Assignment 3

Deadline: May 29, 2021, 11.59 pm

1. Use R to solve the following question. Consider the following training set for a loan classification problem below.

Record	Age	Annual Income	Years of Education	Default (Loan)
1	59	$132,000	18	No
2	41	$85,000	14	Yes
3	29	$60,000	16	Yes
4	25	$35,000	10	Yes
5	49	$170,000	12	No
6	30	$70,000	9	Yes
7	33	$45,000	13	Yes
8	29	$110,000	12	No
9	43	$45,000	12	No
10	56	$112,000	16	No
11	28	$65,000	12	Yes

(a) Standardize the Age, Annual Income and Years of Education variables

(b) Based on the standardized variables, calculate the Euclidean distances between a 45-year old person with 15 years of education and $60,000 annual income and the 11 observations in the record.

(c) Use the K-Nearest Neighbor approach and set K =5 to find out whether a 45-year old person with 15 years of education and $60,000 annual income will default.

2. Consider the following transaction table.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers, Milo}
2	{Bread, Eggs, Milk, Milo}
3	{Milk, Diapers, Eggs}
4	{Beer, Eggs, Diapers, Coke}
5	{Beer, Diapers, Coke }
6	{Milk, Eggs, Bread, Butter, Milo}
7	{Milk, Eggs, Coke}
8	{Eggs, Bread, Coke}
9	{Coke, Milo, Instant Noodles}
10	{Instant Noodles, Coke, Bread}

(a) Explain Support and Confidence and their significances in Market Basket Analysis.

(b) What is the maximum size of frequent itemsets that can be extracted (assuming *minsup* > 0.1)?

(c) Consider *{Diapers}→{Beer}*. Calculate the support and confidence of the rule.

3. Use the Bank dataset.

(a) Select the following eight variables: PersonalLoan, Age, Experience, Income, Family, CCAvg, Education, and CD.Account. Save them in mydata data frame.

[Hint: To select variables from a data frame, we could use the "select" function. Load library(dplyr) and select(name of the data frame, variable 1, variable 2,….)]

(b) Calculate the descriptive statistics for these eight variables. Explain your findings.

(c) Eliminate all the unreasonable data points for Experience variable.

(d) Randomly draw 4500 observations to form a test set and let the rest be the validation set.

(e) Use PersonalLoan as the Y variable and the rest as X variable. Apply the KNN approach to the test and validation sets in the part (d).

(f) Calculate the accuracy rate for your model in the part (e)

4. [Classification Problem]

Refer to https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Pima.tr.html.

You need to load(MASS) and call the dataset out, i.e, mydata<-Pima.tr

(a) Use the summary function to caclulate the descriptive statistics

(b) Draw histograms and boxplots for bmi, bp, and age. Which variables are normally distributed?

(c) Apply the LDA approach to the Pima dataset. Use your LDA model to predict if one would have diabetes and save your results in mydata.lda.

(d) Print out mydata.lda$class. What are these?

(e) Print out mydata$type. What are these?

(f) According to (d) and (e), calculate the accuracy rate.

(g) Use the QDA approach and re-do (c) and (f).

(h) Do you prefer to use LDA or QDA in this classification problem? Why?

(i) Explain in detail the differences between LDA, QDA, and LPM.