Applied Datamining: Homework #2

Due on 9/6/2017

Instructor: Hasan Kurban

Keith Hickman

September 6, 2017

Contents

Problem 1

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

1. A section of highway on a map. - Interval
2. The value of a stock. - Ratio
3. The weight of a person. - Ratio, as we can be infinitely precise about a person's weight.
4. Marital status. - Ordinal. (One would have to be single before married, married before divorced. Thus there is some implicit ordering among the variable.
5. Visiting United Airlines (`https://www.united.com`) the seating is: Economony, Economy plus, and United Business. - Ordinal.

Problem 2

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? A: It depends on the problem set. There are several types of problems that might use address data, including crime statistics, general mapping applications real estate sales, or school districting. I would create different structures for each problem (crime - time and geolocation; school - assigned schools, ratings; real estate sales - home characteristics and other homes sold information).

Problem 3

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? A: The data is used in various applications, including better prediction and diagnostic models for certain types of breast cancer. The data observations are individual patient visits/dianostic tests, and the dimensions are the values held for different types of diagnostic tests. This data has been used in various supervised/unsupervised machine learning tasks. Data are mostly continuous variables, likely ordinal as the higher or lower values could better explain variance or outcomes.

R Code

Using R, show code that answers the following questions:

1. How many entries are in the data set? Answer here . . .

Listing 1: Sample R Script With Highlighting

```
mydata <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases
/breast-cancer-wisconsin/breast-cancer-wisconsin.data")
str(mydata)
```

```
699 observations
```

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

Listing 2: Sample R Script With Highlighting

```
summary(mydata)
sum(is.na(mydata))
```

```
There are no missing data.
```

3. How many malignant and benign identifiers are there? Answere here . . .

Listing 3: Sample R Script With Highlighting

```
table(mydata$V11)

  2   4
458 241
```

```
There are two identifiers of malignancy in the feature labelled V11.
4 = Malignant and 2 = Benign.
There are 241 Malignant cases and 458 benign cases.
```

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.

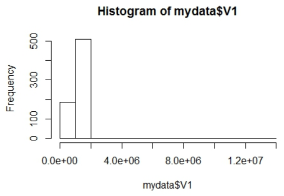
Listing 4: Sample R Script With Highlighting

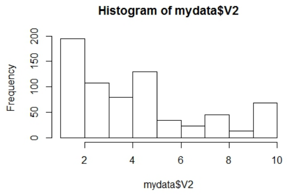
```r
hist(mydata$V1)
hist(mydata$V2)
hist(mydata$V3)
hist(mydata$V4)
hist(mydata$V5)
hist(mydata$V6)
hist(mydata$V8)
hist(mydata$V9)
hist(mydata$V10)
hist(mydata$V11)
```

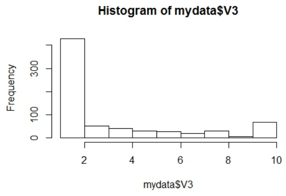
Discussion of Attributes

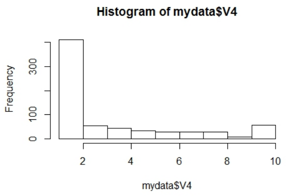
5. Almost all of the data are right-skewed toward lower values of each variable. I would begin standardizing the features, to control for outliers and attempt to normalize the distribution, and then try square root, log, or reciprocal transformations.

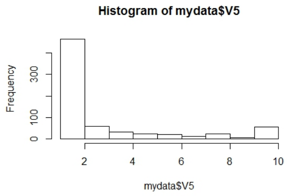
Histograms

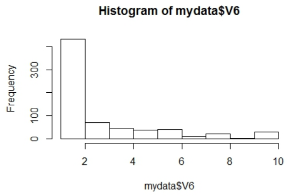


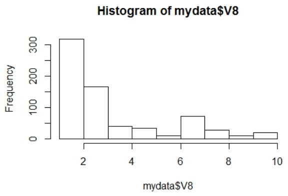


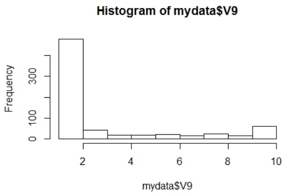










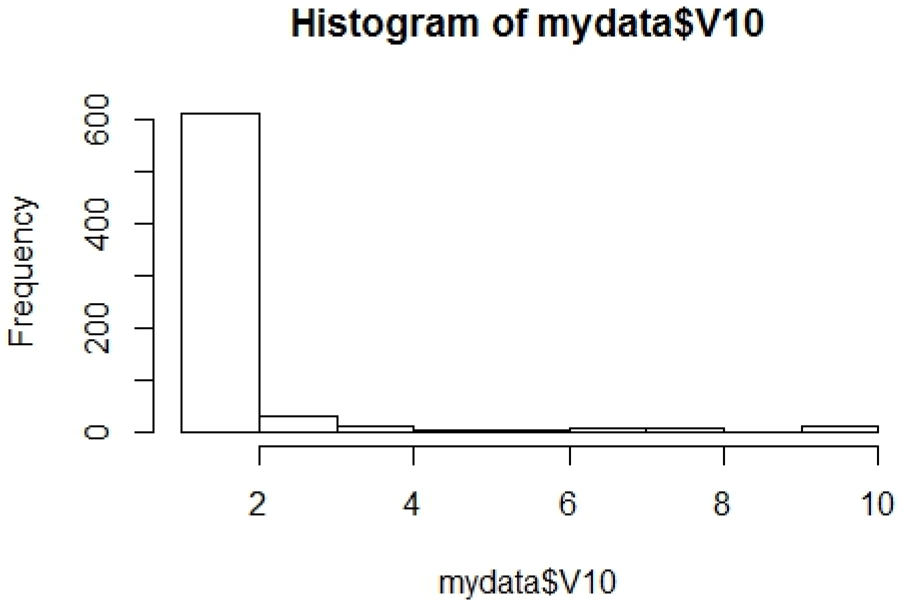


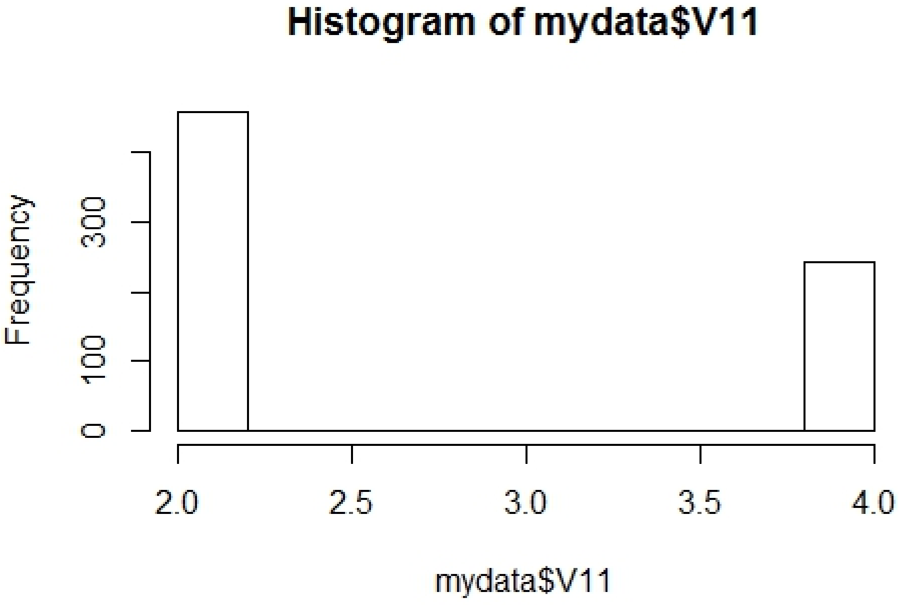
For plots of V10 and V11, see next page.

Discussion of simply removing tuples

Quantify the effect of simply removing the tuples with unknown or missing values. What is the cost in human capital?

In cancer research, removing values with unknown or missing data can be detrimental to the overall outcomes. When creating a model that is designed to diagnose a potentially fatal disease, the models must be extremely accurate. The difference between a 99.5 and 99.99 percent accuracy could cost thousands of dollars in unncessary tests, or worse, a missed positive diagnosis.


Histogram of mydata$V10
Frequency
0
200
400
600
2
4
6
8
10
mydata$V10


Histogram of mydata$V11
Frequency
0
100
300
2.0
2.5
3.0
3.5
4.0
mydata$V11