Data Mining and Business Analytics

Homework 4

In this homework you will do exercises in SVM, Model Comparison, and Clustering. Please follow the instruction given in the homework and answer the question.

Submission format:

1. Please include your answers and the corresponding codes/plots/outputs in **one single word doc or pdf file**. If you are using R markdown that is great, you can submit the pdf output file. If you are not using R markdown, please include the corresponding codes/plots/outputs by screenshots or copy&paste.
2. Please also submit your **R script or R markdown file**. But please note that TA is not responsible to run your code to obtain any output to support your answers.

Notes:

3. This homework is a group assignment.
4. If you encounter any challenges, before coming to TA or me:
   a. Make sure you go through the "R and R helps.pdf" (posted under **R HELPS** Module on Canvas). Please pay special attention to "Few Notes Regarding R" and "Few Tips When R Is Not Working".
   b. Ask Online using the following links
      i. Stackoverflow - http://stackoverflow.com/ - coding and language related questions

Additional Files on canvas for this homework:

1. **Part A: credit_data.csv, new_customers.csv**
2. **Part C: prospects.csv**

For the next part, we will be using the bank loan default data (credit_data.csv). Note that, It is the same dataset you worked on for Part B in HW3.

Dataset Description

Sr. No.	Column	Description
1	months_loan_duration	Duration of the loan
2	credit_history	Previous credit history
3	purpose	Purpose of the loan
4	amount	Amount borrowed
5	employment_duration	Present employment since
6	percent_of_income	Installment rate in percentage of disposable income
7	years_at_residence	Present residence since
8	age	Age of the borrower in years
9	other_credit	Any other credit
10	housing	Housing type of borrower
11	existing_loans_count	Number of existing credits at this bank
12	job	Job type
13	dependents	No, of dependents of the borrower
14	phone	If the borrower has a phone or not

15	default	Classifier-if the Borrower defaulted or not

A) Comparison of Decision Tree, K nearest neighbor, Support vector Machine and combining results. (50 PTS)

0. Read in the data "credit_data.csv" using the following line (note that, "stringsAsFactors = TRUE" ensures the strings are loaded as factors):

```
credit_data  <- read.csv("credit_data.csv", stringsAsFactors = TRUE)
```

And convert variables "credit_history" and "employment_duration" to ordered factors using the following codes:

```
credit_data$credit_history <- ordered(credit_data$credit_history,
                                      levels = c("critical", "poor", "good","very good","perfect"))
credit_data$employment_duration  <- ordered(credit_data$employment_duration,
                                levels = c("unemployed" , "< 1 year", "1 - 4 years", "4 - 7 years","> 7 years"))
```

1. For each of the algorithms - Decision Tree (method="C5.0"), K nearest neighbor (method="knn"), Support vector Machine (method="svmLinear"), obtain accuracy based on confusionMatrix.
   Notes:
   1) Please reset seed begin running each method. In particular:
      - Set seed to 100 for decision tree
      - Set seed to 100 for kNN
      - Set seed to 23 for SVM
   2) Please use 10-fold Cross-validation for all algorithms.
   3) Please set preProcess = c("center", "scale") for all algorithms to preprocess the data by normalization.
   4) For this question, there is no need to reserve a testing sample.
   5) For **KNN**, implement KNN based on K=1 to 50 (as you did in HW3 Question B)
   6) For **SVM linear**, implement with "Cost" parameter being 0.01, 0.1, 1, and 10 by seting tuneGrid=expand.grid(C = c(0.01,0.1,1,10))
   7) For **decision tree** using method = "C5.0" please use the blue lines listed below, since we did not cover this in class. Please use the help function or look at the caret library pdf < https://cran.r-project.org/web/packages/caret/caret.pdf> for further details if you are interested.

```
C5.0_credit <- train(default ~ ., data=credit_data,
                     method = "C5.0",
                     preProcess = c("center", "scale"),
                     tuneGrid = expand.grid( .winnow = c(TRUE,FALSE), .trials=1, .model="tree"),
                     trControl = trainControl(method = "cv", number=10))
```

   8) Use the default accuracy as metric, so no need to set metric explicitly.

2. For each algorithms - Decision Tree, K nearest neighbor, Support vector Machine, also draw insample ROC curve on a same graph.

3. Which algorithm would you choose based on AUC and why?

4. Given 5 customers (in a separate file-"new_customers.csv", set "stringsAsFactors = TRUE" and convert variables "credit_history" and "employment_duration" to ordered factors as you did in step-0), predict their outcome using the three algorithms you trained in step-1 and report predictions. Based on predictions from three algorithms what would you decide for each of these customers using majority voting rule (i.e., predict based on the label suggested by most algorithms)?

5. Compare each individual algorithm prediction accuracy vs majority voting rule accuracy with the true y label (i.e., values in column “default”) given in the new_customers.csv

B) Clustering (30 pts)

Take the following points in two-dimensional space:

(8, 4), (3, 3), (4, 5), (0, 1), (10, 2), (3, 7), (0, 9), (8, 1), (4, 3), (9, 4).

For this exercise, use the Manhattan distance metric:

For instance, the distance from (3,3) to (8,1) is $|3 - 8| + |3 - 1| = 7$.

Beginning with centroids at (1,1) and (8,8), perform two iterations of the 2-means clustering algorithm, that is:

- Iteration 1: Allocate the points to centroids (using Manhattan distance), then find the new centroids.
- Iteration 2: Again allocate the points to the centroids, and then get the new centroids.

Tip: Centroids are defined as the centers within clusters. For instance, if a clusters contains 3 data points, (1,1.5), (2,2.5) and (3,3.5), then the centroid for this cluster is: ((1+2+3)/3,(1.5+2.5+3.5)/3) which is (2,2.5).

If a point is equidistant between the centroids, assign it to the centroid that starts at (1,1). For each iteration, provide a concise table that show (point x, point y, distance to centroid 1, distance to centroid 2, cluster it belongs to). Note that the centroids might be changing from iteration 1 to iteration 2.

C) Clustering with Data (20 pts)

Download the file **prospects.csv**. Columns are as follows.

Name	Model Role	Measurement Level	Description
AGE	Input	Interval	Age in years
INCOME	Input	Interval	Annual income in thousands
MARRIED	Input	Binary	1=married, 0=not married
GENDER	Input	Binary	F=female, M=male
OWNHOME	Input	Binary	1=homeowner, 0=not a homeowner
LOCATION	Rejected	Nominal	Location of residence (A-H)
CLIMATE	Input	Nominal	Climate code for residence (10,20,30)
FICO	Input	Interval	Credit score
ID	ID	Nominal	Unique customer identification number

Notes:

- Use na.omit to remove all rows with NAs.
- Convert GENDER into a dummy variable, i.e., 0 or 1.
- In your clustering model exclude the columns LOCATION and ID.
- Use preProcess() to standardize the data before running Kmeans.

1. Set seed to 100 and k=4. How many points are in each cluster?
2. For each of the four clusters, briefly describe the characteristics of members of that cluster based on the cluster centroid along two features: Gender and Married.