Math 6350

fall 2019 MSDS

Homework 1

On the website www.StatLearning.com

**download** the data set "Auto.csv" (in .csv format, which is essentially equivalent to excel format)

the column names are

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| mpg | cylinders | displacement | horsepower | weight | acceleration | year | origin | name |

**preliminary treatment of the data set**

discard the last three columns (year, origin, name)

find and discard the very few rows which contain missing numerical data

indicate the **number N of cases** = number of numerical rows which are kept

**mpg** = miles per gallon will be the **target** variable

the 5 explanatory variables or **features** will be

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | cylinders | displacement | horsepower | weight | acceleration |

we will denote them **cyl, dis, hor, wei, acc**

1) compute the mean and standard deviation of each feature

2) display the histogram of each feature, and the histogram of mpg

3) display the 5 following scatterplots

(cyl , mpg) , (dis , mpg) , (hor , mpg) , (wei , mpg) , (acc , mpg)

4)

interpret these 5 scatterplots to guess which features may have stronger capacity to predict msg

5) compute the 5 correlations

cor(cyl , mpg) , cor(dis , mpg) , cor(hor , mpg), cor(wei , mpg), cor(acc , mpg)

interpret these correlations to guess which features may have stronger capacity to predict msg

6) compute the covariance matrix COV and the correlation matrix CORR of the 5 features

7) compute the 5 eigenvalues L1 >L2>L3>L4>L5 of CORR

8) verify that L1+L2 +...+L5 =5

9) for i =1 2 3 4 5 compute the ratios Ri = (L1+L2+ ... +Li)/5

10) interpret these 5 ratios

11) reorder the rows of the .csv data set so that the first column msg becomes increasing

separate then the data set into two tables,

the LOWmsg table will include all the cases for which msg is inferior to median(msg)

the HIGHmsg table will include all the cases for which msg is larger than median(msg)

12)

Let F be any one of the five features **cyl, dis, hor, wei, acc**

display side by side

the histogram histlow(F) computed on the F values corresponding to the cases belonging to the table LOWmsg

the histogram histhigh(F) computed on the F values corresponding to the cases belonging to the table HIGHmsg

This will give you 5 pairs of histograms, one pair for each feature F

13) interpret each one of these 5 pairs of histograms to guess which feature has a good capacity to discriminate between high msg and low msg

14) successively, for each one of the five features (denoted F)

compute the mean mlow(F) and standard deviation stdlow(F) of the F values corresponding to the cases belonging to the table LOWmsg

compute the mean mhigh(F) and standard deviation stdhigh(F) of the F values corresponding to the cases belonging to the table HIGHmsg

15) compute | mhigh(F) - mlow(F) | / s(F) where s(F) = (stdlow(F)+ stdhigh(F) )/sqrt(N)

use these numbers to rughly evaluate the capacity of feature F to discriminate between low msg and high msg

Write code to confirm and check the answer and add comments/interpretations.