#1 (10 Points)

Is the following function a proper distance function? Why? Explain your answer.

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_i (x_i - y_i)^2\right)^2$$

Hint: Measure the distance between (0,0), (0,1) and (1,1)

Answer:

1. $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Assume:
x is (0,0), y is (0,1) and z is (1,1)

According to the assumption, get following result:
d(x, y) = d(y, x) = 1 >= 0
d(x, z) = d(z, x) = 4 >= 0
d(y, z) = d(z, y) = 1 >= 0

In addition to d(x, z) = 4 and d(x, y) + d(y, z) = 2

So, as triangle inequality states, this function is not a proper distance function.

# 2 (20 Points)

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912 RMS Titanic sank after colliding with an iceberg, killing hundreds of passengers. Using the "Titanic_rows.CSV" dataset, answer the following questions:

- **What is the probability that a passenger survived? Why?**
- **What is the probability that a passenger survived AND the passenger was staying in the "Crew class" cabin? Why?**
- **What is the probability that a passenger survived GIVEN he/she was staying in the "Crew class" cabin? Why**
- **What is the probability that a passenger survived GIVEN that he ("MALE") was staying in the "2st class" cabin? Why?**
- **Are Survival and Age independent? Why?**
- **Given that a passenger survived, what is the probability that the passenger was a "Female" and was staying in the "2st class" cabin? Why?**

Answer:

To solve this question, I run a R program. It has been included in this .zip file, named **Midterm_2.R.** The comments also show the way I got solutions.

#3 (25 Points)

a) **Company XYZ is targeting professionals between the ages of 25 and 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 = 1 and method = ”unweighted vote” is used

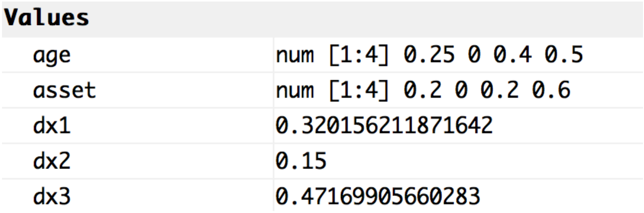
  K = 2 and method = ”unweighted vote” is used

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

Answer:

To solve this question, I run a R program. It has been included in this .zip file, named **Midterm_3.R.**

First, Normalize age and asset, then calculate the distance between x and id = 1, 2, 3. Run Midterm_3.R, I got the following result.



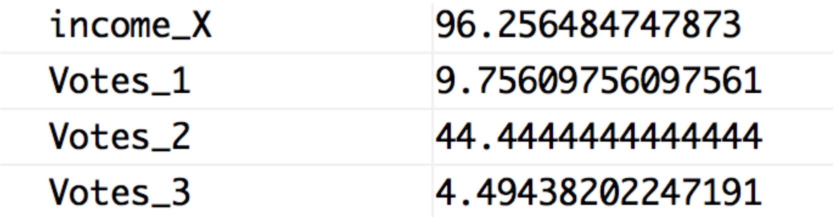
```
Values
  age       num [1:4] 0.25 0 0.4 0.5
  asset     num [1:4] 0.2 0 0.2 0.6
  dx1       0.320156211871642
  dx2       0.15
  dx3       0.47169905660283
```

Thus,

If k = 1, classify id = X according to whichever single point in the training set it is closet to. In this case, X is closest to the id = 2, and therefore X be classified as income 90K.

If k = 2, reclassify X using k-Nearest Neighbor. Now X is closest to id = 2 and id = 1, and therefore X’s income is 95K ((90+100)/2).

If k = 3, reclassify X using a weight voting scheme, this leads to different votes for X. According to the result of programming, the income of X should be about 96K.



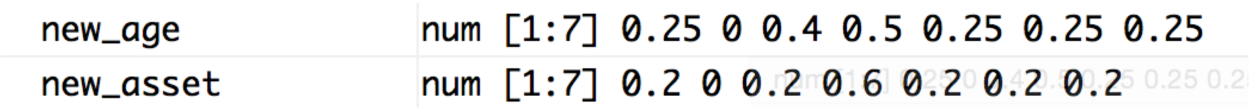
```
income_X     96.256484747873
Votes_1      9.75609756097561
Votes_2      44.4444444444444
Votes_3      4.49438202247191
```

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, see table below.

- **What would be the income category for X if K=3 and distance weighted vote is used? Why?**

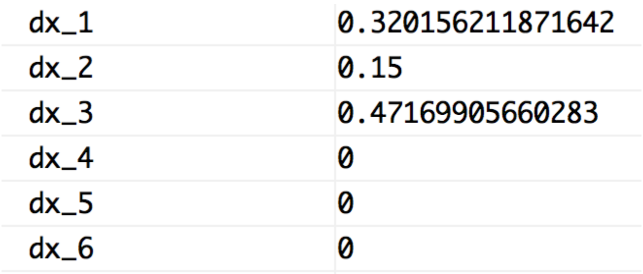
Answer:

First, normalize age and asset.



```
new_age      num [1:7] 0.25 0 0.4 0.5 0.25 0.25 0.25
new_asset    num [1:7] 0.2 0 0.2 0.6 0.2 0.2 0.2
```

Second, calculate the distance between x and id = 1-6

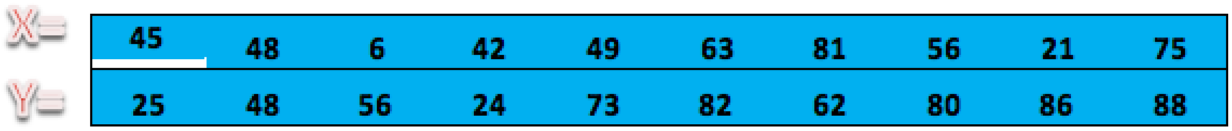


```
dx_1    0.320156211871642
dx_2    0.15
dx_3    0.47169905660283
dx_4    0
dx_5    0
dx_6    0
```

Third, the inverse distance of 0 is undefined using weighted voting. **Thus**, if k = 3, the three closest records should be id = 4,5,6. For these three records, two of the three closet records to X are High.

#4 (15 Points)

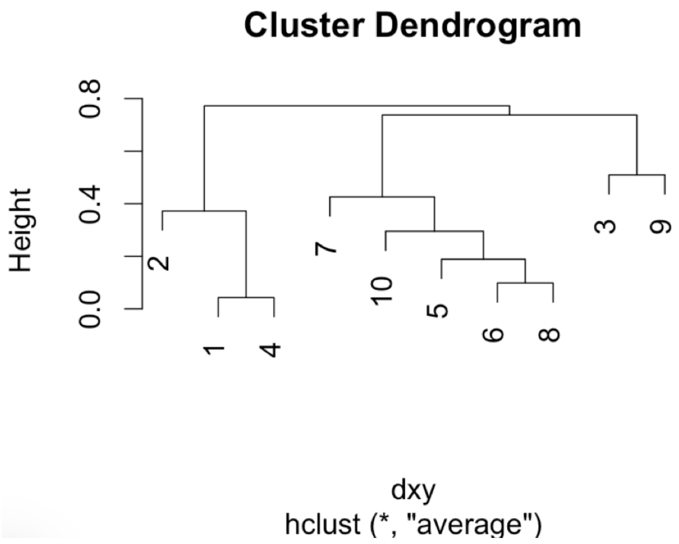
- **Using “hclust” in R cluster the following 10 points.**



X=	45	48	6	42	49	63	81	56	21	75
Y=	25	48	56	24	73	82	62	80	86	88

Answer:

To solve this question, I run a R program. It has been included in this .zip file, named **Midterm_4.R**
Step 1: normalized the data
Step 2: computed the distance by using dist() function
Step 3: used hclust() function to do hierarchical clustering with “average” linkage.
Step 4: plotted the result, as below Cluster Dendrogram.



#5 (10 Points)

Using R perform the following:

a) Load the following CSV file to your R environment:

http://www.math.smith.edu/sasr/datasets/help.csv

b) Create a dataframe of: id, age, "number of days any substance used" (daysanysub), substance, and race group

c) Normalize "number of days any substance used" (daysanysub)

d) Substitute the missing values of "daysanysub" with zero

e) Calculate: Mean, Max, Median, STD of Age

f) Create a categorical variable "age_group" as:

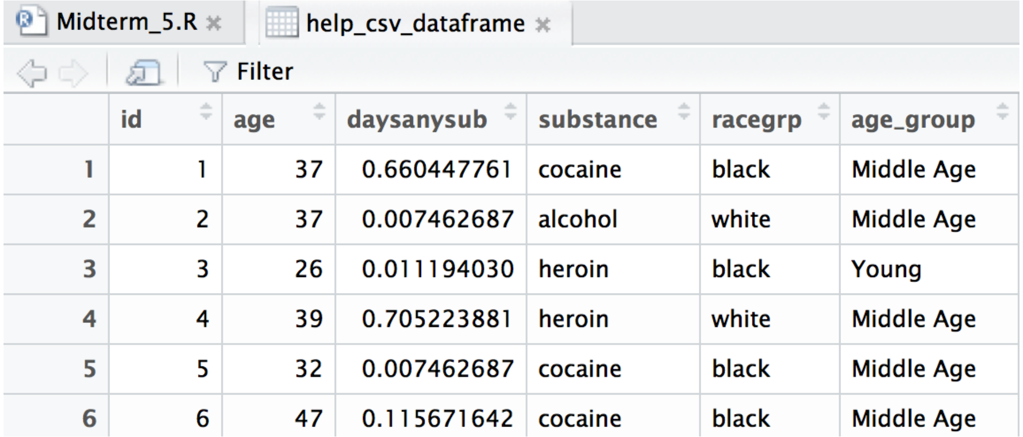
i. From 0 up to and including 30 = "Young"

ii. Over 30 up to and including 60 = "Middle Age"

iii. Older than 60 = "OLD"

Answer:

To solve this question, I run a R program. It has been included in this .zip file, named **Midterm_5.R.** The below picture is a part of final data frame I got after doing these steps.



	id	age	daysanysub	substance	racegrp	age_group
1	1	37	0.660447761	cocaine	black	Middle Age
2	2	37	0.007462687	alcohol	white	Middle Age
3	3	26	0.011194030	heroin	black	Young
4	4	39	0.705223881	heroin	white	Middle Age
5	5	32	0.007462687	cocaine	black	Middle Age
6	6	47	0.115671642	cocaine	black	Middle Age

#6 (20 Points)

It is believed that cancerous tissues have larger nuclei with rougher surfaces. Today, automated image analysis can collect measurements of the nuclei of the cells in a picture of a sample without human intervention. The Wisconsin Breast Cancer Dataset (wisc_bc_data.csv), represents the automated measurements of 569 samples, some benign and some malignant.

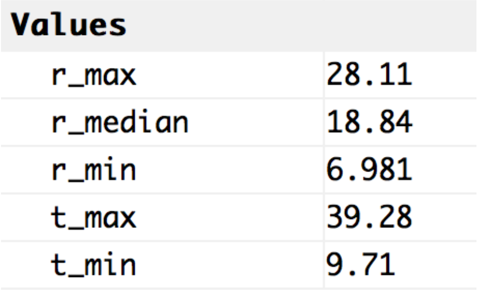
- **Load the "wisc_bc_data.csv" dataset into R**
- **Find max, min and the median of the "radius_mean " and "texture_mean" for all the observations**
- **select every 7th observation (row) to create the test data set.**
- **Use the remaining data as the training dataset**
- **Use "radius_mean ", and "texture_mean" columns in the training dataset and the knn function(k=5) to classify observations in the test dataset as either benign (Diagnosis=B) or malignant(Diagnosis=M)**
- **Measure the performance of knn k=5**

Answer:

To solve this question, I run a R program. It has been included in this .zip file, named **Midterm_6.R.** The below picture is a part of final data frame I got after doing these steps.

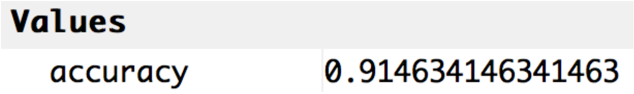
Here are some parts of result for these question.

1. max, min and the median of the "radius_mean " and "texture_mean" for all the observations



Values	
r_max	28.11
r_median	18.84
r_min	6.981
t_max	39.28
t_min	9.71

2. the performance of knn k =5 :



Values	
accuracy	0.914634146341463