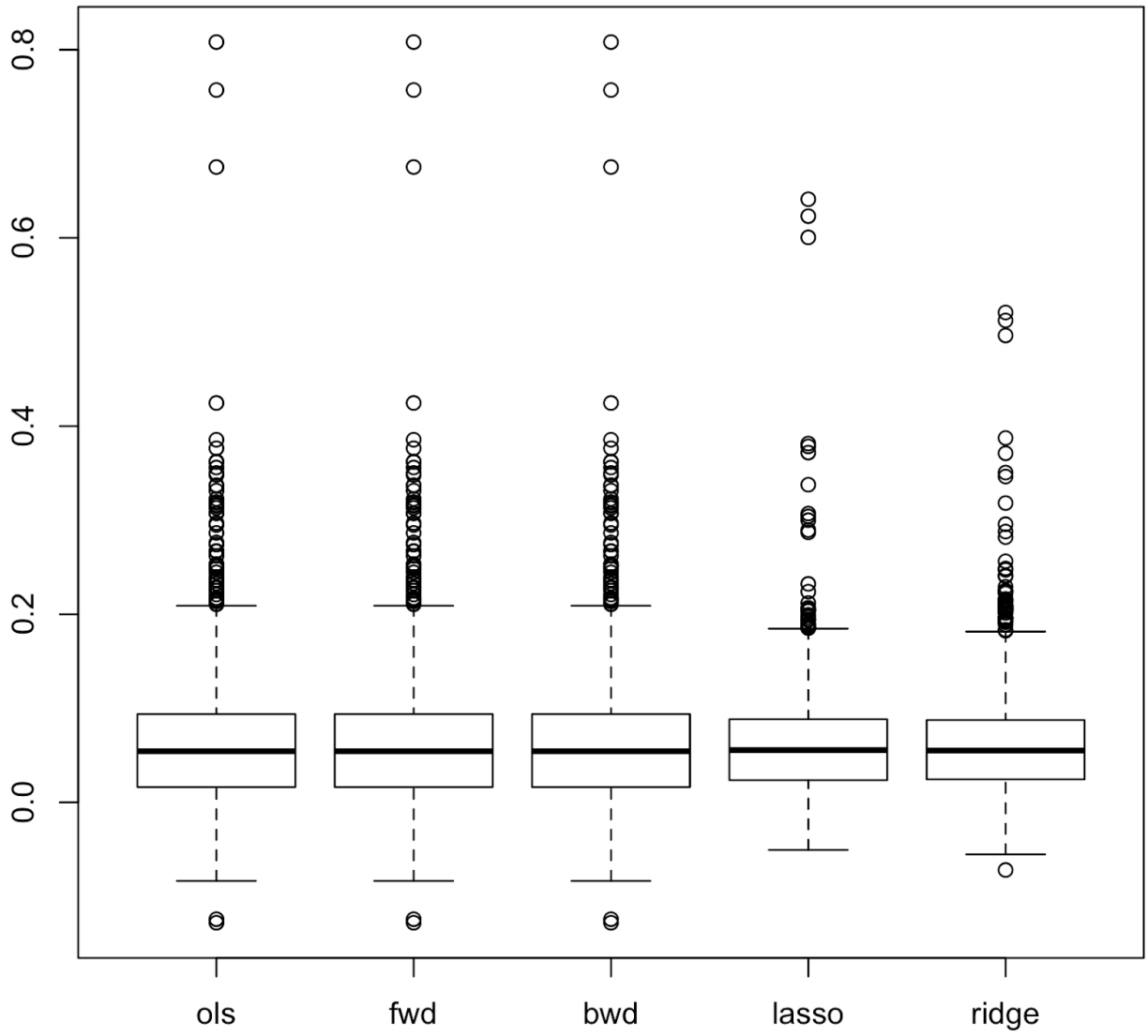
g) Based on the results obtained, all the models show little differences in the test error expect for the PCR model. Therefore we can say that all models excluding PCR, has a high accuracy for predicting college applications.

2. Caravan

To perform the different selection algorithms, the data must be split up into training and testing data sets for our model. Fortunately, the data was pre-split within the ticdata200.txt and ticeval2000.txt. From there we can choose the target data which in this case was the last column, V86 or “CARAVAN Number of mobile home policies 0 - 1” according to the data dictionary.

Using our derived data sets, we can then perform the variable selection algorithms for forwarding selection, backward selection, lasso regression, ridge regression, and OLS. Calculating the test error rate shows, the error rate was comparable among all the algorithms with OLS having only a slightly higher error rate. The estimates values are also predicted for

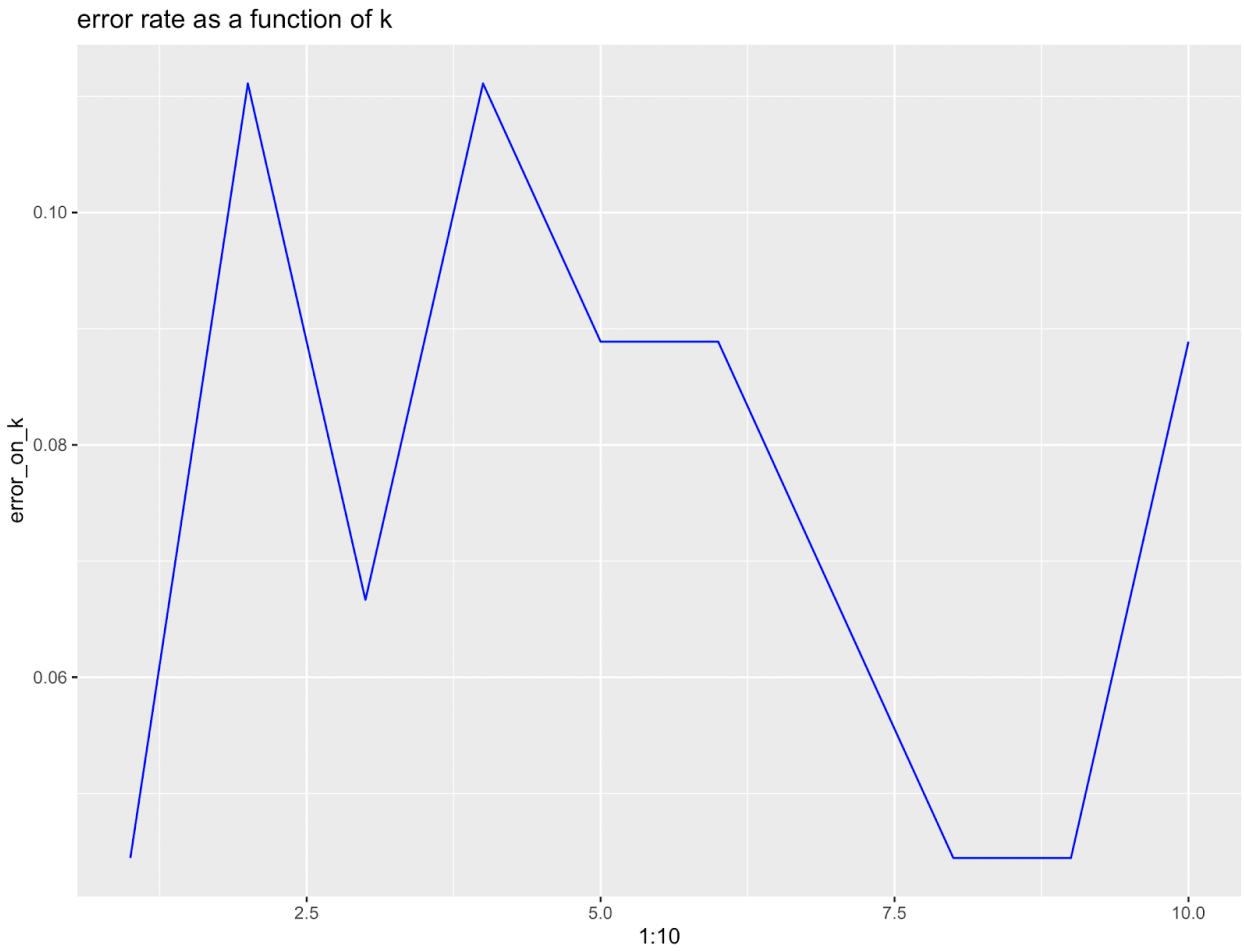
each of the 4000 data points in the test set. The box plot below shows the comparisons for the results of the 5 different algorithms.



3. K-nearest Neighbor on Iris Dataset

First, we load the data from the Iris dataset and analyze its dimensions and variables names. Then we split the data into training and testing with a 70% to 30% ratio respectively.

a) The plot of the error rate as a function of k where k ranged from 1 to 10, can be observed below.



The confusion matrix after performing kNN on the three species can be seen below:

```
> table(test_lable, knn_iris.pred)
             knn_iris.pred
test_lable    setosa versicolor virginica
  setosa          13          0         0
  versicolor       0         12         1
  virginica        0          1        18
```

As the results show our kNN model can discriminate agensit the various species. This is due to outliers that may be present in the iris data set. These outliers can make a drastic difference in the classification results.