STA 545 Statistical Data Mining I, Fall 2020

Homework 1

Stella Liao

September 9, 2020

1. (40 points) Read the materials in the R/Rstudio/Rmarkdown folder in UBlearns.

(d) Change the template to show the information about the wage dataset used in the ISLR textbook. In the ISLR package, the name of this dataset is Wage.

```
library(ISLR)
summary(Wage)
```

```
##       year           age                     maritl           race
##  Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480
##  1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293
##  Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190
##  Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37
##  3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##                education                     region               jobclass
##  1. < HS Grad      :268   2. Middle Atlantic   :3000   1. Industrial :1544
##  2. HS Grad        :971   1. New England       :   0   2. Information:1456
##  3. Some College   :650   3. East North Central:   0
##  4. College Grad   :685   4. West North Central:   0
##  5. Advanced Degree:426   5. South Atlantic    :   0
##                           6. East South Central:   0
##                           (Other)              :   0
##             health      health_ins      logwage           wage
##  1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09
##  2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38
##                                      Median :4.653   Median :104.92
##                                      Mean   :4.654   Mean   :111.70
##                                      3rd Qu.:4.857   3rd Qu.:128.68
##                                      Max.   :5.763   Max.   :318.34
##
```

2. (30 points) Explain whether each scenario is a classification or regression problem. In addition, please provide the sample size n and the number of independent variables p for each scenario.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

regression
n = 500
p = 3

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

classification
n = 20
p = 13

(c) We are interest in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

regression
n = 52
p = 3

3. (30 points) You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Please do not use the examples shown in the slides.

1. Given some socio-economic factors(which are the predictors), like education, gender, race, etc, to predict whether there will be a crime occurring or not(which should be the response).
2. Based on some attributes(which are the predictors) like, salary, debts, age, etc, to categorize applications for new cards into those who have a good credit, bad habit, or fall into a gray area requiring more human analysis(which should be the response).
3. To classify animal species in images(which should be the response), we applied statistical features of raw pixel data, shape and color feature maps and transform coefficient or vectors[1] as predictors.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Please do not use the examples shown in the slides.

1. To predict compressive strength of cement(which should be the response), we applied the amount of cement, age and water content as predictors.
2. To predict water quality index(which should be the response), we applied pH values, the content of Ca, Mg, HCO_3, SO_4, PO_4, etc as predictors.
3. To predict GDP of different countries (which should be the response), we applied education, population, average income, etc as predictors.

(c) Describe three real-life applications in which cluster analysis might be useful. Please do not use the examples shown in the slides.

1. Crime hot spots, which is to identify areas where there are a larger amount of incidences of one particular crime type
2. To identify potentially dangerous zones by clustering observed earthquake epicenters
3. To identify different groups of plants by their features, like their tolerance to soil contamination[2], geographic locations, etc.

References

[1] Alharbi, Fahad et al. "Animal Species Classification Using Machine Learning Techniques". MATEC Web Of Conferences, vol 277, 2019, p. 02033. EDP Sciences, doi:10.1051/matecconf/201927702033.

[2] Potashev, K., Sharonova, N. and Breus, I., 2014. The use of cluster analysis for plant grouping by their tolerance to soil contamination with hydrocarbons at the germination stage. Science of The Total Environment, 485-486, pp.71-82, https://doi.org/10.1016/j.scitotenv.2014.03.067