Homework 1

PStat 131/231 - Spring 2016

Due April 15th, 2016

Instructions

Use **R Markdown** to produce your solution to the hw and provide a pdf version of it.

Download the **HW template.rmd** from GauchoSpace and open it in R-studio. Insert your text and R code (the R code in an R chunk). Answer all questions in the order given. Clearly show the structure of your answers by recalling the specific question you're answering .

Use the packages `ggplot2` and `data.table`.

Question 1

Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their profitability.
(b) Computing the total sales of a company.
(c) Predicting the future stock price of a company using historical records.
(d) Sorting a student database based on student identification numbers
(e) Predicting the outcomes of tossing a (fair) pair of dice.

Question 2

Consider the Boston housing data http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data from the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/ . Data described at http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names

(a) Describe this data set: how many observations? how many variables (or attributes)?, what is the unit analyzed? . . .

(b) Load the data into R and can call it Boston.Housing. Consider the fact that the columns of the dataset are unequally separated by white spaces. Also, to keep standard the analysis use

```r
c("Crime.Rate","ResiLand.Zoned","NonRetail.Bus","Charles.River","Nitr.Oxide","Avg.Rooms",
"Age", "Wigh.Dist","Access.Idex","Tax","Pupil.Teacher","Blck","Lower.Sts","Med.Value")
```

```
##  [1] "Crime.Rate"     "ResiLand.Zoned" "NonRetail.Bus"  "Charles.River"
##  [5] "Nitr.Oxide"     "Avg.Rooms"      "Age"            "Wigh.Dist"
##  [9] "Access.Idex"    "Tax"            "Pupil.Teacher"  "Blck"
## [13] "Lower.Sts"      "Med.Value"
```

as column names.

(c) Produce a histogram of the median value of owner-occupied homes with the title "Histogram of median home value based on Boston Housing Data". Using `binwidth` argument, gradually increase the number of bins (create four different histograms). What happens to the histogram?

(d) Show all histograms plot in one chart.

(e) Using R, compute mean, median, standard deviation and interquartile range of the median home value. What is a good measure of center and spread of your data? Explain why. Note that you are asked to compute median of the median home value. Does this make sense? Explain.

(f) Create 5 equally distributed ranks of Crime.Rate variable. Then use a boxplot to analyze if the median value of the house significantly differs across the levels of each rank of crime rate by town. Hint: Use the `quantile()` function.

Question 3

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set, which is part of the ISLR package. You will also need to download class package for part d).

(a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Make sure that you make mpg01 a factor variable. Also, use the `data.table()` function to create a single data set containing both mpg01 and the other Auto variables.

(b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots and other graphical devices discussed in section may be useful tools to answer this question (you should at least include 3 different graphs). Describe your findings.

(c) Split the data into a training set (75%) and a test set (25%). Call them train.set and test.set, respectively. The `sample()` command may be useful for answering this question.

(d) Using train.set and test.set perform k-NN on the training data, with several values of k, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b) (Justify). What test errors do you obtain? Which value of K seems to perform the best on this data set?

Additional exercises for PStat 231

Question 4

Read the paper "From Data Mining to Knowledge Discovery in Databases" (AI Magazine 17 (3), 1996) which you can find uploaded in GauchoSpace. The paper is 20 years old however it provides a good introduction to the problems we are discussing in the course. After reading, answer the following questions. Be as concise as possible: your answers (all of them) should not take more than one page (with formatting as this homework).

(a) Give your definition of KDD and Data Mining. How are KDD and Data Mining related?

(b) Why, according to the authors, there is a need for data mining?

(c) Among the real applications described in the paper, pick one of your interest and try finding, in the references section, a paper related to the application you're interested in. Get that paper (try inserting the full title of the paper in Google Scholar and typically link to the paper will show or, otherwise, check in the library) and describe with some detail the application discussed therein.