HW1 - Intro to Machine Learning

Saurabh Madaan

09/14/2012

1 Part 1 - HW01pb1data.csv

1.1 Attributes

The dataset has 800 observations of 5 variables. The first 3 columns are *integers* whereas the 4th and 5th columns are *factors*.

In R, this can be seen using the following commands:

```
setwd(”/Users/Saurabh/Documents/ML_UCSC/week1/HW”)
data<−read.csv(”HW01pb1data.csv”,header=FALSE)
class(data)
str(data)
#800 obs. of 5 variables. V1−3 are int, V4−5 are Factors
```

1.2 Reason for Categorical Variables

Columns 4 and 5 mostly have integers in them, but when one looks in to the *levels*, it is quickly visible that they also have strings “thirty five” and “twenty five”, respectively. Consequently, *R* treats them as factors.

Below are the R commands which provide more details:

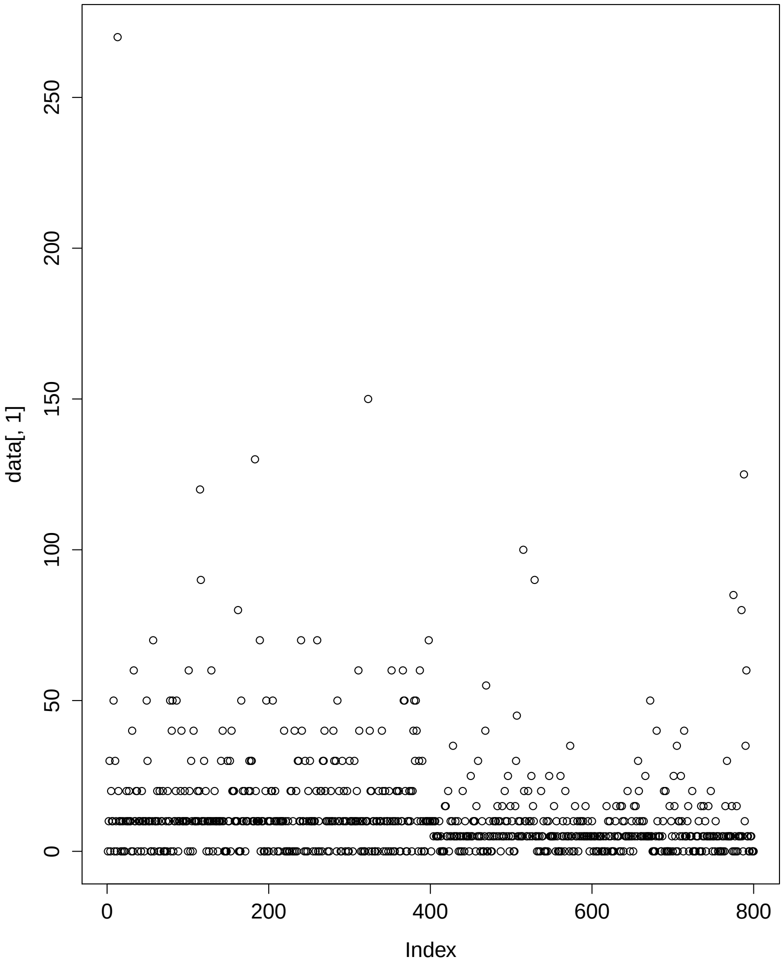
```
l4<−levels(data[,4])
#has integers, and ”thirty five”
which(data[,4]==”thirty␣five”)
#[1] 405

l5<−levels(data[,5])
#has integers, and ”twenty five”
which(data[,5]==”twenty␣five”)
```

```
# [1] 531
```

1.3 Plots for numeric and categorical variables

1. Plot of the 1st column (numerical data)

```
plot(data[,1])
```

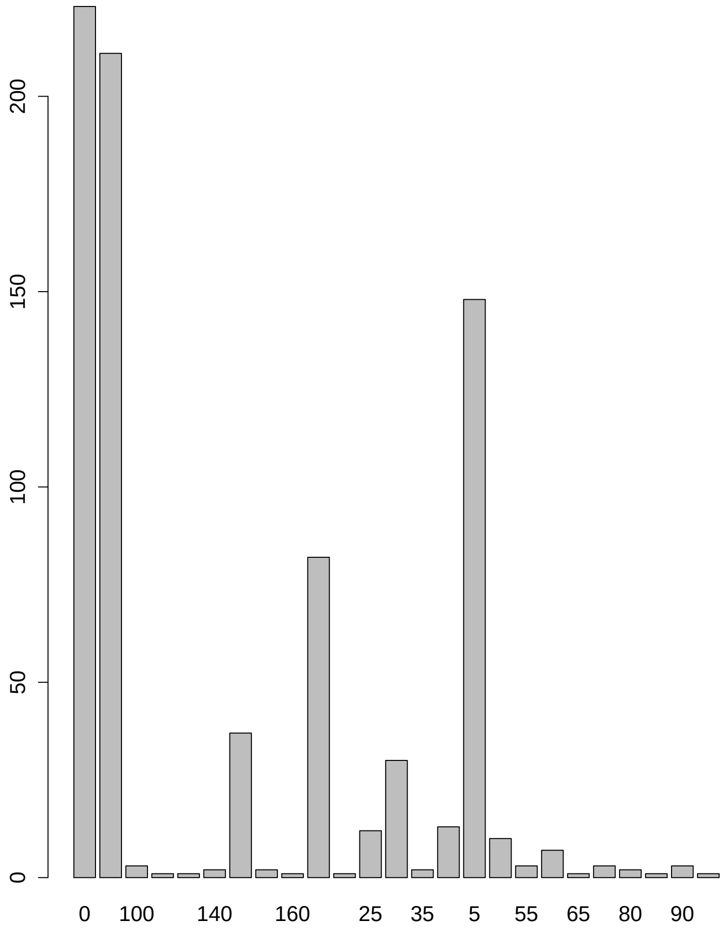


The y-axis of this plot are the values of the first column data, plotted against the index (row number) at which they occur in the dataset. This is a scatter-plot for a numerical data type

.

2. Plot of the 4th column (categorical data)

```
plot(data[,4])
```



This graph is a distribution: it is a histogram with the x-axis showing the values in the 4th column of our dataset, and the y-axis depicts the frequency with which values occur in a specified range. This graph is default since the 4th column data happens to be a categorical variable in our dataset.

2 Part 2 - HW01pb2data.csv

2.1 Extract a random sample of 10k observations

```
setwd(”/Users/Saurabh/Documents/ML_UCSC/week1/HW”)
data<-read.csv(”HW01pb2data.csv”,header=FALSE)
str(data)
# 'data.frame': 2000000 obs. of  1 variable:

nrow(data)
#[1] 2000000

# selecting a sample of 10,000 random records
ss<-seq(1,nrow(data))
rand.ind<-sample(ss,10000,replace=F) #set of random indices for subset
small_data<-data[,1][rand.ind]
length(small_data)
#[1] 10000
```

2.2 Descriptive Stats on Sample

```
#mean, max and other descriptive stats
mean(small_data)
#[1] 9.41002
max(small_data)
#[1] 16.93748
var(small_data)
#[1] 4.004991
quantile(small_data,0.25)
#8.079612
```