**#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:** To check the given function is a distance function or not. There are property of distance which any distance should satisfy for a proper distance.

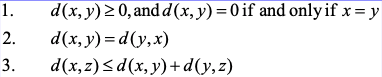
Now calculating the distance between (0,0), (0,1) and (1,1)

(0,0) (0,1) = = 1

(0,0) (1,1) = =

(0,1) (1,1) = = 1

Using the distance properties, check that the distance function satisfies all the property:



**I Property**: First property is satisfied because all the three distances are greater than 0.

**II Property**: Now calculating distance from y to x

(1,0) (0,0) = =1

(1,1) (0,0) = = 4

(1,0) (1,1) = = 1

Here the distance calculated  **=**

Satisfy the II property

**Third property:**

**≤ +**

– (0,0)(0,1) = 1

– (0,1)(1,1) = 1

– (0,0)(1,1) = 4

≤ +

**Third Property doesn’t satisfy.**

**So is not a distance function.**

**# 2** (15 Points)

**A large department store sells sport shirts in three sizes (Small, Medium and Large), three patterns (plaid/Pl, print/Pr, and stripe/Sr), and two sleeve lengths (long and short). The accompanying tables give the proportions of shirts sold falling in the various category combinations.**

* **What is the probability that the next shirt sold is a medium long-sleeved, print shirt? Why?**

**Solution**: P(next shirt sold is a medium long-sleeved, print shirt) = 0.05

It is a Joint probability, considering the **long sleeves** table and checking row M and column Pr.

* **What is the probability that the next shirt sold is a medium print shirt? Why?**

**Solution**: P(next shirt sold is a medium print shirt) = 0.07 + 0.05

= 0.12

Addition of two joint probability one value is from Short sleeves table and other one is from Long sleeves table. Checking row M and column Pr.

* **What is the probability that the next shirt sold is a short-sleeved shirt? A long-sleeved shirt? Why?**

**Solution:** P(next shirt sold is a **short sleeve** shirt) = 0.56

P(next shirt sold is a **long sleeve** shirt) = 0.44

The above values are calculated as the addition of all the values in Short sleeve table and Long sleeved table respectively.

* **Given that the shirt just sold was a short sleeved, plaid, what is the probability that its size was medium?**

**Solution:**

P(shirt sold was short sleeve and plaid) = .15

∴ Considering **Short Sleeve** table, addition of the plaid column values.

P(size was medium) = (Conditional Probability)

= 0.533

* **Given that the shirt just sold was medium, plaid, what is the probability that it was short sleeved? Long-sleeved?**

Solution:

Given: P(shirt sold was medium, plaid) = 0.08+0.1

= 0.18

P(Shirt was **short sleeved**) = (Conditional Probability)

= 0.444

P(Shirt was **long sleeved**) = (Conditional Probability)

= 0.556

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Short Sleeves | |  |  |  |
|  | Pl | Pr | Sr |  |
| S | 0.04 | 0.02 | 0.05 | 0.12 |
| M | 0.08 | 0.07 | 0.12 | 0.27 |
| L | 0.03 | 0.07 | 0.08 | 0.18 |
|  | 0.15 | 0.16 | 0.25 | 0.56 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Long sleeves | |  |  |  |
|  | Pl | Pr | Sr |  |
| S | 0.03 | 0.02 | 0.03 | 0.08 |
| M | 0.1 | 0.05 | 0.07 | 0.22 |
| L | 0.04 | 0.02 | 0.08 | 0.14 |
|  | 0.17 | 0.09 | 0.18 | 0.44 |

**#3 (**15 Points)

1. **Company XYZ is targeting professionals between the ages of 25 to 45 years old with an asset size of 50 to 100K. To estimate the missing income fields, the company is using k-nearest neighbors.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | Age | **Agemmn** | **Asset Size** | **Asset Sizemmn** | **Income** |
| **X** | 30 | 0.25 | 60 | 0.2 | ? |
| **1** | 25 | 0 | 50 | 0 | 100K |
| **2** | 33 | 0.4 | 60 | 0.2 | 90K |
| **3** | 35 | 0.5 | 80 | 0.6 | 150K |

**Normalized Data**

* **What would be the value of income for customer x in the table below if:**

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

**Solution:** ID X is closest to ID 2 and therefore it X salary can be classified as **90K.**

**K =3 and method = ”distance weighted vote” is used?**

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | Distance | Age | **Asset Size** | **Income** |
| **X** |  | 0.5 | 0.333 | **?** |
| **1** |  | 0 | 0 | 100K |
| **2** |  | 0.8 | 0.333 | 90K |
| **3** |  | 1 | 1 | 150K |

Ynew = =

=

=

=96.28

The estimated salary for the professional with ID X will be **100K** because ID 1 is the closest. The salary value of 96.28 makes a significant contribution to the estimated salary value.

**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 |

**Solution:** Customers having id 4,5,6 has asset and age column similar values. If we calculate the Euclidian distance between them, it gives the result 0. Inverse of the distance becomes infinite.

So, we cannot predict the value for the salary of X.

So, income of X will be **UNDEFINED**.