Homework 2 (100 points)

MATH 6333 Statistical Learning

Due Sunday 10/3/2021 by 11:59 PM CT

All R files in this HW are attached with the assignment or in the same folder with this sheet.

Problem 1: (30 points) Write a detailed report to explain every step in the file

Lab 1(Three parts) - R.R that was covered in the fifth week. That is the code between

####################Part 2####################

####################End of file####################

in the R file.

Example of a <u>wrong</u> answer:

<u>Code line:</u> data_orange<-replicate(n=N/2 ,
c(mvrnorm(n=1,mean_orange[sample(nrow(mean_orange),1),],diag(2)/5),1))

<u>Explanation:</u> This step generates 100 orange points. (WRONG ANSWER)

Example of the least <u>correct</u> answer:

sample(nrow(mean_orange),1)

selects one random number from the labels of the rows of mean_orange

mean_orange[sample(nrow(mean_orange),1),]

returns the columns of that randomly selected row in the data frame "mean_orange"

mvrnorm(n=1,mean_orange[sample(nrow(mean_orange),1),],diag(2)/5)

generates one pair of coordinates or a point from a bivariate Gaussian distribution with mean given by the randomly selected row and with variance-covariance matrix given by the 2x2 identity matrix divided by 5.

c(mvrnorm(n=1,mean_orange[sample(nrow(mean_orange),1),],diag(2)/5),1)

concatenates the randomly generated point with a label of 1 for orange color

data_orange<-replicate(n=N/2 ,
c(mvrnorm(n=1,mean_orange[sample(nrow(mean_orange),1),],diag(2)/5),1))

repeats that last step with its nested/internal functions for N/2 times and save the results in data_orange

Problem 2: (20 points)

This problem is based on your understanding of the work done in Lab 1(Three parts) - R.R and similar data to that was generated there. You will produce four files of R codes, each of which is saved after the name of the role played, e.g., contestant 1.R.

In this problem, you will play four different roles: contestant 1, contestant 2, contestant 3 and the judge. Do the following:

1) Import the attached *training_data.xlsx* into R.
2) Take the role of contestant 1 who trains the linear regression using that data.
3) Take the role of contestant 2 who trains the K-nearest neighbor using that data and chooses the appropriate K using Monte-Carlo cross-validation.
4) Take the role of contestant 3 who trains the one-nearest neighbor using that data.
5) Import the attached *testing_data.xlsx* into R.
6) Take the role of the judge and test the three supervised learning methods for classification using the testing error function defined via the 0-1 Loss.
7) Who does win the contest?

Problem 3: (50 points) Chapter 3, Problem 13 on P.126 in ISL. See below.

Provide an R code and a report of the answers.

13. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.

    (a) Using the `rnorm()` function, create a vector, `x`, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X.

    (b) Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $N(0, 0.25)$ distribution—a normal distribution with mean zero and variance 0.25.

    (c) Using `x` and `eps`, generate a vector `y` according to the model

    $$Y = -1 + 0.5X + \epsilon. \tag{3.39}$$

    What is the length of the vector `y`? What are the values of β_0 and β_1 in this linear model?

    (d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

    (e) Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

    (f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.

    (g) Now fit a polynomial regression model that predicts `y` using `x` and $\texttt{x}^2$. Is there evidence that the quadratic term improves the model fit? Explain your answer.

(h) Repeat (a)–(f) after modifying the data generation process in such a way that there is *less* noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

(i) Repeat (a)–(f) after modifying the data generation process in such a way that there is *more* noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

(j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

———————————————————-Good Luck ———————————————-