Stat 432 HW 02

Name: Your Name, netID: yournetID

Summer 2024

Include the R code for this HW.

```
knitr::opts_chunk$set(echo = TRUE)
library(ISLR2)
library(GGally)
```

There are some useful R chunk options that you may use (for this entire semester):

- echo - Display code in output document (default = TRUE)
- include - Include chunk in document after running (default = TRUE)
- message - display code messages in document (default = TRUE)
- results (default = 'markup')
  - 'asis' - passthrough results
  - 'hide' - do not display results
  - 'hold' - put all results below all code
- error - Display error messages in doc (TRUE) or stop render when errors occur (FALSE) (default = FALSE)

See R markdown cheat sheet for more information.

Question 1 (Linear Regression)

We have N observations of $(X_1, X_2, \ldots, X_p, Y)$.

Let us use the following notations:

- $\mathbf{X}$ is a the $N \times (p+1)$ matrix with each row as an input vector (with a 1 in the first position),
- $\mathbf{y}$ be the N-vector of outputs and
- $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$.

Then we may write the multiple linear regression model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon.$$

Show that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

minimizes RSS.

Question 2 (Linear Regression)

This question relates to the College data set, which can be found in the file `College.csv`. It contains a number of variables for 777 different universities and colleges in the US.

(from the previous HW) Use the `read.csv()` function to read the data into R. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.

Before moving on, we're not going to use the college name, so you may remove `X` variable from data.

Also, make sure categorical variables are set as factor variables.

Split your data into two parts: a testing data that contains 100 observations, and the rest as training data. You may use `sample` function to get the indices of the testing data. For this question, you need to set a random seed while generating this split so that the result can be replicated. Use `4322` as the random seed. Report the mean of `Outstate` of your testing data and training data, respectively.

(a) Now, split your training data into two parts: validation data (100 observations), and the rest as estimation data. Use the random seed `4323`.

(b) We're interested in predicting `Enroll`. First, run the linear regression on the estimation data including all variables. What is the feature variable with the highest p-value?

(c) Run the regression again, but this time, without that variable (with the highest p-value from previous regression) and find the feature variable with the highest p-value with the highest p-value in the new regression. Repeat this step until all the variables have p-value less than 0.1.

(d) Find validation MSE of all the models in (b) and (c). Report the model with the smallest validation MSE.

(e) Report your test MSE of your chosen model in part (d).

Question 3 (k-NN)

This question should be answered using the `Carseats` data set form `ISLR2` package.

Make sure all categorical variables are set as factor variables, and omit any missing data.

(a) Set 10% of whole data as a test set, and the rest as a training set. Split the training set into validation set (10% of training data) and the rest of the training set as a estimation set. Use the random seed `4324`.

(b) Conduct the EDA on the training set.

(c) We're going to fit linear regression models to predict `Sales` using `Price`, `US`, and `Advertising`.

Candidate models:

```
model 1: Sales~Price
model 2: Sales~US
model 3: Sales~Advertising
model 4: Sales~Price+US
model 5: Sales~US+Advertising
model 6: Sales~Price+Advertising
model 7: Sales~Price+US+Price*US
model 8: Sales~US+Advertising+US*Advertising
model 9: Sales~Price+Advertising+Price*Advertising
```

Store all regression models in one list. Run the regressions on the estimation data.

(e) Calculate validation MSE of all models. Choose a single model with the lowest validation MSE.

(f) Report your test MSE. Provide a scatter plot of predicted Sales and observed Sales of the test data.

Question 4 (k-NN and decision tree)

This question relates to the Boston data set of ISLR2 package.

```
set.seed(432)
trn.idx=sample(1:nrow(ISLR2::Boston),450)
tst.boston=ISLR2::Boston[-trn.idx,]
trn.boston=ISLR2::Boston[trn.idx,]
```

We are splitting the data into two parts: a testing data that contains 56 observations, and the rest 450 observations as training data.

- The goal is to model `medv` (our response variable) with all the other variables in the data.
- In this HW, we'll not worry about scaling variables. We'll tackle that in the future.

(a) Use the following validation-estimation split.

```
set.seed(1)
val.idx=sample(1:nrow(trn.boston),45)
val.boston=trn.boston[val.idx,]
est.boston=trn.boston[-val.idx,]
```

- Use the estimation data and `knnreg` function of `caret` package to perform KNN.
- Train KNN models using values of `k` from 1 to 100 and calculate validation MSE for each `k`.
- Plot the validation MSE versus `k` and show them in the same graph.

(b) Repeat (a) with different random seeds, (2,3), and see if your answer changes. If so, why does it change?

(c) Use the estimation/validation data from (a) with random seed (1) and `rpart` and `rpart.plot`function to perform decision tree.

- Start with default setting of R.
- Train decision tree models using cp=0, 0.001, 0.01, 0.1.
- Students may explore other tuning parameters as needed.
- Show your tree results using `rpart.plot` function.
- Compute validation MSE versus different cp values.
- Choose cp with lowest validation MSE.

(d) Repeat (c) with estimation/validation set with different random seeds, (2,3), and see if your answer changes. If so, why does it change?