CS 412 Asg5

Guangyu Zhou

Q1

1. What's the information gain for the \Price" attribute? Please show your calculation.

Ans:

In total, P: 7 , NP: 5

Info(D) = I(7, 5) = -7/12*log(7/12) - 5/12*log(5/12) = 0.97987

	Low	Medium	High
P	4	2	1
NP	1	1	3

Info_price(D) = 5/12*I(4, 1) + 3/12*I(2, 1) + 4/12*I(1, 3)
= 5/12*(-4/5*log base 2 (4/5) - 1/5*log base 2 (1/5)) + 3/12*(-1/3*log base 2 (1/3) - 2/3*log base 2 (2/3)) + 4/12*(-1/3*log base 2 (1/3) - 3/4*log base 2 (3/4))
= 0.81024

So Gain(Price) = Info(D) - Info_price(D) = 0.1696

2. Now suppose we want to use Gini Index as attribute selection measure. What's the Gini index for the attribute Parking? What's the reduction in impurity in terms of Gini Index? Please show your calculation.

Ans: Parking partitions D into {No} and {Available}.

	No	Available
P	2	5
NP	2	3

Gini(D) = 1-(5/12)^2 – (7/12)^2 = 0.48611

Gini_ parking (D) = 4/12*Gini(2,2) + 8/12*Gini(5,3)

$= \frac{4}{12} * \left(1 - \frac{1}{2}^2 - \frac{1}{2}^2\right) + \frac{8}{12} * \left(1 - \frac{3}{8}^2 - \frac{5}{8}^2\right) = \frac{23}{48} \approx 0.47917$

Reduction in impurity = Gini(D) - Gini_ parking (D) = 1/144 = 0.00694

3. Based on the training data, we want to construct a Naive Bayes classifier. Please estimate the following terms (No smoothing is required, and please show your calculation):

Ans:

a) Pr(Popularity = `P') = 7/12; and Pr(Popularity = `N') = 5/12

b) Pr(Price = 'Low', Parking = `Available' , Cuisine = `Mexican' | Popularity = `P')
= Pr(Low|P)*Pr(Available|P)*Pr(Mexican|P) = 4/7*5/7*2/7 = 40/343

c) Pr(Price = 'Low', Parking = `Available' , Cuisine = `Mexican' | Popularity = `N')
= Pr(Low|NP)*Pr(Available|NP)*Pr(Mexican|NP) = 1/5*3/5*2/5 = 6/125

4. Suppose a restaurant has the values: Price = 'Low', Parking = `Available' , Cuisine = `Mexican'. Based on the calculation in 3, is this restaurant classified as popular?

P1 = P(P | Price = 'Low', Parking = `Available' , Cuisine= `Mexican')

$$= \frac{\mathrm{P}\left(\text{Price =low ,Parking = `Available}',\text{Cuisine = `Mexican}'\middle|\mathrm{P}\right)*\mathrm{P}(\mathrm{P})}{P(\text{ Price = 'Low',Parking = `Available' ,Cuisine= `Mexican'})}$$

$$= \frac{\left(\frac{40}{343}*\frac{7}{12}\right)}{P(\text{ Price = 'Low',Parking = `Available' ,Cuisine= `Mexican'})}$$

P2 = P(NP | Price = 'Low', Parking = `Available' , Cuisine= `Mexican')

$$= \frac{\mathrm{P}\left(\text{Price =low ,Parking = `Available}',\text{Cuisine = `Mexican}'\middle|\mathrm{NP}\right)*\mathrm{P}(\mathrm{NP})}{P(\text{ Price = 'Low',Parking = `Available' ,Cuisine= `Mexican'})}$$

$$= \frac{\left(\frac{6}{125}*\frac{5}{12}\right)}{P(\text{ Price = 'Low',Parking = `Available' ,Cuisine= `Mexican'})}$$

As P1 > P2, it is classified as Popular.

Q2:

1. According to training data, we know that red is +1, blue is -1.

For each test data, we compute and compare the distance to all training points to get the K=3 nearest neighbors and assign labels according to it.

train				Euclidean dist^2 (Test to train)			
x1	x2	y		1	2	3	4
1	0.5	1		7.73	2.5	4.25	0.29
2	1.2	1		2.74	0.29	1.94	0.68
2.5	2	1		0.53	1	1.25	2.69
3	2	1		0.58	1.25	2.5	4.24
1.5	2	-1		1.93	2	0.25	1.09
2.3	3	-1		0.25	4.04	0.89	5.21
1.2	1.9	-1		2.89	2.5	0.45	0.81
0.8	1	-1		6.5	2.89	2.74	0.16
			Label(y)	1	1	-1	1

Those 3 yellow cells in each column corresponds to the 3 nearest points to current test point. We get the classification labels shown in green, with error rate = 1/4

2. $\text{sign}(\mathrm{w}^{\mathrm{T}}\mathrm{x}) = sign(w_0 + w_1x_1 + w_2x_2) = sign(1 + 0.5 * 0.8 - 1 * 1) = sign(0.4) = 1$ which is not -1. So it is NOT correctly classified.

Adjustment: Update $\mathrm{w} = \mathrm{w} + \eta\mathrm{x}\mathrm{y}$

$\mathrm{w}_0 = w_0 - 0.1 * 1 = 0.9$

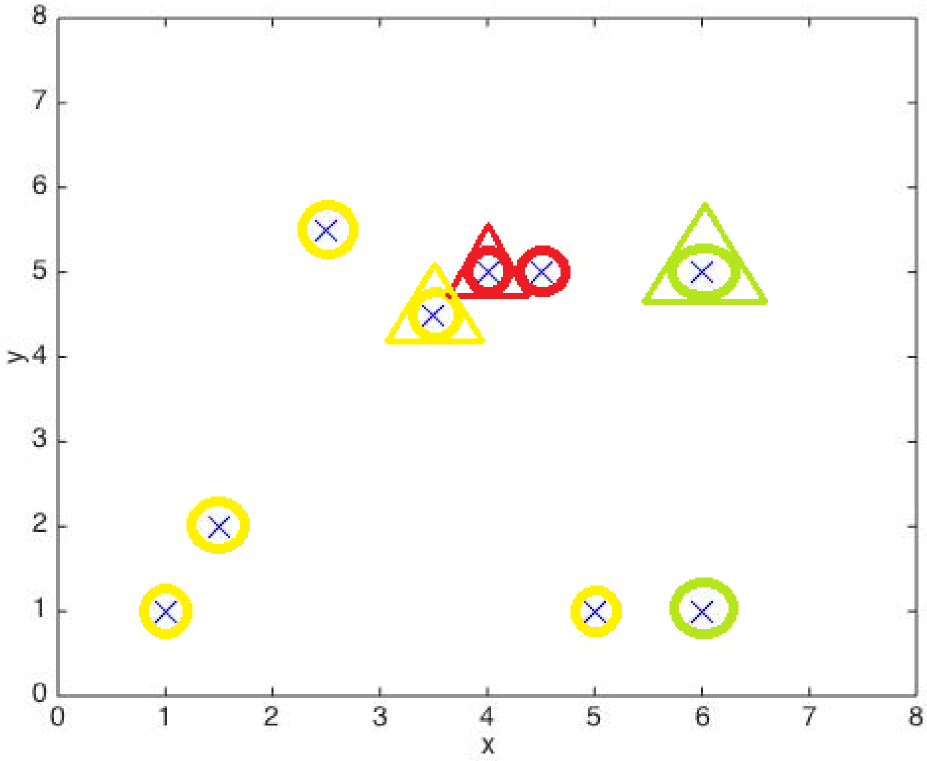
$\mathrm{w}_1 = w_1 - 0.1 * 0.8 = 0.42$

$\mathrm{w}_2 = w_2 - 0.1 * 1 = -1.1$

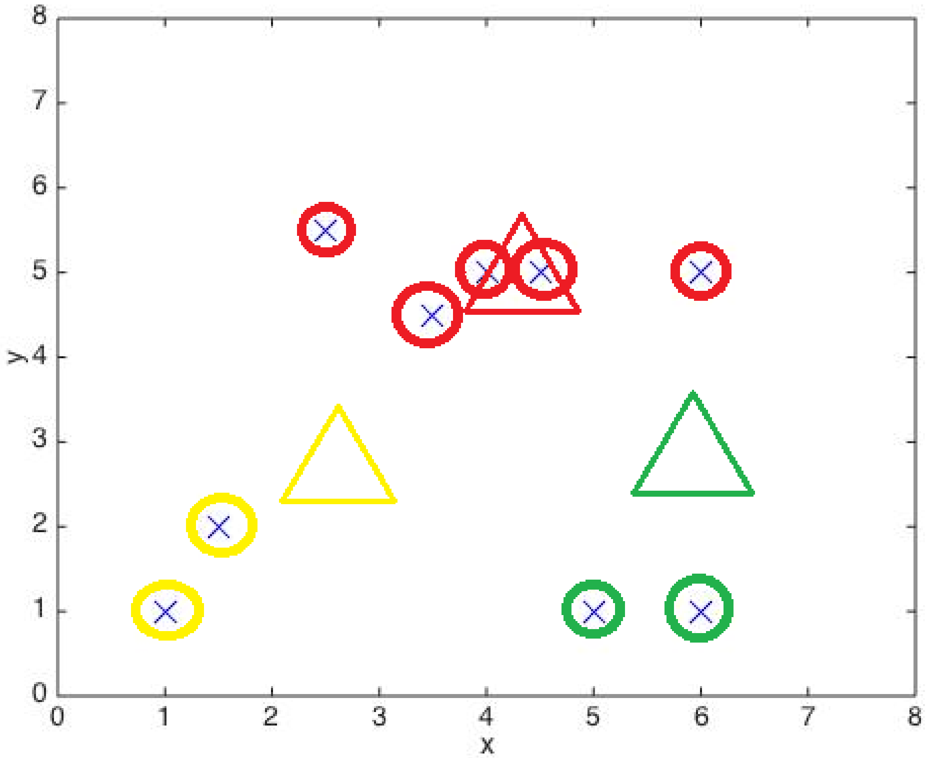
Q3:

1. Note: Triangles represent cluster centers

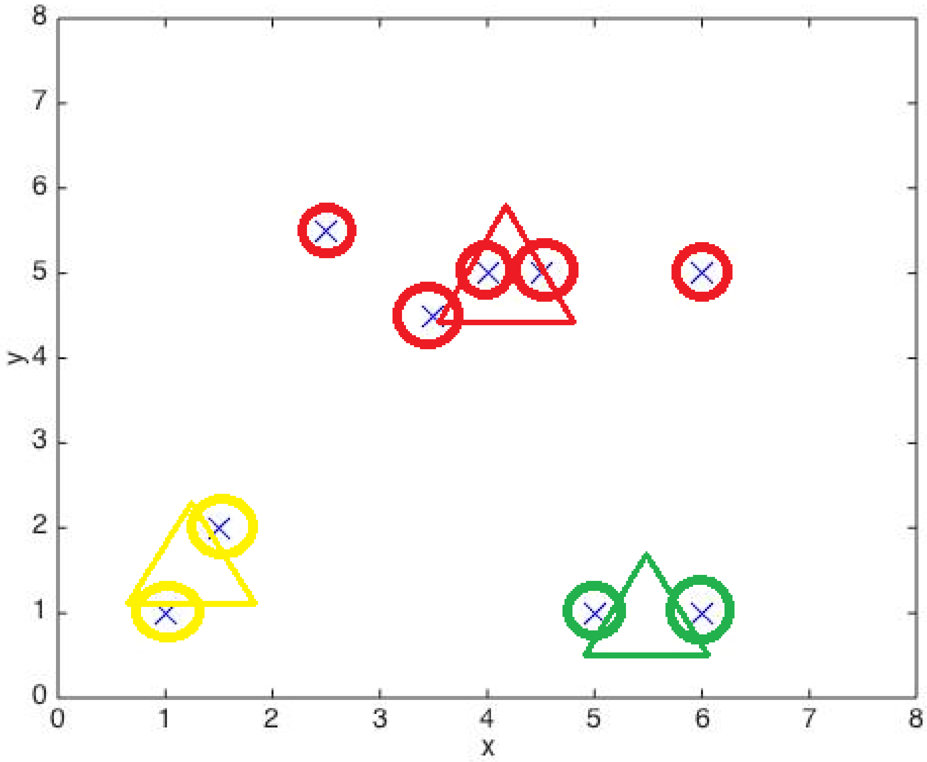
Initial	Center Cord x	Center Cord y
Cluster 1	4	5
Cluster 2	3.5	4.5
Cluster 3	6	5



Round 2:	Center Cord x	Center Cord y
Cluster 1	4.25	5
Cluster 2	2.7	2.8
Cluster 3	6	3

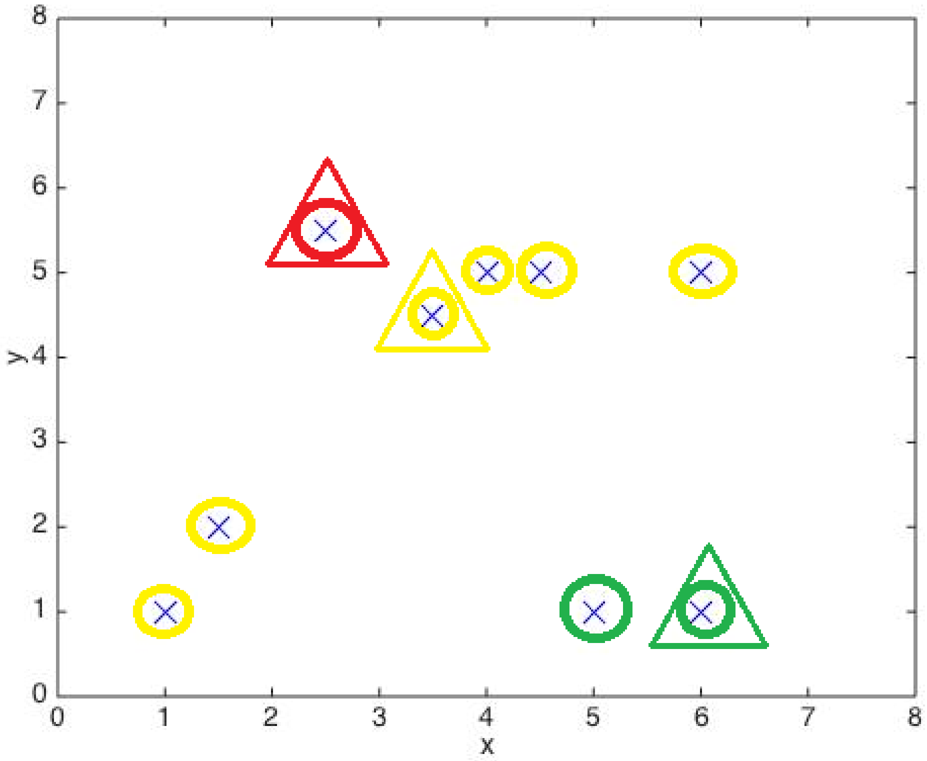


Round 3:	Center Cord x	Center Cord y
Cluster 1	4.1	5
Cluster 2	1.25	1.5
Cluster 3	5.5	1

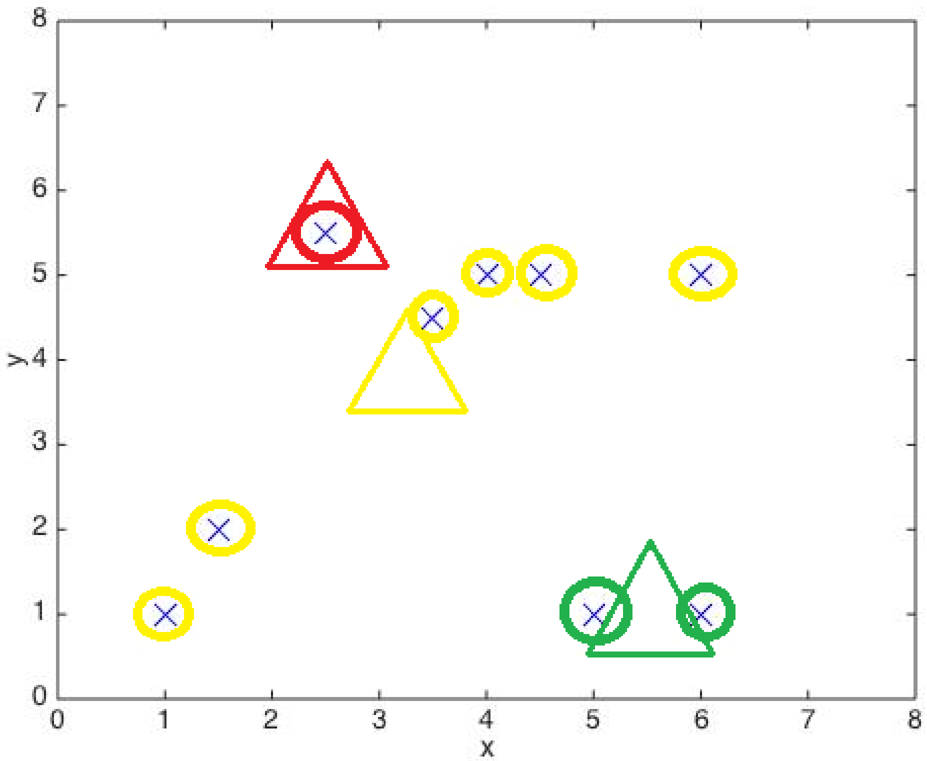


2. Note: Triangles represent cluster centers

Initial	Center Cord x	Center Cord y
Cluster 1	2.5	5.5
Cluster 2	3.5	4.5
Cluster 3	6	1



Round 2	Center Cord x	Center Cord y
Cluster 1	2.5	5.5
Cluster 2	3.416667	3.75
Cluster 3	5.5	1



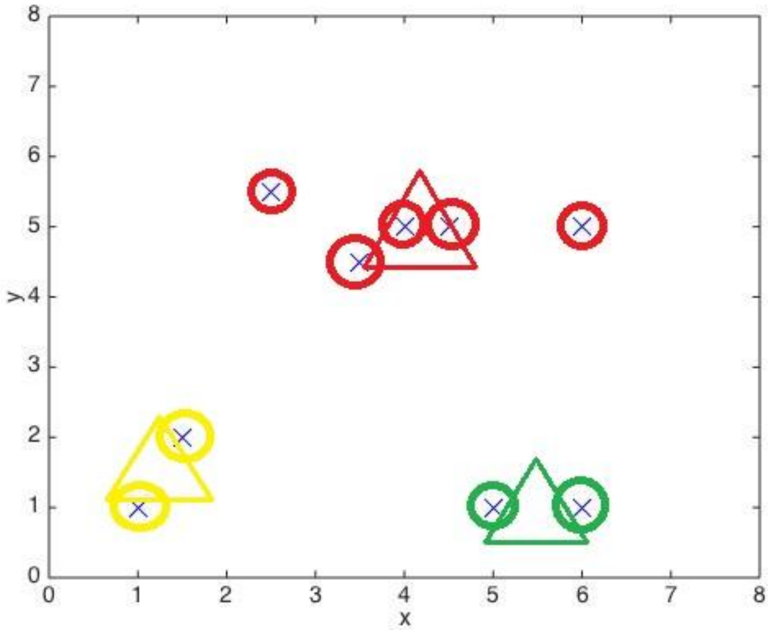
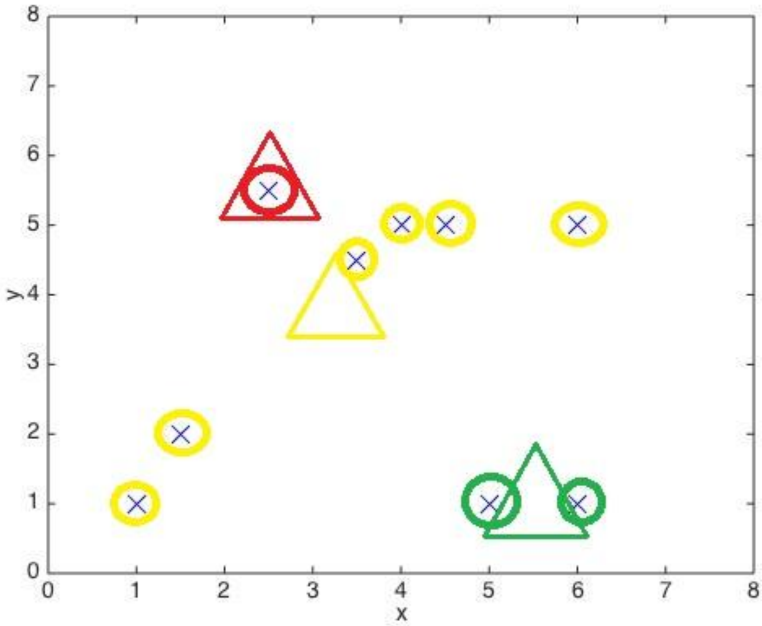
3.

1) They are different because:

It terminates at local optimal. As for the 2nd, the cluster graphs is more distorted by the outliers.

2) The first one is better. As can be seen from the two graphs below:

The 1st one produced high quality clusters with more cohesive within clusters and distinctive between clusters than the 2nd one.

<table>
<tr><td>1st</td><td>
</td></tr>
<tr><td>2nd</td><td>
</td></tr>
</table>

3)
4)