**KDD MIDTERM**

**Neil Gupte**

**10445674**

**#1** (10 Points)

**Is the following function a proper distance function? Why? Explain your answer. Measure the distance between (0, 0, 0), (0, 1, 0), (0, 1, 1), and (1, 1, 1)**

**Answer 1.**

The above function **IS** a valid distance function as it satisfies the following properties :

* Property 1: Distance is always non-negative
* Property 2: Commutative, distance from “A to B” is distance from “B to A”
* Property 3: Triangle inequality holds, distance from “A to C” must be less than or equal to distance from “A to B and B to C”

Let us take 4 points as follows – A(0,0,0), B(0,1,0), C(0,1,1) and D(1,1,1)

Distance of AB = (|0 - 0|2+|0 – 1|2+|0 – 0|2) = 1

Distance of BA = (|0 - 0|2+|1-0|2+|0 – 0|2)= 1

Distance of BC = (|0 - 0|2+|1 – 1|2+|0 – 1|2)= 1

Distance of CB = (|0 - 0|2+|1 – 1|2+|1 – 0|2)= 1

Distance of AC = (|0 - 0|2+|0 – 1|2+|0 – 1|2)=2

Distance of CA = (|0 - 0|2+|1 – 0|2+|1 – 0|2)= 2

Distance of CD = (|0 - 1|2+|1 – 1|2+|1 – 1|2) = 1

Distance of DC = (|1 - 0|2+|1 – 1|2+|1 – 1|2) = 1

Distance of AD = (|0 - 1|2+|0 – 1|2+|0 – 1|2) = 3

Distance of DA = (|1 - 0|2+|1 – 0|2+|1 – 0|2) = 3

Distance of BD = (|0 - 1|2+|1 – 1|2+|0 – 1|2) = 2

Distance of DB = (|1 - 0|2+|1 – 1|2+|1 – 0|2) = 2

Thus all the properties are satisfied.

**#2** (10 Points)

**Load the “COVID19\_v3.CSV” dataset, from the raw\_data module in CANVAS, into R. This is a fictional COVID19 healthcare workers data set. Perform the EDA analysis by:**

**(See the data dictionary at the last page of this exam).**

1. **Summarizing each column (e.g. min, max, mean )**
2. **Identifying missing values**
3. **Displaying the frequency table of “Infected” vs. “MaritalStatus”**
4. **Displaying the scatter plot of “Age”, “MaritalStatus” and “MonthAtHospital”, one pair at a time**
5. **Show box plots for columns: “Age”, “MaritalStatus” and “MonthAtHospital”**
6. **Replacing the missing values of “Cases” with the “mean” of “Cases”.**

**Ans 2. Submitted the code**

**Use EXCEL and the “COVID19\_A.CSV.xlxs” (Excel file containing another variation of the fictional COVID19 dataset) to solve the following two problems.**

**#3** (10 Points)

**Use unweighted Knn (k=3) to classify the following three records (test dataset)**

**Use only excel for this problem.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Exposure** | **MartialStatus** | **MonthAtHospital** | **Infected** |
| 1 | Married | 1 | Yes |
| 3 | Single | 4 | No |
| 2 | Single | 6 | Yes |

**#4** (15 Points)

**Discretize the “MonthAtHospital” into “less than 6 months” and “6 or more months”. Construct a classification and regression tree (CART) to classify infection (“infected’) based on the other variables (only one split level). Use only excel for this problem. Do not use the original MonthAtHospital a predictor.**

**#5** (10 Points)

**Load the CANVAS “COVID19\_v3.CSV” dataset into R/Python. Remove the missing values. Discretize the “MonthAtHospital” into “less than 6 months” and “6 or more months”. Also discretize the age into “less than 35”, “35 to 50” and “51 or over”. Construct a Naïve Bayes model to classify infection (“infected’) based on the other variables. Measure the accuracy of the model.**

**Do not use the original MonthAtHospital and age variables as predictors.**

**#6** (10 Points)

**Load the CANVAS “COVID19\_v3.CSV” dataset into R/Python. Remove the missing values. Discretize the “MonthAtHospital” into “less than 6 months” and “6 or more months”. Also discretize the age into “less than 35”, “35 to 50” and “51 and over”. Construct a CART model to classify infection (“infected’) based on the other variables. Measure the accuracy of the model.**

**Do not use the original MonthAtHospital variable as a predictor.**

**#7 (**10 Points)

**Load the CANVAS fictional “COVID19\_v3.CSV” dataset into R/Python. Remove the missing values. Use unweighted knn(k=5) to predict infection rate (infected) for a random sample (30%) of the data (test dataset).**

**# 8**(10 Points)

**The following table shows the population and the actual current prevalence rate of COVID19 in the US, Italy and Spain.**

**Considering only the three countries (US, Italy and Spain) use the table to answer the following questions:**

1. **Estimate the number of cases in the US, Italy and Spain.**
2. **Given that a person is living in the US, what is the probability that the person is infected with COVID19.**
3. **Given that a person is diagnosed with the COVID19, what is the probability that the person lives in the US.**

|  |  |  |
| --- | --- | --- |
|  | **Population** | **Prevalence** |
|  | rounded to nearest | Cases |
|  | **Million** | **Per Million** |
| **US** | 331 | 381.24 |
| **Italy** | 60 | 1463.97 |
| **Spain** | 47 | 1590.24 |
|  |  |  |
|  |  |  |

**Ans 8.**

**a)Total Cases= ((331\*381.24)+(60\*1463.97)+(47\*1590.24))**

**=288769.92**

**b)P(person living in us is infected)=Total Cases infected in Usa/Total population of USA**

**=126190.44/331000000**

**=0.00038123**

**c)p(person infected lives in Usa)= cases in usa /total cases in all countries**

**= 126190.44/288769.92**

**=0.43699302**

**#9 (**15 Points)

**a) Company XYZ is targeting professionals between the ages of 20 and 50 years old with an asset size of 50k to 100K. To estimate the missing income fields, the company is using k-nearest neighbors. (Use Excel for this problem) What would be the value of income for customer x in the table below if:**

**K = 1 and method = ”unweighted vote” is used**

**K = 2 and method = ”unweighted vote” is used**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | Age | **Asset Size** | **Income** |
| **X** | 30 | 60 | ? |
| **1** | 25 | 50 | 100K |
| **2** | 33 | 60 | 90K |
| **3** | 35 | 80 | 150K |

**b) The company has decided to classify income by category instead of estimating a number. Furthermore, it has obtained additional customer information with the exact profile of customer X.**

* **What would be the income category for X if K=3 and “distance weighted vote” is used? Why?**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | Age | **Asset Size** | **Income** |
| **X** | 30 | 60 | ? |
| **1** | 25 | 50 | Medium |
| **2** | 33 | 60 | Low |
| **3** | 35 | 80 | High |
| **4** | 30 | 60 | Medium |
| **5** | 30 | 60 | High |
| **6** | 30 | 60 | High |