

**CS 513-A: Knowledge Discovery & Data Mining**

*Schaefer School of Engineering & Science at Stevens Institute of Technology*

Spring 2018

**Mid Term Examination**

**Name: Balaji Katakam**

**Date: 04/07/2018**

**Q1. 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:**

According to me, the formula listed above is not a proper distance function.

Let’s take an example from the hint provided in the question.

Let A= (0,0)

Let B= (0,1)

Let C= (1,1).

The distance between point A & point C be AC, can be calculated as follows:

d(AC) = (|0-1|+|0-1|) ^3

d(AC) = 2^3 = 8.

The distance between point A & point B be AB, can be calculated as follows:

d(AB) = (|0-0|+|0-1|) ^3

d(AB) = 1^3 = 1.

The distance between point B & point C be BC, can be calculated as follows:

d(BC) = (|0-1|+|1-1|) ^3

d(BC) = 1^3 = 1.

According to Pythagoras theorem:

d(AB)^2 + d(BC)^2 = d(AC)

Therefore, on substituting the values we get 1 + 1 = 2! = 8

Since, the values do not match the given criteria, the distance function is not correct.

**Q2 – R program**

**Q3 – R program**

**Q4 – Excel Sheet**

**Q5. There are three major manufacturing companies that make a product: manufactures A, B, and C. Manufacturer A has a 50% 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.**

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:**

Since Manufacturer A has 50% market Share, Manufacturer B has 30% market Share thus Manufacturer C has 20% market share.

Considering a random sample of size 1000.

Manufacturer A’s products in the market = N(A) = 50% of 1000 items = 500 items

Manufacturer B’s share of product in the market = N(B) = 30% of 1000 items = 300 items

Manufacturer C’s share of product in the market = N(C) = 20% of 1000 items = 200 items

P(Defective/A) = So, A has 500 pieces: 5% of A’s products are defective = 25 defective products.

P(Defective/B) = B has 300 pieces: 6% of B’s products are defective = 18 defective products.

P(Defective/C) = C has 200 pieces: 8% of C’s products are defective = 16 defective products.

5a.

P(Defective) = ((N[Defective/A]) + (N[Defective/B]) + (N[Defective/C]))/N

P(Defective) = (25 + 18 + 16) / 1000

P(Defective) = 59 / 1000

P(Defective) =0.059 = 5.9%

5b.

probability that a randomly selected product is defective and that it came from A

P (A ∩ Defective) = N (Defective | A) / N

P (A ∩ Defective) = 25/1000

P (A ∩ Defective) = 0.025

P (A ∩ Defective) = 2.5%

5c.

The probability that a defective product came from B = P(B)\*P(B[Defective])

P (B | Defective) = P (Defective | B) / P(Defective)

P (B | Defective) = 18 / 59

P (B | Defective) = 0.3051

P (B | Defective) = 30.51%

5d.

Probability that the product is from manufacturer B = P(B) = 300 / 1000 = 0.3

Probability that the product from manufacturer is defective B = 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)

Thus, these events are not independent of each other.

Q6

(10 Points)

True or False:

1. In data mining, we usually delete all the data rows that contain a

missing value to obtain a clean dataset. - False

2. Supervised data mining methods are those that use expert opinions. - False

3. Usually, low-complexity classifiers/separators need not change very

much to accommodate new data points. - True

4. The optimal level of model complexity is obtained at the minimum

error rate on the training dataset. - False

5. Data mining processes are autonomous, requiring little or no human

oversight. - False