Applied Data Mining: Midterm Exam I

Due on 10/02/2017, 10:00pm (ET)

Instructor: Hasan Kurban

Student Name

September 29, 2017

Directions

This midterm exam is due Monday Oct 2, 2017 10:00p.m (ET). **OBSERVE THE TIME**. Absolutely no midterm exam will be accepted after that time. All the work must be your own. I am providing the LaTeX of this document too. You are not allowed to post questions related to the midterm exam on Canvas (Piazza/Discussion). The last question is a bonus question. If you think that any of the questions is ambiguous, answer it as you understand and explicitly explain your approach.

Problem 1 (20 pt.)

Load mydata.txt into R and answer the following questions:

1. How many entries are in the data set? Answer here ...
2. How many unknown or missing data are in the data set? Remove the tuples with the missing values and call this data, clean.mydata. You will use clean.mydata to answer the questions below:

R script to find missing data

Listing 1: Sample R Script With Highlighting

```
%% You provide code here %%
```

Count of missing values

Answer here...

3. Calculate mean and median of variable V2? Answer here ...
4. Find variance, standard deviation and interquartile range of variable V4? Answer here ...
5. Create a bar plot that shows count of data points for classes "1" and "2" (variable 5). Is the data skewed?

R script

Listing 2: Sample R Script With Highlighting

```
%% You provide code here %%
```

Bar plot

Place images here with suitable captions.

Discussion of Data

Answer here...

6. Create two scatter plots using (1st, 2nd) and (1st, 3rd) variables and color the data points with the class variable (5th variable). Discuss the plots? Is there any pattern?

R script

Listing 3: Sample R Script With Highlighting

```
%% You provide code here %%
```

Scatter plots

Place images here with suitable captions.

Discussion of Scatter Plots

Answer here...

Problem 2 (20 pt.)

Apply PCA to the clean.mydata and answer the following questions: (Remove the class variable, V5, before PCA)

1. How many principal components explain 90% of the variance? Answer here...

R script

Listing 4: Sample R Script With Highlighting

```
%% You provide code here %%
```

2. What are loadings in PCA? Observe loadings and express the principal components using the original variables.

R script

Listing 5: Sample R Script With Highlighting

```
%% You provide code here %%
```

Discussion and results

Answer here...

3. Make a scree plot. Discuss the plot, i.e., what is a scree plot? What is the optimal number of dimensions based on the plot?

R script

Listing 6: Sample R Script With Highlighting

```
%% You provide code here %%
```

Scree plot

Place images here with suitable captions. =

Discussion

Answer here...

4. Make a scatter plot of PC2 and PC3. Do you observe any relationship? i.e., Calculate the correlation between PC2 and PC3? What does it show?

R script

Listing 7: Sample R Script With Highlighting

```
%% You provide code here %%
```

Scatter plot

Place images here with suitable captions.

Discussion

Answer here...

Problem 3 (20 pt.)

Run the R code below:

```
kmeans.mydata <- clean.mydata[,c(1,2)])
```

1. Randomly sample without replacement 300 data points from kmeans.mydata. (call the sampled data, mysample). Cluster mysample with K-means. Include the R code and answer the questions below:

R script

Listing 8: Sample R Script With Highlighting

```
%% You provide code here %%
```

2. Explain *iter.max* and *algorithm* parameters of kmeans function in R and run k-means on mysample data set where $nstart = 35$ and $k = 2$. Report total within squares error and within squares error for each cluster.

R script

Listing 9: Sample R Script With Highlighting

```
%% You provide code here %%
```

Discussion and results

Answer here...

3. Make a plot of data points and color the observation according to the cluster labels obtained.

R script

Listing 10: Sample R Script With Highlighting

```
%% You provide code here %%
```

Cluster plot

Place images here with suitable captions.

4. Run k-means on mysample data set where $nstart = 35$ and $k = 4$. Report total within squares error and within squares error for each cluster.

R script

Listing 11: Sample R Script With Highlighting

```
%% You provide code here %%
```

Results

Answer here...

5. Make a plot of data points and color the observation according to the cluster labels obtained.

R script

Listing 12: Sample R Script With Highlighting

```
%% You provide code here %%
```

Cluster plot

Place images here with suitable captions.

6. Compare (2) and (4). Answer here...

Problem 4 (20 pt.)

The file `rainfalldataraw.txt` are data collected in a cloud-seeding experiment over the span of approximately seven years using five sites. Total rainfall is collected for a fixed, arbitrary, and uniform period of time where seeding is compared to unseeding. In particular,

- SEEDED is a Boolean {U,S}, U for unseeded and S for seeded.
- SEASON is a string indicating one of four seasons {FALL, WINTER, SUMMER, SPRING}
- The remaining columns {A,B,C,D,E} are the sites of experiments and the total rainfall recorded.

1. In the listing below, which line number effectively reads the data into an R data.frame?

```
teach <- read.table(file="c:/R/rainfalldataraw.txt", header=FALSE, sep=",")
teach <- read.table(file="c:/R/rainfalldataraw.txt", header=TRUE, sep=",")
teach <- read.table(file="c:/R/rainfalldataraw.txt", header=TRUE, sep="")
```

   Answer here...

2. Give a select operation on the data.frame that gives the rows whose E variable values are greater than 4, but less than 5.

R script

Listing 13: Sample R Script With Highlighting

```
%% You provide code here %%
```

3. Give the code that produces the histogram of variable D.

R script

Listing 14: Sample R Script With Highlighting

```
%% You provide code here %%
```

4. How many tuples (or records) are in the data? Answer here...

5. Identify the data that is either missing or likely corrupted.

R script to find missing data

Listing 15: Sample R Script With Highlighting

```
%% You provide code here %%
```

Results for missing data

Answer here...

6. Preprocess the data, addressing the problems above and save the file as `rainfixed.txt` as a csv file. Explain explicitly what you have done in preprocessing this file.

R script for preprocessing data

Listing 16: Sample R Script With Highlighting

```
%% You provide code here %%
```

Results and discussion for preprocessing step

Answer here...

7. Using any techniques you've learned, answer this question to a policy maker:

   > When there isn't enough rain, we lose a substantial amount of tax revenue, because farmers lose crops and livestock. We think this cloud-seeding works and want to seed these areas A,B,C,D,E when the rainfall is significantly less than average. What does your datamining tell us?

   State any explicit assumptions that you need to answer the question thoughtfully and fully. You should provide ample evidence of a thorough analysis of the data. Answer here...

Problem 5 (10 pt.)

Assume four pieces of data $x_1 = (.5, 2000, -100); x_2 = (.2, 3000, -200); x_3 = (4, 4000, -100), x_4 = (.14, 4400, -140)$. You've been hired to datamine this data using Euclidean distance. How would you preprocess this *before* datamining and explain why. What are the two closest data points?

R script

Listing 17: Sample R Script With Highlighting

```
%% You provide code here %%
```

Discussion an Results

Answer here...

Problem 6 (5 pt.)

You're given a sample of data: 15,2,44,21,40,20,19,18. Calculate the sample mean and sample variance. Answer here...

R script

Listing 18: Sample R Script With Highlighting

```
%% You provide code here %%
```

Problem 7 (5 pt.)

Choose *all* that apply. Which of the following statistical measures can be observed on a box plot?

(a) Mode

(b) Mean

(c) Median

(d) Outliers

(e) Noise

(f) Maximum element

(g) Minimum element

(h) Variance or covariance.

Answer here ...

Problem 8 (5 pt.)

Choose *all* that apply. The most common methods of removing outliers are:

(a) Removing tuples with missing values.

(b) Duplicating tuples without missing values.

(c) Observing the probability of existing values in

(d) Assuming a uniform distribution on existing values, then populating values using this probability.

(e) Putting in `Null` and ignoring the value in the operations.

Answer here ...

Problem 9(10 pt.)

Swiss bank data contains various lengths measurements on 200 Swiss bank notes. Load the Swiss bank data as follows:

```
> install.packages("alr3")
> library("alr3")
> head(banknote)
  Length  Left Right Bottom  Top Diagonal Y
1  214.8 131.0 131.1    9.0  9.7    141.0 0
2  214.6 129.7 129.7    8.1  9.5    141.7 0
3  214.8 129.7 129.7    8.7  9.6    142.2 0
4  214.8 129.7 129.6    7.5 10.4    142.0 0
5  215.0 129.6 129.7   10.4  7.7    141.8 0
6  215.7 130.8 130.5    9.0 10.1    141.4 0
```

Use whatever graphical techniques you think appropriate to investigate whether there is any pattern or structure in the data. Do you observe something suspicious?

R script

Listing 19: Sample R Script With Highlighting

```
%% You provide code here %%
```

Plots

Place images here with suitable captions.

Discussions

Answer here...

Bonus question (10 pt.)

Here is some data taken from a website that measures where a person clicks on a page:

user	location
1	UL, UL, UL, LR, M
2	UL, LR, M, M
3	LL, UR, M, M

where UL = upper left, LR = lower right, M = middle, LL = lower left, LR = lower right. Find the entropy of location.

R script

Listing 20: Sample R Script With Highlighting

```
%% You provide code here %%
```