Assignment 4 - Harnoor Rangi

Question 1
1.1

First, we compute the support for all the 1-item sets:

1-ITEM SETS	SUPPORT
milk	0.5
bread	0.4
eggs	0.4
coffee	0.3
juice	0.3
cookies	0.2
butter	0.2

The min support required is 0.2, so all 1-item sets satisfy this requirement, i.e. they are all frequent.

For the next iteration, we examine 2-item sets composed of the frequent 1-item sets. The number of potential 2-item sets is 21 (i.e., 7 items taken 2 at a time). The 2-item sets that satisfy the min support of 0.2 are the following:

2-ITEM SETS	SUPPORT
milk,bread	0.4
milk,eggs	0.3
bread,eggs	0.3

For the next iteration, we examine 3-item sets composed of the frequent 2-item sets. The 3-item sets that satisfy the min support of 0.2 are the following:

3-ITEM SETS	SUPPORT
milk,bread,eggs	0.3

1.2

Answer: There is only one frequent itemset of size 3, i.e., {milk,bread,eggs}

Confidence = Support/ Number of times it occurs

Milk , Bread —> eggs = ¾ = 0.75

Bread —> milk , eggs = ¾ = 0.75

Question 2: Decision Tree

We start by computing the entropy for the entire set. We have 7 positive samples and 3 negative samples.

The entropy, I(7,3), is -(7/10 * log (7/10) + 3/10 * log(3/10)) = 0.88

AGE ATTRIBUTE
We consider the first attribute AGE.
There are 4 values for age 20..30 appears 5 times
I(s11, s21) = -(4/5 * log(4/5) + 1/5 * log(1/5)) = 0.72

31..40 appears 2 times
I(s12, s22) = -(1/2 * log(1/2) + 1/2 * log(1/2)) = 1

41..50 appears 2 times
I(s13, s23) = -(2/2 * log(2/2) = 0

51..60 appears 1 time
I(s14, s24) = -(1/1 * log(1/1) = 0

E(AGE) = 5/10 * 0.72 + 2/10 * 1 + 2/10 * 0 + 1/10 * 0 = 0.56
GAIN(AGE) = 0.88 - 0.56 = 0.32

CITY ATTRIBUTE
We consider the second attribute CITY. There are 3 values for city
LA occurs 2 times
I(s11, s21) = -(1/2 * log(1/2) + 1/2 * log(1/2)) = 1

NY occurs 7 times
I(s12, s22) = -(2/7 * log(2/7) + 5/7 * log(5/7)) = 0.86

SF occurs 1 times
I(s13, s23) = -(1/1 * log(1/1) = 0

E(CITY) = 2/10 * 1 + 7/10 * 0.86 + 1/10 * 0 = 0.80
GAIN(CITY) = 0.88 - 0.80 = 0.08

GENDER ATTRIBUTE
We consider the third attribute GENDER. There are 2 values
F occurs 7 times
I(s11, s21) = -(2/7 * log(2/7) + 5/7 * log(5/7)) = 0.86

M occurs 3 times
I(s12, s22) = -(1/3 * log(1/3) + 2/3 * log(2/3)) = 0.92

E(GENDER) = 0.88
GAIN(GENDER) = 0

EDUCATION ATTRIBUTE
We consider the fourth attribute EDUCATION. There are 3 values

HS occurs 2 times
I(s11, s21) = -(2/2 * log(2/2) = 0

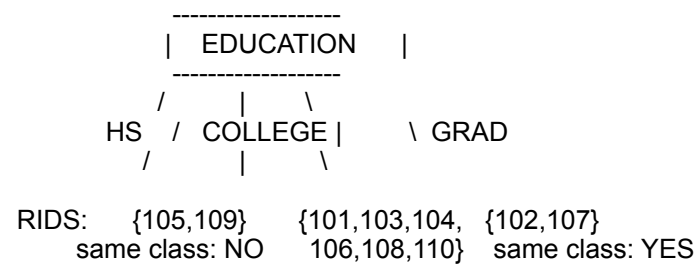
COLLEGE occurs 6 times
I(s12, s22) = -(1/6 * log(1/6) + 5/6 * log(5/6)) = 0.65

GRAD occurs 2 times
I(s13, s23) = -(2/2 * log(2/2) = 0

E(EDUCATION) = 0.39
GAIN(EDUCATION) = 0.49

The greatest gain is for the EDUCATION attribute.

The tree at this point would look like the following:



Only the middle node is not a LEAF node, so continue with those records and consider only the remaining attributes.

The entropy, I(5,1), is -(5/6* log (5/6) + 1/6 * log(1/6)) = 0.65

<u>AGE</u>
We consider the first attribute AGE. There are 4 values for age
20..30 appears 3 times

I(s11, s21) = -(3/3 * log(3/3) = 0
31..40 appears 1 time

I(s12, s22) = -(1/1 * log(1/1) = 0
41..50 appears 1 time

I(s13, s23) = -(1/1 * log(1/1) = 0
51..60 appears 1 time

I(s14, s24) = -(1/1 * log(1/1) = 0

E(AGE) = 0
GAIN(AGE) = 0.65

<u>CITY</u>
We consider the second attribute CITY. There are 2 values for city
NY occurs 1 time
I(s11, s21) = -(1/1 * log(1/1) = 0

SF occurs 5 times
I(s12, s22) = -(1/5 * log(1/5) + 4/5 * log(4/5)) = 0.72

E(CITY) = 0.60
GAIN(CITY) = 0.05

We consider the third attribute GENDER. There are 2 values

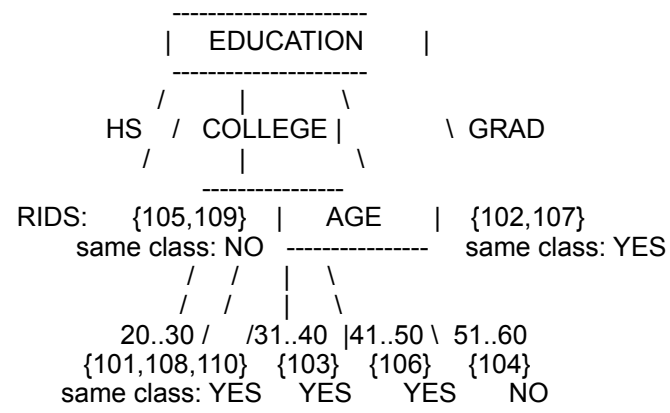
F occurs 5 times
I(s11, s21) = -(1/5 * log(1/5) + 4/5 * log(4/5)) = 0.72

M occurs 1 time
I(s12, s22) = -(1/1 * log(1/1) = 0

E(GENDER) = 0.60
GAIN(GENDER) = 0.05

The greatest gain is for the AGE attribute.

The tree at this point would look like the following :



Question 3

3.1

We start by specifying the centroid for each of the 3 clusters.
C1's centroid is (8,4) , i.e., record with rid = 1
C2's centroid is (2,4) , i.e., record with rid = 3
C3's centroid is (2,8) , i.e., record with rid = 5

We now place the remaining records in the cluster whose centroid is closest.

The distance between record 2, i.e., point (5,4), and centroid for C1 is

$$\text{SQROOT}(|8-5|^2 + |4-4|^2) = 3$$

The distance between record 2 and centroid for C2 is

$$\text{SQROOT}(|2-5|^2 + |4-4|^2) = 3$$

The distance between record 2 and centroid for C3 is

$$\text{SQROOT}(|2-5|^2 + |8-4|^2) = 5$$

The distance between record 4, i.e., point (2,6), and centroid for C1 is

$$\text{SQROOT}(|8-2|^2 + |4-6|^2) = 6.3$$

The distance between record 4 and centroid for C2 is

$$\text{SQROOT}(|2-2|^2 + |4-6|^2) = 2$$

The distance between record 4 and centroid for C3 is

$$\text{SQROOT}(|2-2|^2 + |8-6|^2) = 2$$

The distance between record 6, i.e., point (8,6), and centroid for C1 is

$$\text{SQROOT}(|8-8|^2 + |4-6|^2) = 2$$

The distance between record 6 and centroid for C2 is

$$\text{SQROOT}(|2-8|^2 + |4-6|^2) = 6.3$$

The distance between record 6 and centroid for C3 is

$$\text{SQROOT}(|2-8|^2 + |8-6|^2) = 6.32$$

Record 6 is closest to centroid of cluster C1 and is placed there.

#first iteration

	2	4	6
1	3	6.3	2
3	3	2	6.3
5	5	2	6.3

We now recalculate the cluster centroids:

C1 contains records {1,2,6} with a centroid of
((8+5+8)/3, (4+4+6)/3) = (7, 4.67)

C2 contains records {3,4} with a centroid of
((2+2)/2, (4+6)/2) = (2, 5)

C3 contains record {5} with a centroid of (2, 8)

We now make a second iteration over the records, comparing the distance of each record with the new centroids and possibly moving records to new clusters. As it turns out, all records stay in their prior cluster assignment. Since there was no change, the algorithm terminates.

3.2

Classification is a supervised learning method where the possible outcome is known in advance, whereas clustering is a unsupervised learning technique where the output is not known.