**Instructions:**

**You can use Word, Excel, Power Point and/or R to answer the questions in this exam. 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. You must supply the R program, and the program 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 Submissions” in CANVAS.**

**A researcher has prepared the following table of patient counts for an addiction clinic.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ethnicity** | **Age Category** | **Alcohol** | **Cocaine** | **Heroin** | **Row Total** |
| **Black** | **Old** | **30** | **48** | **17** | **95** |
|  | **Young** | **20** | **77** | **13** | **110** |
| **Hispanic** | **Old** | **7** | **0** | **10** | **17** |
|  | **Young** | **8** | **7** | **14** | **29** |
| **White** | **Old** | **55** | **2** | **12** | **69** |
|  | **Young** | **31** | **15** | **34** | **80** |
| **Column Total** |  | **151** | **149** | **100** | **400** |

**Problem #1:** (15 Points)

1. **What is the probability that a patient is a Heroin addict?**
   1. **P(heroin addict) = 100/400 = 0.25**
2. **What is the probability that a patient is white?**
   1. **P(white) = 149/400 = 0.3725**
3. **What is the probability that a patient is white and a Heroin addict?**
   1. **P(heroin addict & white) = 46/400 = 0.115**
4. **What is the probability that a patient is white, young and a Heroin addict?**
   1. **P(white & young & heroin addict) = 34/400 = 0.085**
5. **Given that a patient is white and young what is the probability that the patient is addicted to cocaine?**
   1. **P(cocaine addict | white) = 15/80 = 0.1875**
6. **Are race and type of addiction independent?**
   1. **P(white)\*P(heroin) ?= P(white and heroin) ->** 
      1. **0.093125 != 0.115**
   2. **Therefore race and type of addiction are not independent.**

**Problem #2** (30 Points)

**Use the table above and Excel to classify patient addiction type (target values = alcohol, cocaine, heroin) using Ethnicity and Age Category:**

**Construct a classification and regression tree (CART) (one level/split only).**

**Look at midterm.xlsx (Sheet CART)**

**Problem 3:** (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)**

Distance has 3 properties that must hold true:

1. Distance is always non-negative:
   1. (0,0)->(0,1): (0+1)^3 = 1
   2. (0,1)->(0,0): (0+1)^3 = 1
   3. (0,0)->(1,1): (1+1)^3 = 8
   4. (1,1)->(0,0): (1+1)^3 = 8
   5. (0,1)->(1,1): (1+0)^3 = 1
   6. (1,1)->(0,1): (1+0)^3 = 1
2. Distance from A to B = distance from B to A
   1. (0,0)->(0,1) = (0,1)->(0,0)
   2. (0,0)->(1,1) = (1,1)->(0,0)
   3. (0,1)->(1,1) = (1,1)->(0,1)
3. Distance from A to C is <= distance from A to B to C
   1. (0,0)->(0,1)->(1,1) < (0,0)->(1,1) **[1+1 < 8]**
   2. This does not hold, so it is not a proper distance function.

**Problem #4** (20 Points)

**a) Company XYZ is targeting professionals between the ages of 25 and 45 years old with an asset size of 50k 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 = 1 and method = ”unweighted vote” is used**

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | Age | **Asset Size** | **Income** |
| **X** | 28 | 60 | ? |
| **1** | 25 | 50 | 100K |
| **2** | 33 | 60 | 90K |
| **3** | 35 | 80 | 150K |

**Look at midterm.xlsx (Sheet kNN)**

**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** | 28 | 60 | ? |
| **1** | 25 | 50 | Medium |
| **2** | 33 | 60 | Low |
| **3** | 35 | 80 | High |
| **4** | 28 | 60 | Medium |
| **5** | 28 | 60 | High |
| **6** | 28 | 60 | High |

**Look at midterm.xlsx (Sheet kNN)**

**Problem #5** (25 Points)

**The Wisconsin “breast cancer dataset” in CANVAS was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The features in the dataset, have been normalized between 1 and 10.**

Features Domain

-- -----------------------------------------

Sample code number id number

F1. Clump Thickness 1 - 10

F2. Uniformity of Cell Size 1 - 10

F3. Uniformity of Cell Shape 1 - 10

F4. Marginal Adhesion 1 - 10

F5. Single Epithelial Cell Size 1 - 10

F6. Bare Nuclei 1 - 10

F7. Bland Chromatin 1 - 10

F8. Normal Nucleoli 1 - 10

F9. Mitoses 1 - 10

Diagnosis Class: (2 for benign, 4 for malignant)

**Perform the following:**

1. **Load the Breast cancer dataset**
2. **Read the data and:**
   1. **Store every fifth record in a “test” dataset starting with the first record**
   2. **Store the rest in the “training” dataset**
3. **Use knn with k=5 and classify the test dataset**

**> knn1**

**[1] 2 4 2 2 4 2 2 4 4 2 4 4 4 4 2 2 2 4 2 2 2 4 4 4 4 2 4 2 2 4 2 4 4 2 2 2 4**

**[38] 4 4 4 4 4 2 4 4 4 4 2 4 4 2 4 4 2 4 4 4 4 4 2 4 2 2 4 4 4 4 4 2 4 2 2 2 2**

**[75] 2 2 2 2 2 2 4 2 4 2 4 4 2 2 2 2 4 2 2 2 2 2 2 2 2 2 4 2 2 4 2 2 2 2 2 2 4**

**[112] 4 2 2 2 4 2 2 4 4 2 2 4 2 4 2 2 2 2 2 2 2 2 4 2 2 4**

**Levels: 2 4**

1. **Measure the performance of knn**

**> err1**

**[1] 0.05109489**

**> table(Prediction=knn1,Actual=test[,10] )**

**Actual**

**Prediction 2 4**

**2 74 4**

**4 3 56**

**Look at midterm.R**