ALI ABDULLAH AHMAD
CWID-20031246
CS513-A

#1 (10 Points)
Is the following function a proper distance function? Why? Explain your answer. Measure the distance from (0, 0, 0) to (0, 1, 0) and from (0, 1, 0) to (0, 0, 0)

ANS

This is not a proper distance function because it can give us a negative distance which is not possible. The function involves summing cube differences, which can be negative. The cube root of a negative number is also negative. Since a distance function must be non-negative.

Computing the Given Distances

a. distance from (0, 0, 0) to (0, 1, 0)
   $= ((0-0)^3+(0-1)^3+(0-0)^3)^{1/3}$
   $= (0+(-1)+0)^{1/3}=(-1)^{1/3}$
   $= -1$
   This also proves that the distance function is not correct as distance cannot be negative

b. from (0, 1, 0) to (0, 0,0)
   $((0-0)^3+(1-0)^3+(0-0)^3)^{1/3}$
   $= (0+1+0)^{1/3}$
   $= 1$

 # 2 (10 Points)
An employee of a company is traveling to either England, Italy, or Spain. The employee can travel to only one country. There is a 50% chance the employee will go to England and a 20% chance to Italy.
Assume the chances of contracting COVID to be proportional to the prevalence of the disease in each country, given in the table below. For example, the chances of contracting COVID in England is 1,200/1,000,000 What are the chances that the employee will contract COVID while travelling?
Assume that the employee has traveled to Europe and contracted COVID, what is the probability that he/she traveled to England?

ANS

- The probability of traveling to each country:
  P(E)=0.50 (England)

P(I)=0.20 (Italy)
P(S)=1−(P(E)+P(I))=0.30 (Spain)

- The probability of contracting COVID in each country (prevalence per million):
  P(C|E)= 1200 / 1000000 = 0.0012
  P(C|I)= 1500 / 1000000 = 0.0015
  P(C|S)= 1600 / 1000000 = 0.0016

Compute Total Probability of Contracting COVID

Using the law of total probability:

$$P(C)=P(C|E)P(E)+P(C|I)P(I)+P(C|S)P(S)$$

Substituting the values:

$$P(C)=(0.0012\times0.50)+(0.0015\times0.20)+(0.0016\times0.30)$$

$$=0.0006+0.0003+0.00048$$

$$\mathbf{=0.00138}$$

the probability that the employee will contract COVID while traveling is 0.138%

Probability of Traveling to England Given Infection

We use Bayes' Theorem: $P(E|C)=P(C|E)P(E) / P(C)$

Substituting the values: $P(E|C)= (0.0012\times0.50) / 0.00138$

$= 0.0006 / 0.00138 \approx 0.4348$

Thus, given that the employee has contracted COVID, the probability that they travelled to England is 43.48%.

#3 (10 Points)
Load the "hepatitis_A.csv" dataset, from the Raw_data module in CANVAS, into Python (see the data dictionary at the bottom of this document). This is a dataset used for predicting "patient mortality".
Perform the EDA analysis by:
I. Summarizing each numerical column (e.g., min, max, mean)
II. Displaying scatter plots of "BILIRUBIN"," SGOT" and "ALBUMIN" one pair at a time
III. Showing box plots for columns "BILIRUBIN"," SGOT" and "ALBUMIN"

ANS - in Q3_and_Q4_ipynotebook

#4 (15 Points)
Load the "hepatitis_A.csv" dataset, from the Raw_data module in CANVAS, into Python/R (Excel file containing another variation of the hepatitis dataset). This is a dataset used for predicting "patient mortality". Construct a CART model to classify "patient mortality" based only on the "SEX","Age_Quartile", "STEROID", "FATIGUE" and "MALAISE" attributes

ANS – in Q3_and_Q4_ipynotebook

#5 (15 Points)
Load the "hepatitis_B2.csv" dataset, from the Raw_data module in CANVAS, into Python. This is a variation of the hepatitis dataset used for predicting "Patient mortality". Construct a knn model to classify "patient death" based on only the AGE, SEX, ASCITES, BILIRUBIN, ALK_PHOSPHATE, SGOT, ALBUMIN attributes (K=1,3,5,7)

ANS- in Q5_KNN.ipynb

#6 (20 Points)
Use Excel and the k-Nearest Neighbors (k-NN) algorithm with k=1,2,3 to classify the first five rows of hepatitis_C2.csv as test data. Use the remaining records as training data. Evaluate the model's performance by measuring, accuracy, precision, recall, and F1-score.

ANS – in Ali_Abdullah_Ahmad_Q6_KNN.xlsx

#7 (20 Points)
Use Excel and the training data in hepatitis_D2.csv to construct the first-level split of a CART (Classification and Regression Tree) classification model.

ANS – in Ali_Abdullah_Ahmad_Q7_CART.xlsx