KDD Midterm

#1 (10 Points)

Is the following function a proper distance function? Why? Explain your answer.

$$d(\mathbf{x},\mathbf{y}) = (\sum_i | x_i - y_i |)^3$$

Hint: Measure the distance between (0,0), (0,1) and (1,1)

Solution:

Let us assume that,X=(0,0),Y=(0,1) and Z=(1,1)

For any distance function to work the following conditions must be satisfied:

$d(x,y) \geq 0$	1.	Non-negativity or separation axiom
$d(x,y) = 0 \Leftrightarrow x = y$	2.	Identity of indiscernible
$d(x,y) = d(y,x)$	3.	Symmetry
$d(x,z) \leq d(x,y) + d(y,z)$	4.	Subadditivity or triangle inequality

Using given distance function,

The distance between X (0,0) & Y (0,1) => d (x, y)
$= (|0 - 0| + |0 - 1|)^3$
$= (0+ 1)^3$
$= (1)^3$
$= 1$

The distance between Y (0,1) & X (0,0) => d (y, x)
$= (|0 - 0| + |1 - 0|)^3$
$= (0 + 1)^3$
$= (1)_3$
$= 1$

The distance between Y (0,1) & Z (1,1) => d (y, z)
$= (|0 - 1| + |1 - 1|)^3$
$= (1 + 0)^3$
$= (1)^3$
$= 1$

KDD Midterm

The distance between Z (1,1) & Y (0,1) => d (z, y) $= (|1-0| + |1-1|)^3$
$= (1+0)^3$
$= (1)^3$
$= 1$

The distance between Z (1,1) & X (0,0) => d (z, x) $= (|1-0| + |1-0|)^3$
$= (1 + 1)^3$
$= (2)_3$
$= 8$

The distance between X (0,0) & Z (1,1) => d (x, z) $= (|0-1| + |0-1|)^3$
$= (1 + 1)^3$
$= (2)^3$
$= 8$

Checking validity of the distance function properties on the distance values calculated using given distance function.

1. d (x, y) ≥ 0, d (y, x) ≥ 0, d (y, z) ≥ 0, d (z, y) ≥ 0, d (z, x) ≥ 0, d (x, z) ≥ 0.
   Clearly $d(x, y) \geq 0$ and $d(x, y) = 0 \Leftrightarrow x = y$ are satisfied.

2. d (x, y) = d (y, x), d (y, z) = d (z, y), d (z, x) = d (x, z)
   Clearly $d(x, y) = d(y, x)$ is satisfied.

3. d (x, z) = 8, d (x, y) = 1, d (y, z) = 1
   $d(x, z) \leq d(x, y) + d(y, z)$
   8 ≤ 1 + 1
   8 ≤ 2 which is false. So, condition 4 failed d
   (z, x) = 8, d (z, y) = 1, d (y, x) = 1.
   $d(z, x) \leq d(z, y) + d(y, x)$
   8 ≤ 1 + 1
   8 ≤ 2 which is false. So, condition 4 failed here as well.

As per above calculations and observations, given distance function satisfies the first 3 conditions but fails to meet the last condition (Triangle inequality). Therefore, given function is not a proper distance function.

Tejashree Prabhu 10450404

KDD Midterm

 # 2 (15 Points)

There are three major manufacturing companies that make a product: Manufacturers A, B, and C. Manufacturer A has a 60% market share, and Manufacturer B has a 30% market share. 5% of A's products are defective, 6% of B's products are defective, and 8% of C's products are defective.

- **a)** What is the probability that a randomly selected product is defective? P(Defective)?
- **b)** What is the probability that a randomly selected product is defective and that it came from A? P(A and Defective)?
- **c)** What is the probability that a defective product came from B? P(B/Defective)?
- **d)** Are these events (being defective and coming from B) independent? Why?

Solution:

Let's assume there are 1000 items of the product in the market => N = 1000

Based on Market Share,

A has 50% of market share. => N(A) = 60% of 1000 = 600

B has 30% of market share. => N(B) = 30% of 1000 = 300

Remaining are from C => N(c) = 1000-600-300 = 100

Number of defective pieces by manufacturer are as follows:

A's defective products = N (Defective | A) = 5% of 600 items = 30

B's defective products = N (Defective | B) = 6% of 300 items = 18

C's defective products = N (Defective | C) = 8% of 100 items = 8

a) P(Defective) = (N (Defective | A) + N (Defective | B) + N (Defective | C)) / N

= (30 + 18 + 8) / 1000 = 56 / 1000 = 0.056 = 5.6%

KDD Midterm

b) $P(A \cap \text{Defective}) = N(\text{Defective} \mid A) / N = 30 / 1000 = 0.030 = 3\%$

c) $P(B \mid \text{Defective}) = P(\text{Defective} \mid B) / P(\text{Defective}) = 18 / 56 = 0.3214 = 32.14\%$

d) $P(B) = 300 / 1000 = 0.3$

$P(\text{Defective}) = 56 / 1000 = 0.056$

For events to be independent => $P(B \cap \text{Defective}) = P(B) * P(\text{Defective})$

$P(B) * P(\text{Defective}) = 0.3 * 0.056 = 0.0168$

$P(B \cap \text{Defective}) = 18 / 1000 = 0.018$

Since, $P(B \cap \text{Defective}) \neq P(B) * P(\text{Defective})$

Therefore, the events are **not independent** of each other.

KDD Midterm

#3 (20 Points)

The following training dataset (table) is a subset of "census data" for workers with the following characteristics:

- Age: between 15 and 65
- Education: between 0 and 16 years
- Average hours worked per week: between 0 and 80 per week

Age	Education	Gender	Hours worked	Income
39	13	Male	40	<=50K
50	13	Male	13	<=50K
38	9	Male	40	<=50K
53	7	Male	40	<=50K
28	13	Female	40	<=50K
37	14	Female	40	<=50K
49	5	Female	16	<=50K
52	9	Male	45	>50K
31	14	Female	50	>50K
42	13	Male	40	>50K
37	10	Male	80	>50K
30	13	Male	40	>50K
23	13	Female	30	<=50K
32	12	Male	50	<=50K
40	11	Male	40	>50K
34	4	Male	45	<=50K
25	9	Male	35	<=50K
32	9	Male	40	<=50K
38	7	Male	50	<=50K
43	14	Female	45	>50K
40	16	Male	60	>50K
54	9	Female	20	<=50K
35	5	Male	40	<=50K
43	7	Male	40	<=50K
59	9	Female	40	<=50K
56	13	Male	40	>50K
19	9	Male	40	<=50K
54	10	Male	60	>50K
39	9	Male	80	<=50K

Tejashree Prabhu 10450404

KDD Midterm

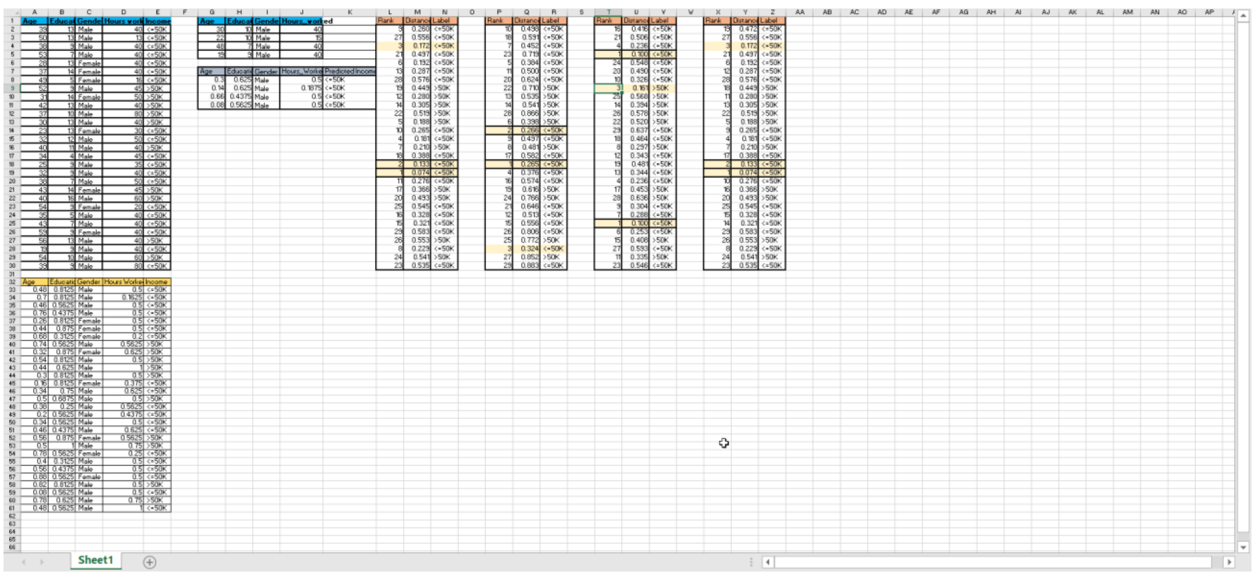
Use **EXCEL**, weighted knn (k=3) and the above training dataset to predict the income level of the following people (test dataset).

Age	Education_Years	Gender	Hours_worked
30	10	Male	40
22	10	Male	15
48	7	Male	40
19	9	Male	40

Remember, you can copy and paste data into Excel.

Solution: file: Q3_KNN.xlsx

Output



KDD Midterm

The following questions (#4 through #6) refer to various subsets of the "Census Data". The original dataset was produced by the "Census Bureau". Use the dataset and methods mentioned below to predict salary level.

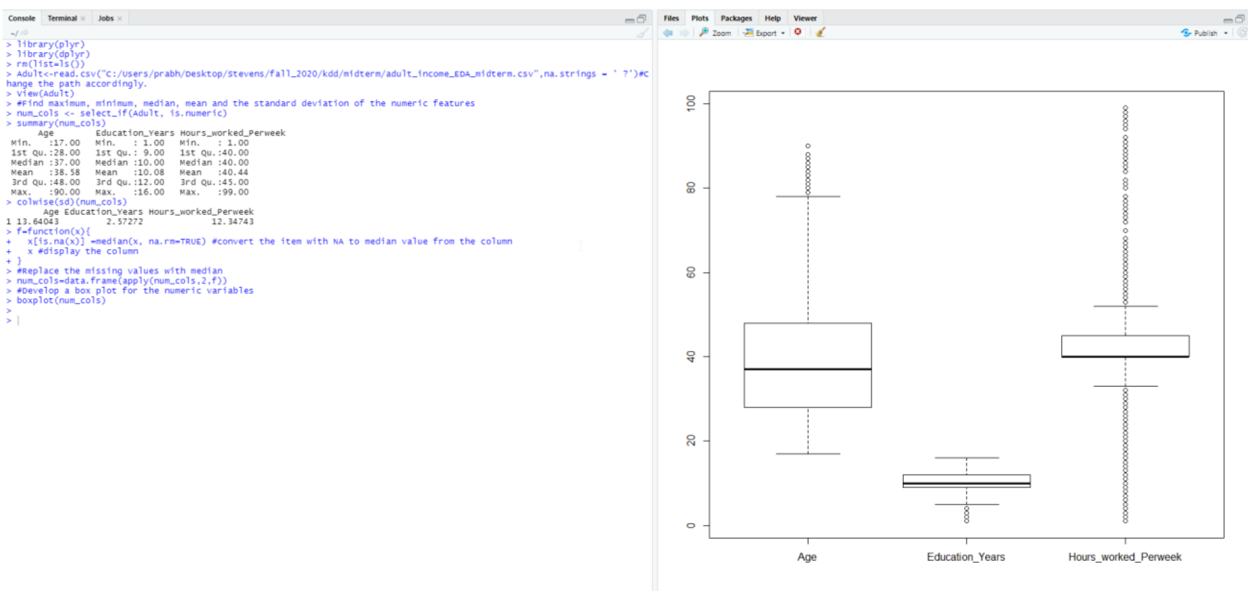
#4 (15 Points)

Load the "Adult_income_EDA.csv" and perform the following exploratory data analysis:

- **Find maximum, minimum, median, mean and the standard deviation of the numeric features**
- **Replace the missing value with the median of the numbers**
- **Develop a box plot for the numeric variables**

Solution: file: Q4_EDA.R

Output



KDD Midterm

#5 (20 Points): **Naïve Bayes:**

Load the "Adult_Income_Bayes.csv"

a) **Remove any row with missing values**
b) **Select every fourth record, starting with the first record, as the test dataset and the remaining records as the training dataset**
c) **Perform Naïve Bayes**
d) **Score the test dataset**
e) **Measure the error rate.**

Solution: file: Q5_NB.R

Output

```
Console   Terminal ×   Jobs ×
~/
> ##############################################
> rm(list=ls())
> library(e1071)
> library(class)
> Adult<-read.csv("C:/Users/prabh/Desktop/Stevens/fall_2020/kdd/midterm/adult_income_Bayes_V2.csv",na.strings =
 ' ?')#Change the path accordingly.
> View(Adult)
> #a)Remove any row with missing values
> omit<-na.omit(Adult)
> View(omit)
> #b)Select every fourth record, starting with the first record, as the test dataset and the remaining records
 as the training dataset
> index<-seq(1,nrow(omit),by=4)
> test<-omit[index,]
> training<-omit[-index,]
> #c)Perform NaÃ¯ve Bayes
> model<-naiveBayes(Income~.,training)
> #Prediction using test
> prediction<-predict(model,test)
> #d)Score the test dataset
> conf_matrix<-table(prediction,test$Income)
> conf_matrix

prediction  <=50K  >50K
     <=50K   4593   481
     >50K    1078  1389
> prop.table(table(prediction,test$Income))

prediction      <=50K       >50K
     <=50K 0.60907042 0.06378464
     >50K  0.14295186 0.18419308
> #e)Measure the error rate
> wrong_prediction<-sum(prediction!=test$Income)
> wrong_prediction
[1] 1559
> wrong_prediction_rate<-wrong_prediction/length(prediction)
> wrong_prediction_rate
[1] 0.2067365
>
> |
```

KDD Midterm

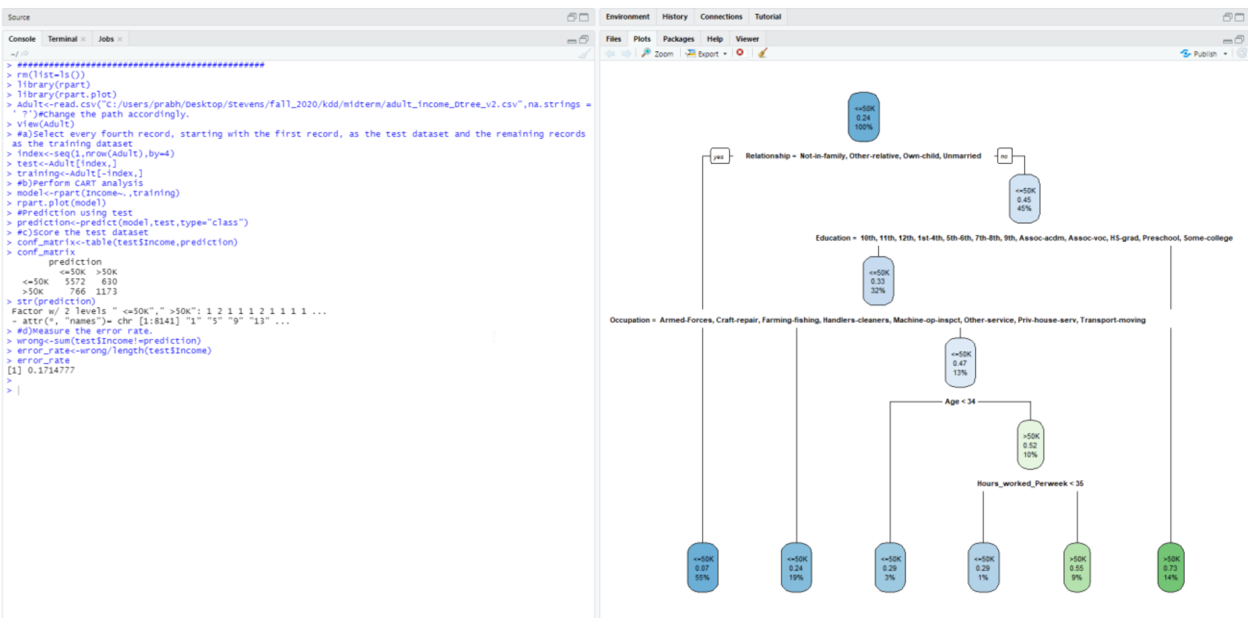
#6 (20 Points): **CART Analysis:**

Load the "Adult_Income_Dtree.csv"

a) Select every fourth record, starting with the first record, as the test dataset and the remaining records as the training dataset
b) Perform CART analysis
c) Score the test dataset
d) Measure the error rate.

Solution: file: Q6_CART.R

Output



Tejashree Prabhu 10450404