Homework 01

Subhankar Ghosh

Question 1

(a) Perform a descriptive analysis on all variables. Comment on any potential issues and address them if needed.

The Columns in the Boston Housing dataset are as follows:

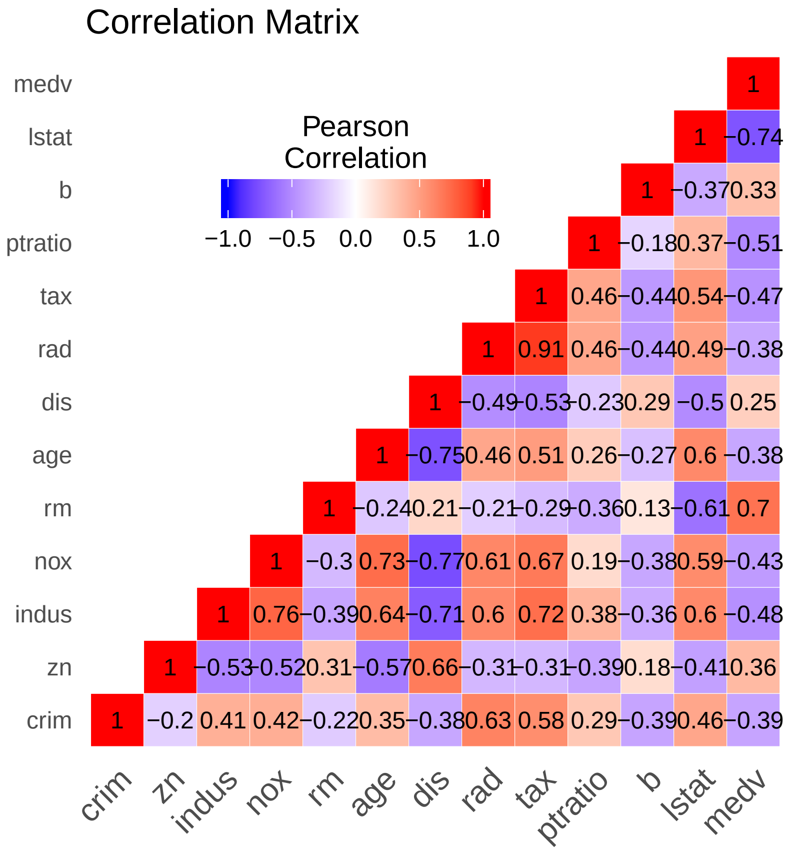
- crim : per capita crime rate by town
- zn : proportion of residential land zoned for lots over 25,000 sq.ft
- indus : proportion of non-retail business acres per town
- chas : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- nox : nitric oxides concentration (parts per 10 million)
- rm : average number of rooms per dwelling
- age : proportion of owner-occupied units built prior to 1940
- dis : weighted distances to five Boston employment centres
- rad : index of accessibility to radial highways
- tax : full-value property-tax rate per USD 10,000
- ptratio : pupil-teacher ratio by town
- b : 1000(B - 0.63)ˆ2 where B is the proportion of blacks by town
- lstat : percentage of lower status of the population
- medv : median value of owner-occupied homes in USD 1000's

Let us look at the Min, Mean, Median, Max values of the variables

Table 1: Summary of variables

crim	Min. : 0.00632	Median : 0.25651	Mean : 3.61352	Max. :88.97620
zn	Min. : 0.00	Median : 0.00	Mean : 11.36	Max. :100.00
indus	Min. : 0.46	Median : 9.69	Mean :11.14	Max. :27.74
chas	0:471	NA	NA	NA
nox	Min. :0.3850	Median :0.5380	Mean :0.5547	Max. :0.8710
rm	Min. :3.561	Median :6.208	Mean :6.285	Max. :8.780
age	Min. : 2.90	Median : 77.50	Mean : 68.57	Max. :100.00
dis	Min. : 1.130	Median : 3.207	Mean : 3.795	Max. :12.127
rad	Min. : 1.000	Median : 5.000	Mean : 9.549	Max. :24.000
tax	Min. :187.0	Median :330.0	Mean :408.2	Max. :711.0
ptratio	Min. :12.60	Median :19.05	Mean :18.46	Max. :22.00
b	Min. : 0.32	Median :391.44	Mean :356.67	Max. :396.90
lstat	Min. : 1.73	Median :11.36	Mean :12.65	Max. :37.97
medv	Min. : 5.00	Median :21.20	Mean :22.53	Max. :50.00

SOme of the variables like **crim** has a very high Maximum value compared to its mean and 3rd quantile so there might be some leverage points due to this.



- We can see a very **strong correlation of 0.91** between *rad* and *tax* which makes sense because we would expect full-value property-tax rate to go up as accessibility to radial highways increase.
- We can see a **high negative correlation of -0.77** between *nox* and *dis* so we can say that the concentration of nitrogen oxide increases near the employment centers.

(b) Perform the best subset selection using BIC criterion. Report the best model (the selected variables and their parameters).

Performing best subset selection using BIC we get the best model as:

```
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + b + lstat, data = BostonHousing)
```

With a BIC value of:

```
[1] 1636.479
```

Variables and parameters of best BIC model

Table 2: Variables and parameters of best BIC model

	best_bic.coefficients
(Intercept)	36.3411450
crim	-0.1084133
zn	0.0458449
chas1	2.7187163
nox	-17.3760234

	best_bic.coefficients
rm	3.8015788
dis	-1.4927115
rad	0.2996085
tax	-0.0117780
ptratio	-0.9465246
b	0.0092908
lstat	-0.5225535

(c) Perform i) forward stepwise selection using AIC criterion; and ii) backward stepwise selection using Marrow's C_p criterion. Compare these two models with the model in part b).

(i) Performing best subset selection using AIC we get the best model as:

```
lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age + 
    dis + rad + tax + ptratio + b + lstat, data = BostonHousing)
```

With an AIC value of

```
[1] 1589.643
```

Table 3: Variables and parameters of best AIC model

	best_aic.coefficients
(Intercept)	36.4594884
crim	-0.1080114
zn	0.0464205
indus	0.0205586
chas1	2.6867338
nox	-17.7666112
rm	3.8098652
age	0.0006922
dis	-1.4755668
rad	0.3060495
tax	-0.0123346
ptratio	-0.9527472
b	0.0093117
lstat	-0.5247584

(ii) Performing best subset selection using C_p we get the best model as:

lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + b + lstat, data = BostonHousing)

With a minimum C_p value of:

We observe that the best model for obtained by using C_p and BIC are the same with 11 predictor variables.

But the AIC model includes all 13 predictor variables.

(d) Comment on the advantages and disadvantages of the selection algorithms (best subset, forward, backward and stepwise). If you get different results using these three algorithms (assume that you use the same selection criterion), would you prefer some over others? Why?

Advantages

Best-Subset Selection: If we consider all combinations then it will consider 2^n combinations (where n is the number of independent variables) which is computationally infeasible if n is very large. But for small n it goes through all possible subsets of n in an incremental/decreasing fashion so it gets the best subset.

Forward Selection: Since it starts from intercept model, it works well for cases where number of predictor variables are more than the observations. Forward stepwise will have lower variance.

Backward Selection: It is computationally less expensive and tends to give better models.

Disdvantages

Best-Subset Selection: It can't work for large p(p>40) as it will become computationally infeasible since it goes through all 2^p subsets still it becomes computationally infeasible.

Forward Selection: As forward stepwise is a more constrained search, it will have more bias than models chosen by other selections.

Backward Selection: Backward selection can only be used when N > p(where N is the number of records and p is the number of variables). Since it starts from full model, it will require atleast n=p, i.e. a full rank matrix to give unique Beta values.

I would prefer Best-Subset Selection Algorithm over all others as it checks for all subsets of the predictor variables.

(e) Comment on the advantages and disadvantages of the three selection criteria (AIC, BIC, C_p). If you get different results using these three criteria (assuming the same algorithm), would you prefer some over others? Why?

AIC works better when the data is best modelled by a nonparametric model and BIC when the data is best modelled by a parametric model. When n is large, the costs incured by BIC is a lot more than AIC (or C_p). So AIC tends to pick a larger model than BIC. C_p works similar to AIC.

If the model is a parametric model then I would prefer AIC over others and in case of non-parametric model I would prefer BIC over AIC and C_p.

Question 2

(a) Provide a short summary of the dataset and the research goal.

Data Summary: Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255. The training and test data sets have 785 columns. The first column consists of the class labels (see above), and represents the article of clothing.

Research Goal: In this exercise the research goal is to do image classification by implementing a fast knn algorithm. This dataset is one of the most standard datasets to do image classification. KNN is a very well know approach to classify based on neighbourhood approach.

(b) Write your R code for kNN. This is a fairly large dataset, so you should consider writing an efficient algorithm. If it is very slow, you can consider doing part d) first. In addition, how do you deal with ties when k is even.

Function to get the misclassification rate

```
## Function to compute the Misclassification rate
## Input variables:
## pred: predicted values
## actual: actual values
get_misclassification <- function(pred, actual) {
    mean(pred != actual)
}
```

Implementation of knn model

```
## Function to implement knn
# k : choice of k
# x_test : Matrix of predictor variables of testset
# x_train : Matrix of predictor variables of trainset
# y_test : vector of class of testset
# y_train : vector of class of trainset
# return :: predictions for the x_test data
myknn <- function(k, x_test, x_train, y_test, y_train)
{
  y_pred = rep(0, nrow(x_test))
  for(i in 1:nrow(x_test))
  {
    ## Euclidean distance
    one_dist = sqrt(colSums((t(x_train) - x_test[i,])^2))
    one_df = as.data.frame(one_dist)
    a = y_train[head(sort(one_df$one_dist, index.return = TRUE, decreasing = FALSE)$ix, k)]
    y_pred[i] = a[max(table(a))]
  }
  y_pred
}
```

In case of ties we can select the best class randomly from the tied classes.

(c) Fit your kNN model to the training data, and predict the labels using the testing data. Tune the parameter k to obtain the best testing error and comment on the effect of k. Summarize the performance of the final model. What is the degrees of freedom?

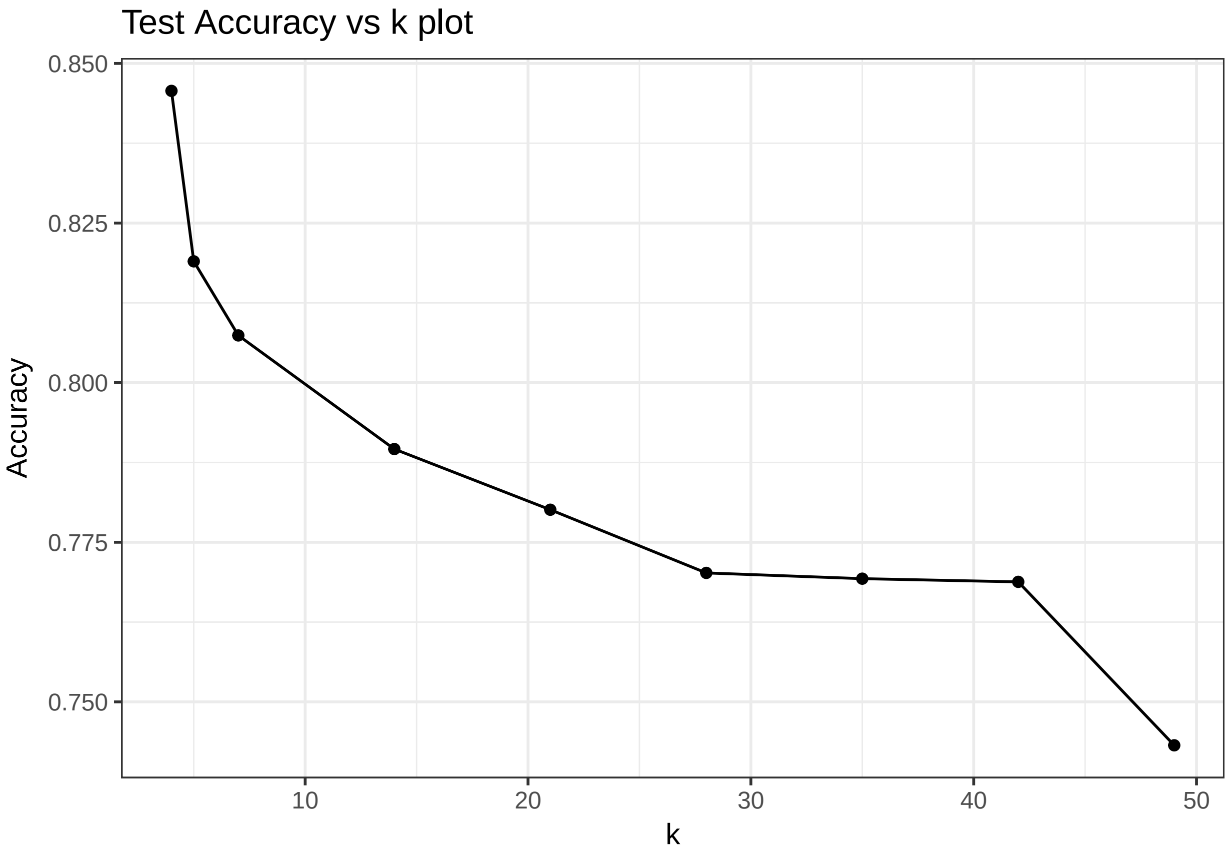


Table 4: Performance summary of knn as k varies

k	4	5	7	14	21	28	35	42	49
Accuracy	0.8457	0.8190	0.8074	0.7896	0.7801	0.7702	0.7693	0.7688	0.7432

From the plot of Test Accuracy vs k we can see that the highest accuracy of 0.8457 at k=4. Since the degree of freedom of knn is given by N/k the degree of freedom of the best model is 15000.

(d) Can you suggest some approaches that can speed up the computation (even at some minor cost of prediction accuracy)?

For speeding up the computation we can do the following:

- Instead of using a loop to iterate over the all the predictor variables to calculate the euclidean distance we can sweep through the predictor variables in R so that the computation for all the predictor variables can happen parallely.
- Since we have a large number of predictor variables we can apply PCA to reduce dimension of the dataset and speed up the computation.