CSCI-B 365 Introduction to Data Analysis and Mining
Homework 2
Computer Science
Spring 2018
Indiana University,
Bloomington, IN

Siyi Xian
siyixian@indiana.edu

Friday, Feb 9, 2018 10:00 p.m.

All the work herein is solely mine.

Contents

Problem 1

Textbook exercises, chapter 2, pages: 91-93

1. Exercise 12 **(10 points)**

   (a) Noise is not, but outliers are.

   (b) Noise objects can be outliers, because the original values can be modified to values which are far away from mainly data range.

   (c) Noise objects are not always outliers, because original values can be modified to some values that are similar to correct data.

   (d) Outliers are not always noise. They are mainly be a set of data which is a lot more different than others.

   (e) Yes

2. Exercise 15 **(10 points)**

   For the first scheme, the data get from the set is exactly the same in each group. However, the second scheme, the amount from each group is only roughly the same.

3. Exercise 16 **(10 points)**

   (a) Words that appear in every documents, weight will be too small. Otherwise, the weight will be extremely high.

   (b) To distinguish documents from each others which the original data cannot observe the similarity.

4. Exercise 18 **(15 points)**

   (a)
   ```
   Hamming distance   = number of different bits
                      = 3
   Jaccard Similarity = "1-1" matches /(bits - "0-0" matches)
                      = 2 / 5
                      = 0.4
   ```

   (b) The Hamming distance is similar to the Simple Matching Coefficient, because SMC = Hamming distance / number of bits.
   The Jaccard Similarity is similar to the cosine
   because both of them does not consider "0-0" matches.

(c) Jaccard is better for comparing how similar two organisms of different species. The reason is that the result we want to find is how many genes are in common for different species.

(d) Hamming distance is better to compare genetic makeup of two organisms of the same species, because for the same species, most of their genes are the same. So, it is only necessary to compare the differences.

5. Exercise 19 **(15 points)**

(a) cos(x, y) = 1
corr(x, y): undefined
Euclidean(x, y) = 2

(b) cos(x, y) = 0
corr(x, y) = ?1
Euclidean(x, y) = 2
Jaccard(x, y) = 0

(c) cos(x, y) = 0
corr(x, y)=0
Euclidean(x, y) = 2

(d) cos(x, y) = 0.75
corr(x, y) = 0.25
Jaccard(x, y) = 0.6

(e) cos(x, y) = 0
corr(x, y) = 0

Problem 2

For the following data, give the best taxonomic type (interval, ratio, nominal, ordinal):

1. A section of highway on a map.

   Highway on a map is a nominal data, because different name of highways represent different high ways. For example I-45 is a different highway with I-37, and both of them are unique highways.

2. The value of a stock.

   Stock is a ratio data. Difference can show the profit or loss, and ratio can show changes between weeks, months, or years.

3. The weight of a person.

   Weight can be a ordinal data which can be a sorting standers for people.

4. Marital status.

   Marital status is a nominal data which can distinguish people in to couple of groups.

5. Visiting United Airlines (https://www.united.com) the seating is: Economony, Economy plus, and United Business.

   Seating is a nominal data which can distinguish people in to couple of groups.

   (10 points: each question is worth 2 points)

Problem 3

You are datamining with a column that has physical addresses in some city with the same zipcode. For example,

```
55 WEST CIR
2131 South Creek Road
Apt. #1 Fountain Park
1114 Rosewood Cir
1114 Rosewood Ct.
1114 Rosewood Drive
```

What structure would you create to mine these? What questions do you think you should be able to answer? **(10 points)**

```
Road Name -> Road Type (Ct., Dr, Ln, and etc.) ->  House Number -> Apartment Number
```

Problem 4

The Wisconsin Breast Cancer data set is very famous. Here is the URL `https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`. In the Data Folder are multiple files. Here is the beginning of an R session that allows us to read this data from the web into our local R session:

```
> install.packages("data.table")
> library(data.table)
> install.packages("curl")
> mydata <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases/
                   breast-cancer-wisconsin/breast-cancer-wisconsin.data")
> head(mydata)
        V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11
1: 1000025  5  1  1  1  2  1  3  1   1   2
2: 1002945  5  4  4  5  7 10  3  2   1   2
3: 1015425  3  1  1  1  2  2  3  1   1   2
4: 1016277  6  8  8  1  3  4  3  7   1   2
5: 1017023  4  1  1  3  2  1  3  1   1   2
6: 1017122  8 10 10  8  7 10  9  7   1   4
>
```

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?

```
Samples arrive periodically as Dr. Wolberg reports his clinical cases.
The database therefore reflects this chronological grouping of the data.
The database are using the result of experientment.
```

R Code

Using R, show code that answers the following questions:

1. How many entries are in the data set? 699

Listing 1: Entries

```r
length(mydata$V1)
```

2. How many unknown or missing data are in the data set? 16

Listing 2: Missing Data

```r
sum(mydata == "?")
```

3. How many malignant and benign identifiers are there? *Malignant* : 241 *Benign* : 458

Listing 3: Malignant and Benign

```r
sum(mydata$V11 == 4) # malignant
sum(mydata$V11 == 2) # benign
```

4. Make a histogram of each attribute and discuss the distribution of values *e.g.*, are uniform, skewed, normal. Place images of these histograms into the document. Show the R code that you used below and discussion below that.

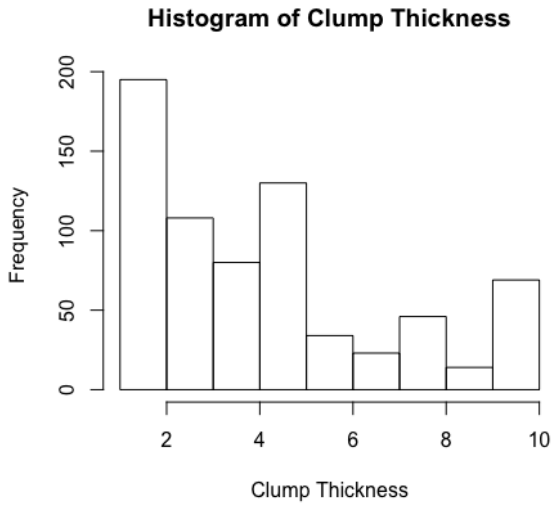
Listing 4: Histogram

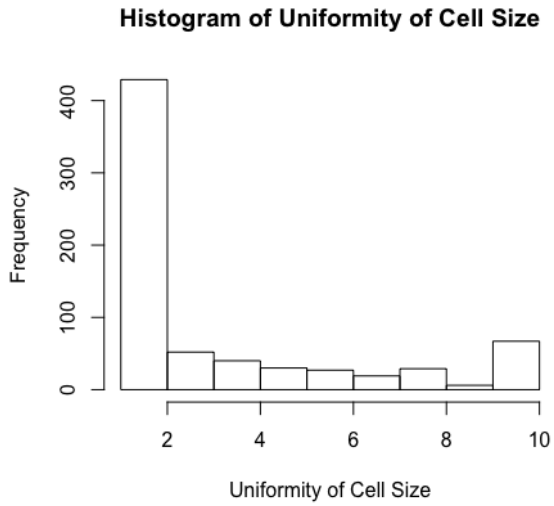
```
meanV7 <- mean(as.numeric(mydata[mydata$V7 != "?",]$V7))
mydata[which(mydata$V7 == "?"),"V7"] <- meanV7

hist(mydata$V2,
     main = "Histogram of Clump Thickness",
     xlab = "Clump Thickness")
hist(mydata$V3,
     main = "Histogram of Uniformity of Cell Size",
     xlab = "Uniformity of Cell Size")
hist(mydata$V4,
     main = "Histogram of Uniformity of Cell Shape",
     xlab = "Uniformity of Cell Shape")
hist(mydata$V5,
     main = "Histogram of Marginal Adhesion",
     xlab = "Marginal Adhesion")
hist(mydata$V6,
     main = "Histogram of Single Epithelial Cell Size",
     xlab = "Single Epithelial Cell Size")
hist(as.numeric(mydata$V7),
     main = "Histogram of Bare Nuclei",
     xlab = "Bare Nuclei")
hist(mydata$V8,
     main = "Histogram of Bland Chromatin",
     xlab = "Bland Chromatin")
hist(mydata$V9,
     main = "Histogram of Normal Nucleoli",
     xlab = "Normal Nucleoli")
hist(mydata$V10,
     main = "Histogram of Mitoses",
     xlab = "Mitoses")
hist(mydata$V11,
     main = "Histogram of Class",
     xlab = "Class")
```

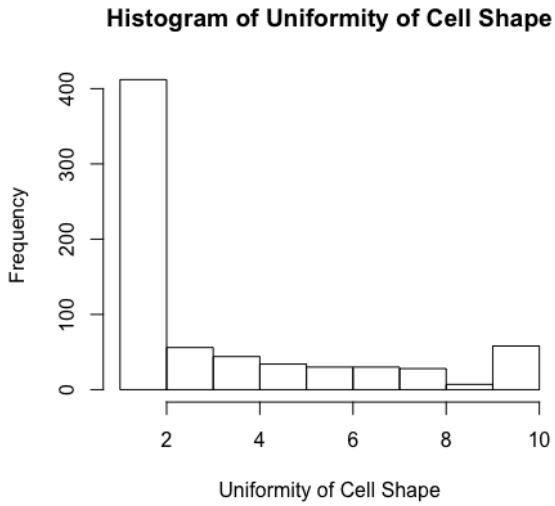
Discussion of Attributes

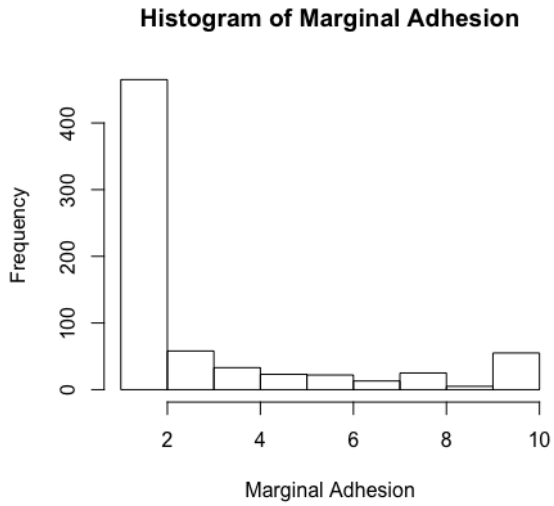
Positive Skewed

Histograms

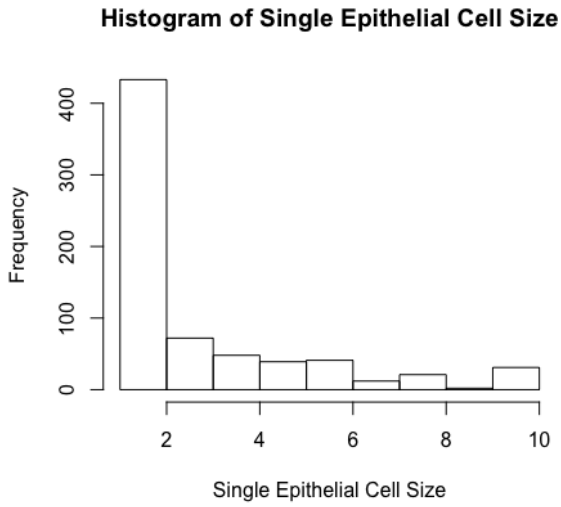


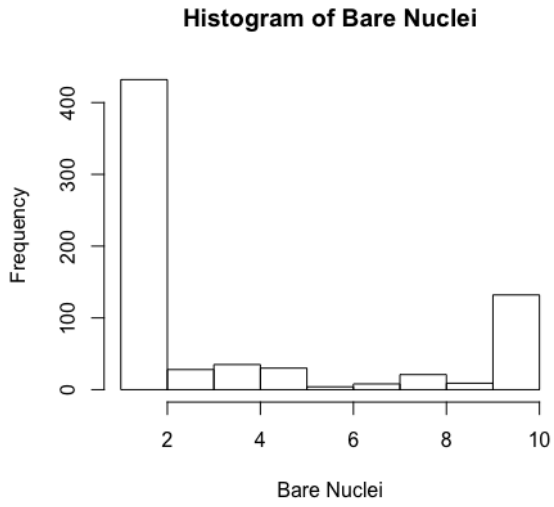


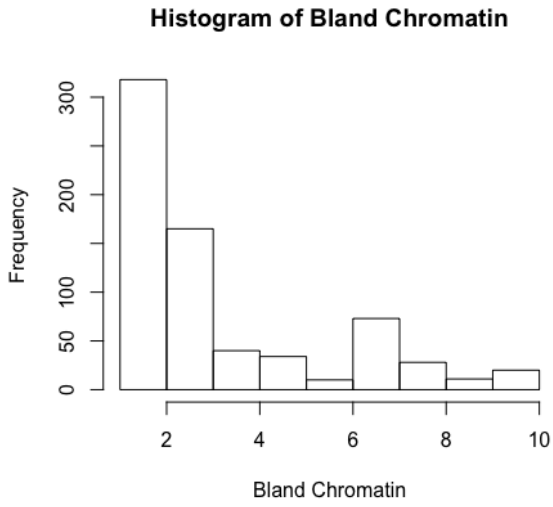


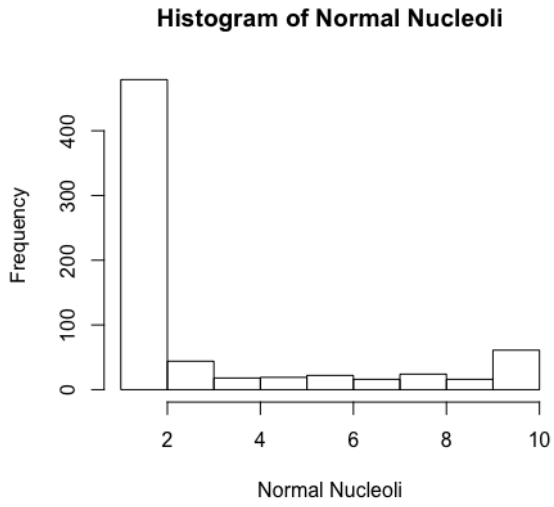


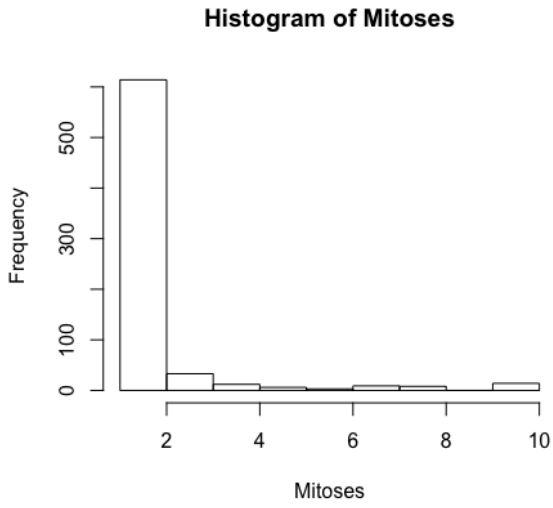
Problem 4 continued on next page...


Histogram of Single Epithelial Cell Size
Frequency
0
100
200
300
400
2
4
6
8
10
Single Epithelial Cell Size


Histogram of Bare Nuclei
Frequency
0
100
200
300
400
2
4
6
8
10
Bare Nuclei


Histogram of Bland Chromatin
Frequency
0
50
100
200
300
2
4
6
8
10
Bland Chromatin


Histogram of Normal Nucleoli
Frequency
0
100
200
300
400
2
4
6
8
10
Normal Nucleoli


Histogram of Mitoses
Frequency
0
100
300
500
2
4
6
8
10
Mitoses


Histogram of Class
Frequency
0
100
200
300
400
2.0
2.5
3.0
3.5
4.0
Class

Discussion of simply removing tuples

Quantify the affect of simply removing the tuples with unknown or missing values. What is the cost in human capital? **(20 points)**

When simply removing those tuples, the other information the contains in the tuple that are not missing are gave up either. So it will waste a lot by doing that.

What to Turn-in

Please follow the syllabus guidelines in turning in your homework. I am providing the LaTeX of this document too. This homework is due Friday, Feb 9, 2018 10:00p.m. **OBSERVE THE TIME**. Absolutely no homework will be accepted after that time. All the work should be your own. Submit a .zip file that includes the files below. Name the .zip as "usename-section number", i.e., hakurban-B365.

1. The *tex and *pdf of the written answers to this document.