Jingxuan Ai 10431517

**#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)**

We can see that this function fulfills:

1. d >= 0, d(x,y) = 0 only when x = y.
2. d(x,y) = d(y,x)
3. For an triangle, two side’s sum must be larger than the third side.

Then we apply rule 3 to this function:

d((0,0)&(0,1)) = 1, d((0,0)&(1,1)) = 4, d((0,1)&(1,1)) = 1

1 + 1 < 4, which is against rule 3. Thus this is not a proper 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?**
* **What is the probability that the next shirt sold is a medium print shirt? Why?**
* **What is the probability that the next shirt sold is a short sleeved shirt? A long-sleeved shirt? Why?**
* **Given that the shirt just sold was a short sleeved, plaid, what is the probability that its size was medium?**
* **Given that the shirt just sold was medium, plaid, what is the probability that it was short sleeved? Long-sleeved?**

|  |  |  |  |
| --- | --- | --- | --- |
| **Short Sleeves** | |  |  |
|  | Pl | Pr | Sr |
| S | 0.04 | 0.02 | 0.05 |
| M | 0.08 | 0.07 | 0.12 |
| L | 0.03 | 0.07 | 0.08 |
|  |  |  |  |
| **Long Sleeves** |  |  |  |
|  | Pl | Pr | Sr |
| S | 0.03 | 0.02 | 0.03 |
| M | 0.1 | 0.05 | 0.07 |
| L | 0.04 | 0.02 | 0.08 |

1. Long-sleeved => Medium => Print Shirt (Long Sleeves => M => Pr). It’s 0.05
2. (Long-Sleeves, M, Pr) + (Short-Sleeves, M, Pr) = 0.05 + 0.07 = 0.12
3. 0.56 for adding Short Sleeves table together. 0.44 for adding Long Sleeves table together.
4. 0.08 / (0.04 + 0.08 + 0.03) = 0.53. (M, Pl) / (SML, Pl).
5. P(ShortSleeves) = 0.08/(0.08+0.1) = 0.44. P(LongSleeves) = 0.1 / (0.08+0.1) = 0.56.

**#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.**

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

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

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

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

After normalization:

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Age’** | **Asset Size’** | **Income** |
| **X** | 0.25 | 0.2 | ? |
| **1** | 0 | 0 | 100K |
| **2** | 0.4 | 0.2 | 90K |
| **3** | 0.5 | 0.6 | 150K |

When k = 2 and method is “unweighted vote”:

d(x,1) =

d(x,2) =

d(x,3) =

(x,2) and (x,1) are two nearest points. Income of x is 90k or 100k with equal possibility 50%, thus, Income is 95k

When k = 3 and method is “distance weighted vote”:

Vote(1) = 1 / d(x,1)^2 = 9.76

Vote(2) = 1 / d(x,2)^2 = 44.44

Vote(3) = 1 / d(x,3)^2 = 4.49

Income = (9.76\*100+44.44\*90+4.49\*150) / (9.76+44.44+4.49) = 96.25k

Three nearest points are (x,1), (x,2) and (x,3). The Income of x is 96.25k

**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 |

d(x,1) =

d(x,2) =

d(x,3) =

d(x,4) =

d(x,5) =

d(x,6) =

Vote(1) = 1 / d(x,1)^2 = 0.008

Vote(2) = 1 / d(x,2)^2 = 0.1111

Vote(3) = 1 / d(x,3)^2 = 0.00235

Vote(4) = 1/ d(x,4)^2 = +

Vote(5) = 1/ d(x,4)^2 = +

Vote(6) = 1/ d(x,4)^2 = +

(x,4), (x,5) and (x,6) are three nearest points. Income of X is high.

**#4 (**10 Points)

* **Use R to create a vector of the following 20 numbers**
* **Find maximum, minimum, median, mean and the standard deviation of the follow 20 numbers.**
* **Replace the missing value with the mean of the numbers**
* **Use R to develop a box plot for these numbers**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **45** | **48** | **6** | **42** | **49** | **63** | **81** | **56** | **21** | **75** |
| **25** | **48** | **56** | **24** | **73** | **82** | **NA** | **80** | **86** | **88** |

**The following two questions refer to the “IBM\_attrition\_v1.csv” dataset on canvas which is a subset of the “IBM attrition” dataset. The original dataset is used in IBM ML labs to uncover the factors that lead to employee attrition (attrition=yes). The dataset is a fictional data set created by IBM data scientists.**

**In R ==================**

**#5** (30 Points): **Classification using K Nearest Neighbor:**

**Load IBM\_attrition\_v1.csv into R**

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

**In R ==================**

**#6** (20 Points): **Naïve Bayes:**

**Load IBM\_attrition\_v1.csv into R**

1. **Remove any row with missing values**
2. **Select every third record as the test dataset and the remaining records as the training dataset**
3. **Preform Naïve Bayes using only the following columns: “JobSatisfaction”, “Single” and “Gender”**
4. **Score the test dataset**
5. **Measure the error rate.**

**In R ==================**