STAT 6340 (Statistical and Machine Learning, Fall 2024)

Mini Project 2

Instructions:

- Due date: Sep 18, 2024 (Use Submission Link on eLearning to upload by 11:59 pm)
- Total points = 30
- Submit a typed report.
- It is OK to consult with others regarding this project or use any publicly available resources, but you must write code and answers on your own. Violation of this policy will be considered academic dishonesty, and you will referred to appropriate university authorities. If you don't do the work by yourself, how will you learn?
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Section 1. Answers to the specific questions asked
  Section 2: Code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.
- Section 1 of the report must be limited to *four* pages. Also, only those software outputs should be provided in this section that are referred to in the report.
- You can code in `R` or `Python`.
- You are encouraged to use `R Markdown` or `Jupyter Notebook` to write your report.
- Bonus (4 points): Use *both* `R` and `Python` to code, compare the results, and provide an explanation if the results are not identical. This work should be described on a separate page titled Bonus Work with two sections — one to describe the comparison and the other for the code. The first section should be limited to one page.

---

1. (12 points) Consider the forest fires dataset available on eLearning as `forestfires.csv`. The dataset is based on the forest fires in the Montesinho Natural Park in Portugal. It consists of 517 instances, with each instance representing a forest fire event. A description of the variables is given in Table 1. We would like to understand how the burned area of the forest (`area`) is related to the other predictors in the dataset.

   (a) Perform an exploratory analysis of data and report the findings that interest you.

   (b) Is `area` appropriate as a response variable or a transformation is necessary? In case a transformation of response is necessary, try the natural log transformation and use it for the rest of this problem. (Note: If `area` has any zero values in the data, add 1 to `area` before applying the log transformation.)

   (c) Do part (a) of Exercise 15 in Chapter 3 for these data.

   (d) Do part (b) of Exercise 15 in Chapter 3 for these data.

Header	Description
X	x-axis spatial coordinate within the Montesinho park map: 1 to 9
Y	y-axis spatial coordinate within the Montesinho park map: 2 to 9
month	month of the year: 'jan' to 'dec'
day	day of the week: 'mon' to 'sun'
FFMC	FFMC index from the FWI system: 18.7 to 96.20
DMC	DMC index from the FWI system: 1.1 to 291.3
DC	DC index from the FWI system: 7.9 to 860.6
ISI	ISI index from the FWI system: 0.0 to 56.10
temp	temperature in Celsius degrees: 2.2 to 33.30
RH	relative humidity in %: 15.0 to 100
wind	wind speed in km/h: 0.40 to 9.40
rain	outside rain in mm/m2 : 0.0 to 6.4
area	the burned area of the forest (in ha): 0.00 to 1090.84

Table 1: List of variables in the forest fires data

(e) Build a "reasonably good" multiple regression model for these data. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions.

(f) Write the final model in equation form, being careful to handle qualitative predictors (if any) properly.

(g) Use the final model to predict the burned area of the forest for a scenario where the quantitative predictors are set to their sample means and qualitative predictors (if any) are set to the most frequent category.

(h) Are there any outliers in the data on which the model in (g) was built? If so, assess the impact of the outliers on the estimated coefficients by refitting the model without the outliers. Which model would you finally suggest?

2. (10 points) Consider the Pima Indians Diabetes dataset. All patients in the dataset are of Pima Indian heritage (subgroup of Native Americans) and are females of ages 21 and above. The dataset consists of 768 observations with 9 variables and is divided into a training set (700 observations) and a test set (68 observations). The binary response variable is `Outcome` (1: diabetes present, 0: otherwise). There are 8 quantitative predictors. Standardize the predictors before performing any analysis for (b)–(e).

(a) (8 points) Perform an exploratory analysis of the data by examining appropriate plots and comment on how helpful these predictors may be in predicting response.

HW2 HW3

(b) Perform an LDA on the training data. Compute the confusion matrix and overall misclassification rate based on both training and test data.

(c) Repeat (b) using QDA.

(d) Repeat (b) using naive Bayes classifier.

(e) Repeat (b) using KNN with K chosen optimally on the test set. Do you see any issues with how this K is chosen?

(f) Compare the results in (b)–(e). Which method would you recommend? Justify your conclusions.

3. Consider the business school admission data available on eLearning as `admission.csv`. The admission officer of a business school has used an "index" of undergraduate grade point average (GPA, X_1) and graduate management aptitude test (GMAT, X_2) scores to help decide which applicants should be

admitted to the school's graduate programs. This index is used categorize each applicant into one of three groups — admit (group 1), do not admit (group 2), and borderline (group 3). We will take the first five observations in each category as test data and the remaining observations as training data.

(a) Perform an exploratory analysis of the training data data by examining appropriate plots and comment on how helpful these predictors may be in predicting response.

HW 2 QUES 2

(b) Perform an LDA using the training data. Superimpose the decision boundary on an appropriate display of the data. Does the decision boundary seem sensible? In addition, compute the confusion matrix and overall misclassification rate based on both training and test data. What do you observe?

(c) Repeat (b) using QDA.

(d) Compare the results in (b) and (c). Which classifier would you recommend? Justify your conclusions.