**Course: CS- 513**

**Instructions:**

**You can use Word, Excel, Power Point and R to answer the questions in this test. There are a total of five (5) multi-part questions, with point values noted for each question.**

**Please show your calculations, or the details of your program(s) for each problem. The R programs should be commented so that each step is clearly explained.**

**Combine all your answers/files into a single zipped file and post the zipped file to “Midterm” in CANVAS.**

**#1 (**10 points**)**

**For the experiment consisting of a single die toss, let**

**A = {outcome is <= 3}**

**B = {1,2,5,6}**

**C = {outcome is odd}**

**Please answer each of the following three True/False questions. Show your work.**

**a).  True or False?**

**b).  True or False?  
c). **  **True or False?**

**#2** (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)**

**#3 (**20 points**)**

**Using R perform the following**

1. **Load the following CSV file to your R environment:** [**http://www.math.smith.edu/sasr/datasets/help.csv**](http://www.math.smith.edu/sasr/datasets/help.csv)
2. **Create a dataframe of: id, age, “number of days any substance used” (daysanysub), substance, and race group**
3. **Normalize “number of days any substance used” (daysanysub)**
4. **Substitute the missing values of “daysanysub” with zero**
5. **Calculate: Mean, Max, Median, STD of Age**
6. **Create a categorical variable “age\_group” as:** 
   * 1. **From 0 up to and including 30 =”Young”**
     2. **Over 30 up to and including 60 =”Middle Age”**
     3. **Older than 60 =”OLD”**
7. **Create “training” and “test” datasets by:**

**Choosing every *third* record as “test” and the remaining records as “training”.**

**#4** (20 Points)

**A telecommunications company is concerned about the number of customers leaving their business (Chur=True). Using past data, an analyst has prepared the table below. Using the table below, calculate the following probabilities:**

**P(Churn=True)**

**P(Churn=False)**

**P(International Plan=Yes)**

**P(Voice Plan=Yes)**

**P(International Plan=Yes, Voice Plan=Yes)**

**Are “Voice Plan” and “International plan” independent?**

**P( (International Plan=Yes, Voice Plan=Yes)/Churn=True)**

**P( (International Plan=Yes, Voice Plan=Yes)/Churn=False)**

**P( Churn=False/(International Plan=Yes, Voice Plan=Yes ))**

**P( Churn=False/(International Plan=Yes, Voice Plan=Yes ))**



**#5** (40 Points)

1. **A telecommunications company is analyzing its customers’ data for those customers that had between 0 and 175 “Day”, “Eve” and “Night” calls. To estimate the missing “Night Calls” field, the company is using k-nearest neighbors.**

* **What would be the value of “Night Calls” for customer x in the table below if:**

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

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer** | **Day Calls** | **Eve Calls** | **Night Calls** |
| A | 110 | 99 | 91 |
| B | 123 | 103 | 103 |
| C | 71 | 88 | 89 |
| D | 113 | 122 | 121 |
| E | 98 | 101 | 118 |
| X | 114 | 110 | ? |

1. **The company has decided to classify “Night Calls” 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 “Night Call” category for X if K=3 and distance weighted vote is used? Why?**

