STAT4001 Data Mining and Statistical Learning

Homework 1

Due on Oct.19 (Monday)

Instructions:

- Please write down **detailed** calculations and derivations to receive credits.
- Please include **all the R codes and outputs** (including numerical values, plots and so on) in your homework as **pdf form**. You may use R markdown to summarize, or you can simply print screen for ALL R codes and outputs.
- Please submit **only one single pdf file** through blackboard by 5pm on the date the assignment is due.
- File formats other than pdf form (both for the written part and R code, output part) will NOT be graded. Re-submission may be treated as late submission with partial marks deducted.

1. Consider the simple linear regression setting: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $RSS = \sum_{i=1}^{n}(y_i - \hat{y})^2$, $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$.
   Derive that $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = r^2$, where r is the correlation between X and Y (you may refer to lecture notes Chapter 2 page 15).

2. Cross-validation
   Consider the following dataset with three observations. $Y = (1.2, 1.8, 3.2)$, $X = (0.4, 0.8, 1.2)$ and the model $Y = \beta_0 + \beta_1 X$. Calculate the LOOCV (Leave-One-Out Cross-Validation) error.

3. Bootstrap
   Given the following data:

Observation	x	y
i=1	4.3	2.4
i=2	2.1	1.1
i=3	5.3	2.8

Use R *'sample'* function to generate 3 bootstrap samples and calculate the standard error (SE) of $\hat{\alpha}$. Please do NOT use any packages, write your own codes using the formula in the lecture notes Chapter 4 page 13$\sim$14.

Remark: Since the sample size is 3, which is very small, you may encounter (1,1,1), (2,2,2) or (3,3,3) when you generate bootstrap samples. In this case, you will get "NaN" in calculating $\hat{\alpha}$ due to the undefined denominator. You may just discard (1,1,1), (2,2,2) or (3,3,3) and generate one more sample. However, theoretically speaking, you can not simply discard the sample. In the real case, the sample size will be larger, and it will rarely encounter such numerical issue. This question is just a toy example to illustrate the idea of bootstrap.

4. Newton's method for finding MLE
Denote the likelihood as $L(\beta_0, \beta_1)$, and the log-likelihood as $l(\beta_0, \beta_1)$.
Given the training data $(x_i, y_i), i = 1, ..., n$, consider the logistic regression

$$P(y_i = 1 \mid x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}},\ P(y_i = 0 \mid x_i) = \frac{1}{1 + e^{\beta_0+\beta_1 x_i}}$$

We can rewrite this model into

$$P(y_i \mid x_i) = \frac{e^{(\beta_0+\beta_1 x_i)y_i}}{1 + e^{\beta_0+\beta_1 x_i}},\ y_i = 0, 1$$

Now we assume that $\beta_0 = c_0$ is known , our goal is to find the MLE for β_1 by finding the root of $l^{'}(\beta_1) = 0$.

(a) Write down the likelihood $L(\beta_1)$, log-likelihood $l(\beta_1)$, $l^{'}(\beta_1)$, $l^{''}(\beta_1)$ and the formula for updating $\beta_1^{(t+1)}$ given $\beta_1^{(t)}$ using Newton's method.

(b) Write R code to implement it. Stop the iteration if the difference between two successive likelihood is less than a cutoff (i.e. $|L(\beta_1^{(t+1)}) - L(\beta_1^{(t)})| \leq cutoff$). (Please do NOT use any packages in R for logistic regression or Newton's method, write your own code.)
Hint:

- *In practice, we will rewrite $\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$ as $\frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$ to avoid $\frac{\infty}{\infty}$, please do the similar adjustments when writing your own code.*
- *Use if-else conditional statement and function 'break' to stop the iteration.*
- *Use functions 'list' and 'save' to save multiple outputs.*

(c) Fix $\beta_0 = -0.66$, set $cutoff = 10^{-14}$, run your code using dataset 'HW1Q4data1.Rdata'. Report the trace of $\hat{\beta}_1$ and the trace of the log-likelihood.

(d) Fix $\beta_0 = 0$, set $cutoff = 10^{-4}$, run your code using dataset 'HW1Q4data2.Rdata' (this dataset is well-seperated) to see $\hat{\beta}_1$ diverges. Report the trace of $\hat{\beta}_1$ and the trace of the log-likelihood.

(e) Nonuniqueness of solution $\hat{\beta}_1$
Consider the following model, with $\boldsymbol{c}$ **is known**

$$P(y_i \mid x_i) = \frac{e^{\beta_1(x_i+c)y_i}}{1+e^{\beta_1(x_i+c)}},\ y_i = 0, 1$$

i. Write down the likelihood $L(\beta_1)$, log-likelihood $l(\beta_1)$, $l^{'}(\beta_1)$, $l^{''}(\beta_1)$ and the formula for updating $\beta_1^{(t+1)}$ given $\beta_1^{(t)}$ using Newton's method.

ii. Write R code to implement it. Stop the iteration when the difference between two successive likelihood is less than a cutoff (i.e. $|L(\beta_1^{(t+1)}) - L(\beta_1^{(t)})| \leq cutoff$). (Please do NOT use any packages in R for logistic regression or Newton's method, write your own code.)

iii. Fix $c = 0.5$, set $cutoff = 10^{-8}$, run your code using dataset 'HW1Q4data2.Rdata' to get $\hat{\beta}_1$. Report the trace of $\hat{\beta}_1$ and the trace of the log-likelihood.

(f) Write down the decision boundaries for (d) and (e)(iii).

Remark: The model in (e) is equivalent to the model in (a) with $\beta_0 = c\beta_1$. The model in (d) and the model in (e) give the same likelihood, but different decision boundaries. Hence, for well-seperated dataset, the parameter estimates for logistic regression may not be unique, and the classification performance will be affected by this numerical issue.

5. Implement K-Nearest Neighborhood through R

(a) Given a dataset $X_{n\times p}$ (n observations, p features), $Y_{n\times 1}$ (class labels, $y_i = 0, 1$), write a function to classify one new point x_{new} (i.e. to predict y_{new}). (Please do NOT use any packages in R, write your own code.)
Hint:

i. Find all the distances between x_{new} and $X_{n\times p}$

ii. Pick up class labels of the nearst K points using 'rank()'

iii. Predict y_{new} by majority vote

(b) Run your function using dataset 'HW1Q5data.Rdata' to classify x_{new}, with $K = 8$.

6. Use the 'Weekly data set' in textbook *'An Introduction to Statistical Learning, with Applications in R'* page 171, Question 10 to implement the following. Show the estimated parameters and compare the results of different methods through the predicted error.
The data set is called '$Weekly$' in R, and it is in the ISLR package. Please first install the package 'ISLR', and then run the code '$library(ISLR)$' to load the data.

(a) Logistic Regression

(b) Linear Discriminant Analysis

(c) Quadratic Discriminant Analysis

(d) K-Nearest Neighbors

Hint: You may refer to the tutorial notes 'Tutorial 02', or just follow the instructions in Question 10. If you have your own thoughts and ideas, or explore some new areas about these four methods, you are highly welcomed and encouraged to include new analysis!

- End -