Math 6350 fall 2021 Homework 1 (HW1)

due date : sunday sept12 , 2021 midnight ; email HW1 reports to robertazencott@gmail.com

the data set "cleanDataAuto.csv" will sent to the class by email ; you can also download the "auto" data set (with a few more irrelevant columns) from the website www.StatLearning.com ; the column names in"cleanDataAuto.csv" are

mpg cylinders displacement horsepower weight acceleration name

discard the last column ("name" = carmodel name);

display the **number M of cases** = number of numerical rows

each case is recorded by one row

mpg = miles per gallon will be the **target** or **response** variable ;

the data table displays one column of M values for mpg and 5 columns of M values for the 5 explanatory variables or **features**

features will be denoted by **F1= cyl, F2= dis, F3= hor, F4= wei, F5= acc**

Preliminary statistical data analysis

1) for each feature F compute/display its mean mF, standard deviation stdF, rangeF,

2) display the histogram histF of each feature F, and the histogram hist(mpg) of mpg

display the probability density function (PDF) of a normal density function with mean mF and standard deviation stdF

compare visually histF and this PDF

3) display the 5 **scatterplots** (cyl , mpg) , (dis , mpg) , (hor , mpg) , (wei , mpg) , (acc , mpg)

display the table of 5 correlations

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

interpret the 5 scatterplots and the table of 5 correlations to guess which features may have stronger capacity to predict msg

4) compute the 5x5 correlation matrix CORR(Fi,Fj) of the 5 features; interpretation of that matrix

5) Display the quantiles Q1% Q2% ... Q100% of the response variable **mpg** as an increasing quantile curve.

6) compute 5 linear regressions Y= A1*F1 +B1 +errorterm ….. Y= A5*F5 +B5 +errorterm

give the values of A1,B1, RMSE1 … A5,B5, RMSE5 this defines 5 linear predictors of Y denoted PRED1Y= A* F1 +B1 , ….. , PRED5Y= A5*F5 +B5

Compute the relative accuracies of these 5 predictors : RMSE1 / mean(Y) , … , RMSE5 / mean(Y)

display each linear graphs PREDjY versus Fj on same graph as corresponding scatterplot (Fj,Y)

Interpretation of these results

Automatic Classification of data by kNN technique

8) extract from the data set three disjoint tables of cases ,

LOWmpg table = {all cases for which mpg <= quantile Q33%}

MEDmpg table = {all cases for which quantile Q33% < mpg <= quantile Q66%}

HIGHmpg table = {all cases for which mpg > quantile Q66%}

9) For each feature F= F1 , ..., F5 , display side by side ,

the histogram hist.low(F) of F values for all cases in LOWmpg

the histogram hist.high(F) of F values for all cases in HIGHmpg

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

interpret each such pair of histograms to guess which features may have a good capacity to discriminate between high mpg and low mpg

10) for each feature F ,

compute the mean mL and standard dev. stdL of F values for all cases with LOWmpg;

compute the mean mH and standard dev. stdH of F values for all cases with HIGHmpg

compute 90% confidence intervals around mL and mH and compare them to evaluate the "power" of feature F to discriminate between low mpg and high mpg ;

compare these qualitative discriminating powers between the five feature

11*) application of the automatic classifier* kNN :

randomly partition 80%/20% each one of the 3 classes CL1= LOWmpg, CL2 =MEDmpg, CL3 =HIGHmpg

regroup these three partitions to construct then a global trainingset TRAIN and a global test set TEST of sizes 80% xM and 20%xM

fix k= 5

apply the kNN algorithm to this data set compute the two accuracies AccTrain and AccTest of automatic classification by kNN on the two sets TRAIN and TEST

compare these two accuracies

12) repeat the preceding kNN implementation for k = 3, 5, 7 , 9, 11, 13 , 15, 17 , 19, 29 , 39

plot on same graph the two accuracy curves AccTrain and AccTest versus k to select a best value for k