**Instructions:**

**You can use Word, Excel, Power Point and R to answer the questions in this test. There are a total of eight (8) multi-part questions, with point values noted for each question.**

**Please show your calculations, or the details of your program(s) for each problem. The R programs should be commented so that each step is clearly explained.**

**Combine all your answers/files into a single zipped file and post the zipped file to CANVAS.**

**#1** (10 Points)

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

****

**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:**

1. Non-negativity or separation axiom



1. Identity of indiscernibles
2. Symmetry
3. 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 |
| 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 are satisfied.

and



1. d (x, y) = d (y, x), d (y, z) = d (z, y), d (z, x) = d (x, z)

Clearly  is satisfied.

1. d (x, z) = 8, d (x, y) = 1, d (y, z) = 1



8 ≤ 1 + 1

8 ≤ 2 which is false. So, condition 4 failed d (z, x) = 8, d (z, y) = 1, d (y, x) = 1.



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.

**# 2** (10 Points)

**There are three major manufacturing companies that make a product: Manufacturers A, B, and C. Manufacture A has a 50% market share, and Manufacture 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.**

1. What is the probability that a randomly selected product is defective? P(Defective)?
2. What is the probability that a randomly selected product is defective and that it came from A? P(A and Defective)?
3. What is the probability that a defective product came from B? P(B/Defective)?
4. 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,

* 1. has 50% of market share. => N(A) = 50% of 1000 = 500
  2. has 30% of market share. => N(B) = 30% of 1000 = 300

Remaining are from C => N(c) = 1000-500-300 = 200

Number of defective pieces by manufacturer are as follows:

A’s defective products = N (Defective | A) = 5% of 500 items = 25

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

C’s defective products = N (Defective | C) = 8% of 200 items = 16

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

= (25 + 18 + 16) / 1000 = 59 / 1000 = 0.059 = 5.9%

1. P (A ∩ Defective) = N (Defective | A) / N = 25 / 1000 = 0.025 = 2.5%
2. P (B | Defective) = P (Defective | B) / P(Defective) = 18 / 59 = 0.3051 = 30.51% **d)** P(B) = 300 / 1000 = 0.3

P(Defective) = 59 / 1000 = 0.059

For events to be independent => P (B ∩ Defective) = P(B) \* P(Defective)

P(B) \* P(Defective) = 0.3 \* 0.059 = 0.0177

P (B ∩ Defective) = 18 / 1000 = 0.018

Since, P (B ∩ Defective) ≠ P(B) \* P(Defective)

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

**#3 (**10 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 |

Use **EXCEL**, weighted knn (k=2) 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.**

**The following questions 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. You can find the raw data in CANVAS under the “Raw\_data” module.**

**#4 (**10 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**

**#5** (15 Points): **Classification 1 Knn**

**Load the “Adult\_income\_knn.csv”**

1. **Remove any row with missing values**
2. **Select every fifth record, starting with the first record, as the test dataset and the remaining records as the training dataset**
3. **Perform K Nearest Neighbor ( K=1,3,5 unweighted)**
4. **Score the test dataset**
5. **Measure the error rate.**

**#6** (15 Points): **Naïve Bayes:**

**Load the “Adult\_Income\_Bayes.csv”**

1. **Remove any row with missing values**
2. **Select every fifth record, starting with the first record, as the test dataset and the remaining records as the training dataset**
3. **Perform Naïve Bayes**
4. **Score the test dataset**
5. **Measure the error rate.**

**#7** (15 Points): **CART analysis:**

**Load the “Adult\_Income\_Dtree.csv”**

1. **Select every fifth record, starting with the first record, as the test dataset and the remaining records as the training dataset**
2. **Perform CART analysis**
3. **Score the test dataset**
4. **Measure the error rate.**

**#8** (15 Points): **C5.0 analysis:**

**Load the “Adult\_Income\_Dtree.csv”**

1. **Select every fifth record, starting with the first record, as the test dataset and the remaining records as the training dataset**
2. **Perform C5.0 analysis**
3. **Score the test dataset**
4. **Measure the error rate.**