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*(-1/4*log base 2 (1/4) - 3/4*log base 2 (3/4))

= 0.800803

So Gain(Price) = Info(D) - Info_price(D) =0.179067

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')

$$= \frac{P(\text{Price = low , Parking = `Available', Cuisine = `Mexican'} \mid P) * P(P)}{P(\text{ Price = 'Low', Parking = `Available' , Cuisine= `Mexican'})}$$

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

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

$$= \frac{P(\text{Price = low , Parking = `Available', Cuisine = `Mexican'} \mid NP) * P(NP)}{P(\text{ Price = 'Low', Parking = `Available' , Cuisine= `Mexican'})}$$

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

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. The error is the 4th point label, which should be -1 in this case, with error rate = 1/4

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

Adjustment: Update $\text{w} = \text{w} + \eta\text{xy}$

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

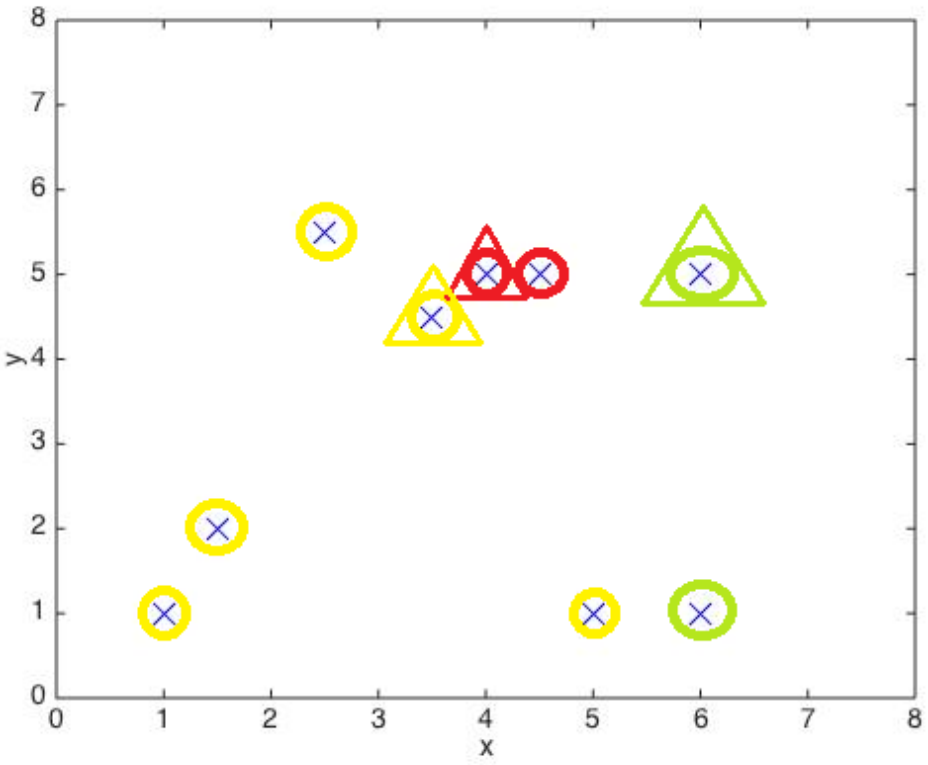
$\text{w}_1 = w_1 - 0.1 * 0.8 = 0.42$

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

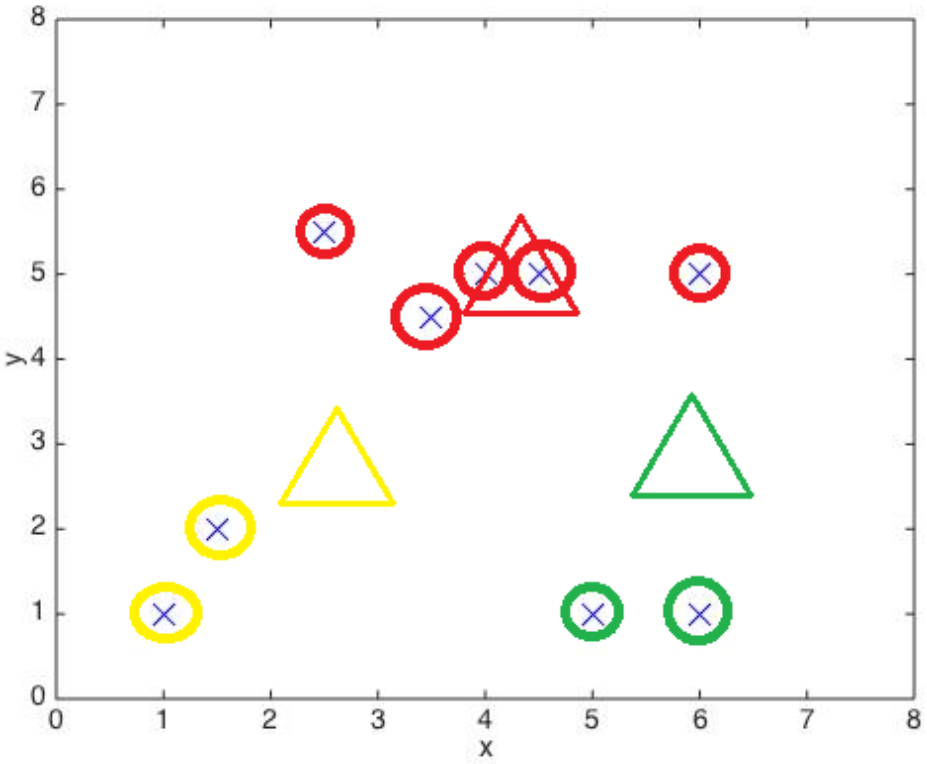
Q3:

1. Note: Triangles represent cluster centers

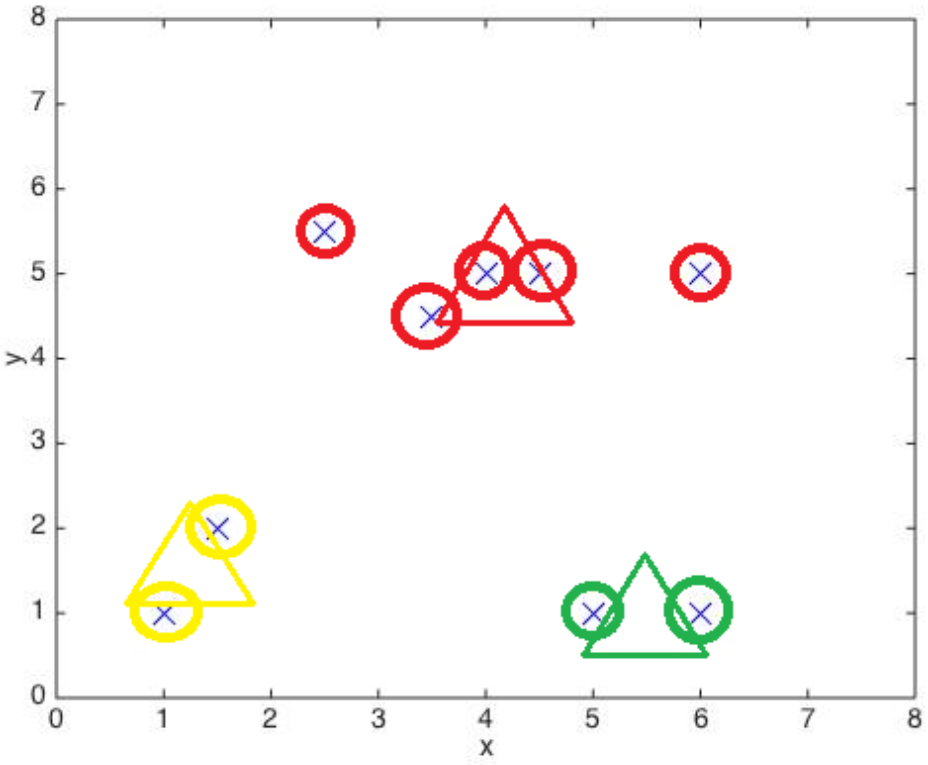
Initial	Center Cord x	Center Cord y
Cluster 1	4	5
Cluster 2	3.5	4.5
Cluster 3	6	5



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

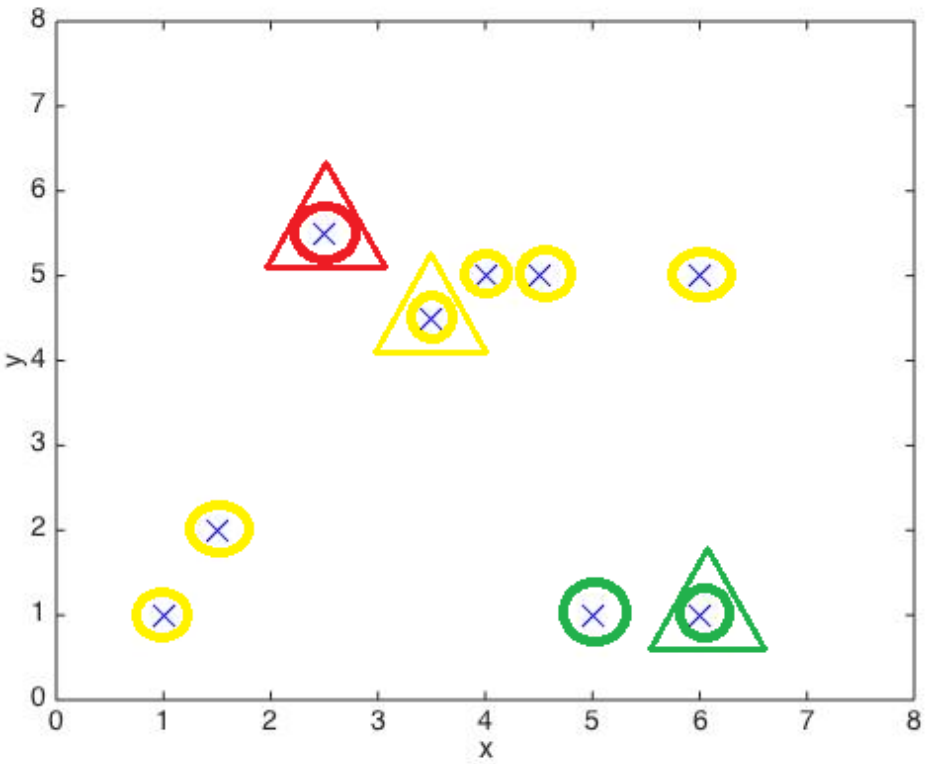


Round 3:	Center Cord x	Center Cord y
Cluster 1	4.1	5
Cluster 2	1.25	1.5
Cluster 3	5.5	1

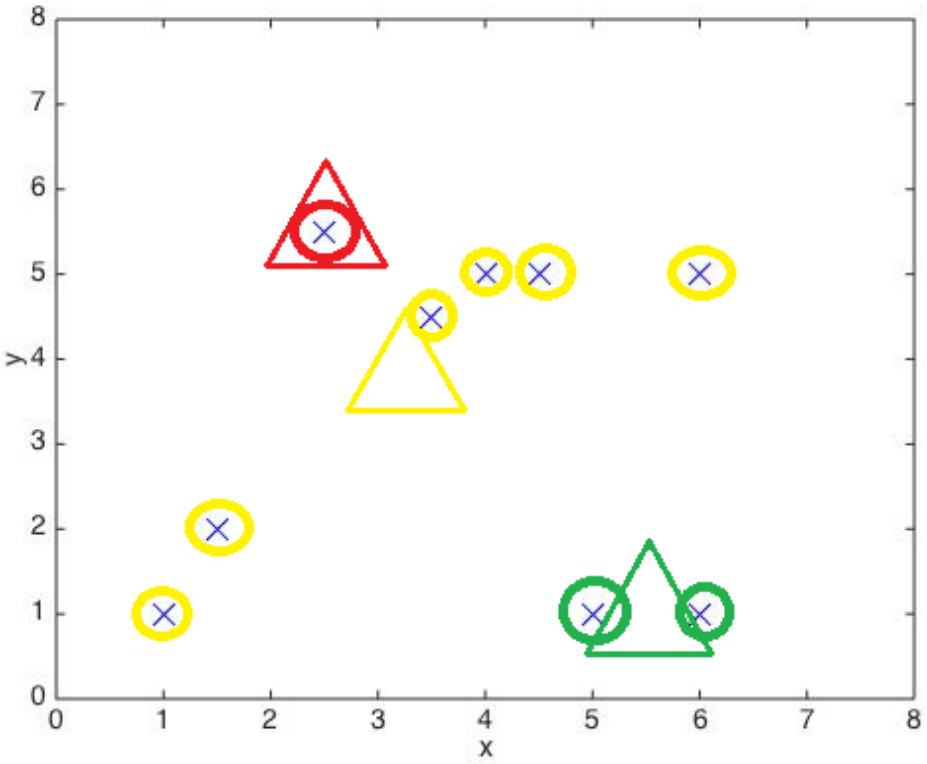


2. Note: Triangles represent cluster centers

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



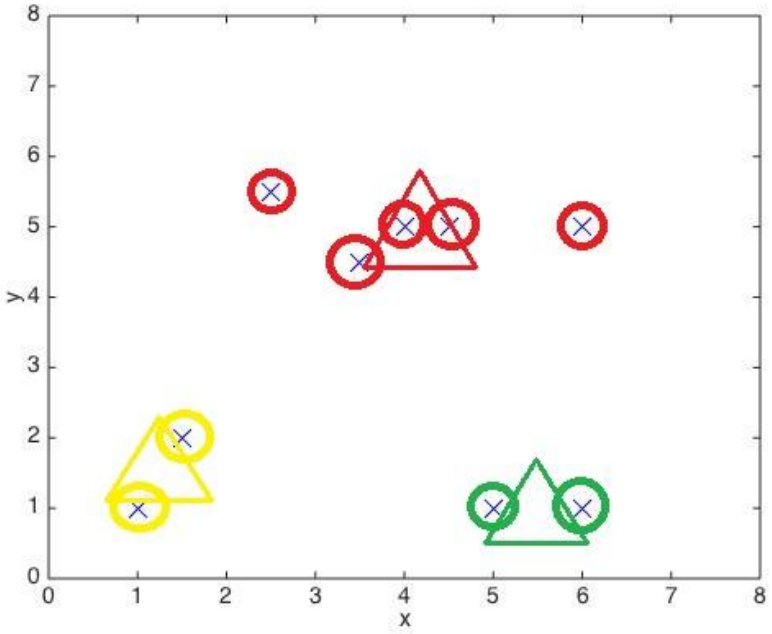
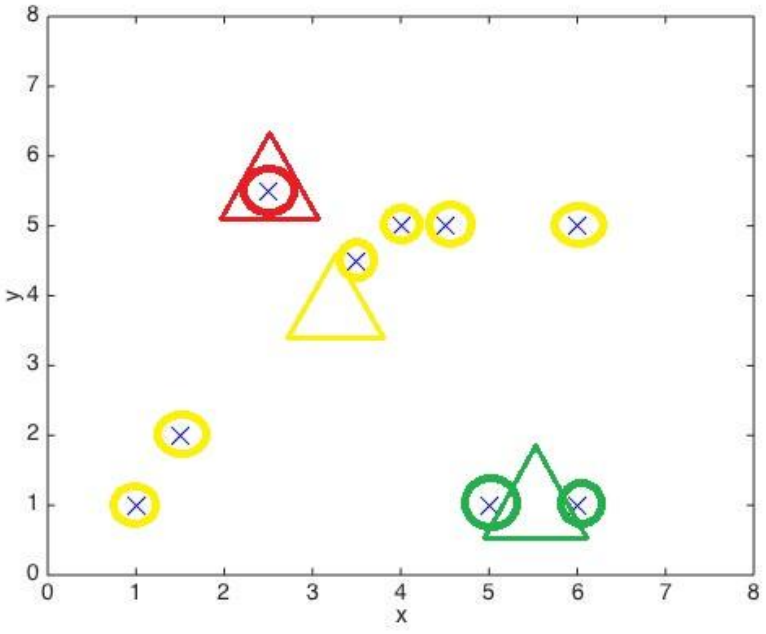
3.

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.

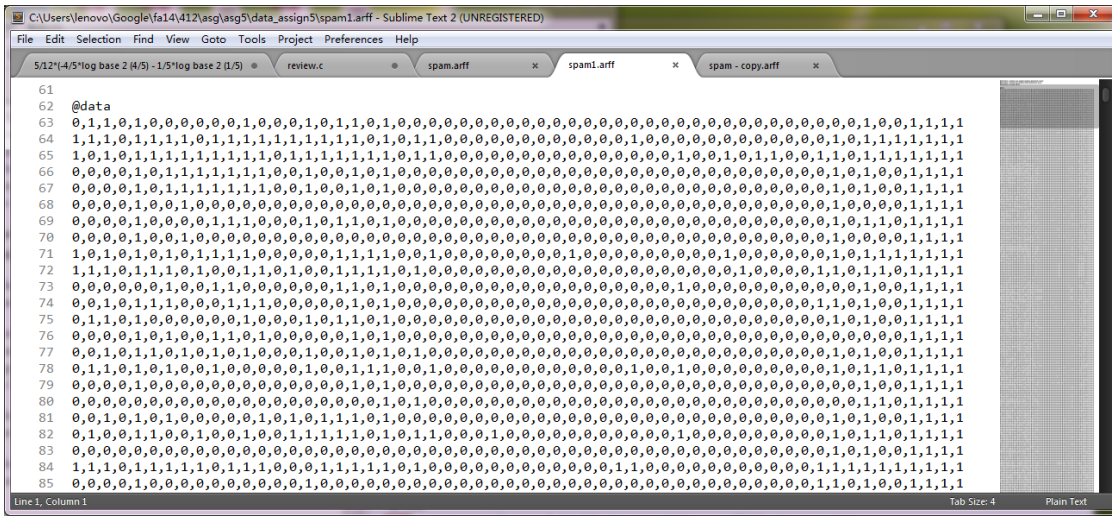
1st	
2nd	

3) Quality measurement criteria: We can calculate the total distance among each clusters and add them up. The smaller this value is, the better the cluster is. For this example, the 1st case is better, because the 2nd one have spread yellow cluster.

4) We can try to exclude outliers from the initial cluster center or try out multiple starting points and choosing the clustering with lowest cost (just like the above process).

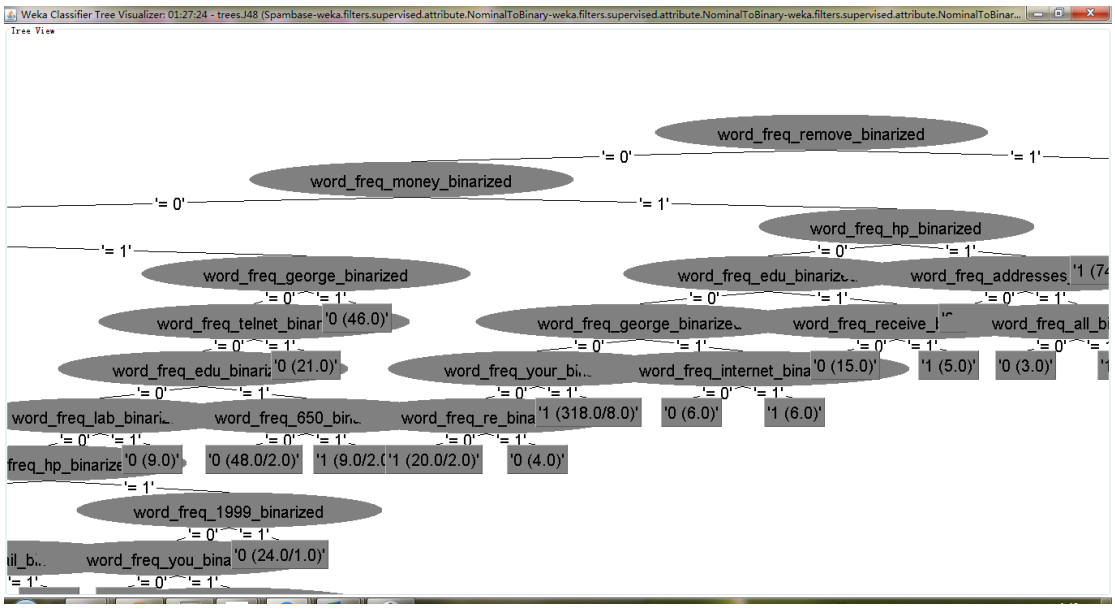
Mini-Mp:

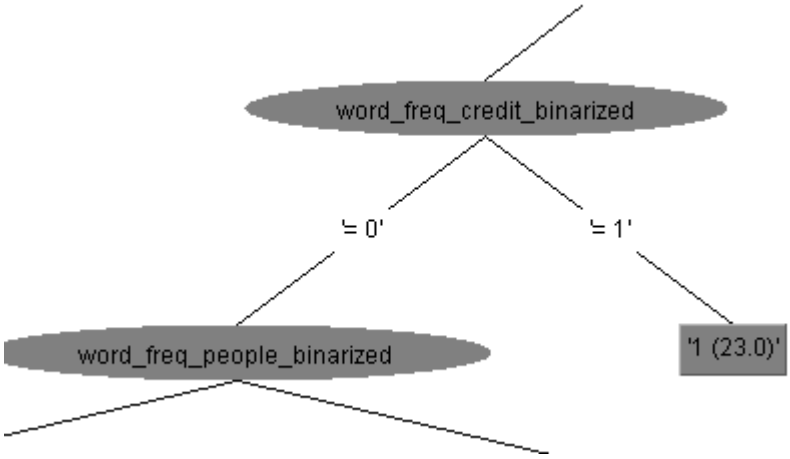
1. Binary Attribute:



2. J48 Result:

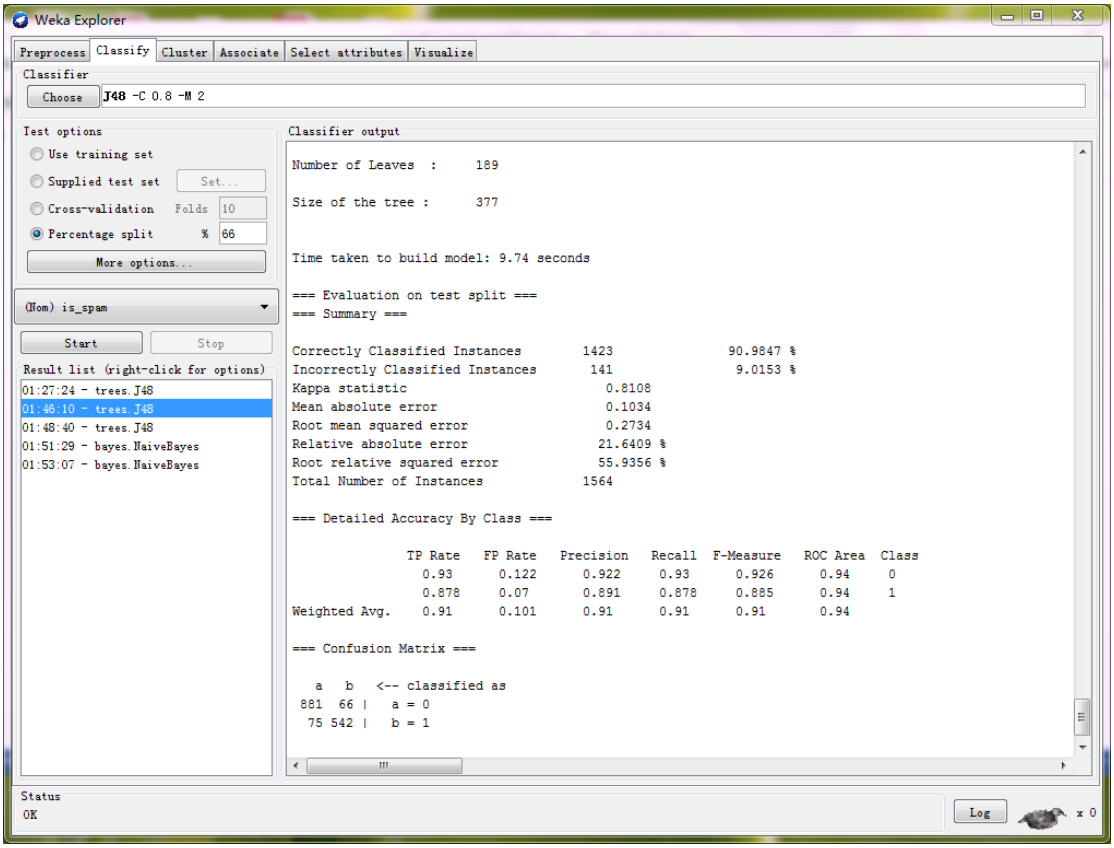
```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        1426               91.1765 %
Incorrectly Classified Instances       138                8.8235 %
Kappa statistic                          0.8131
Mean absolute error                      0.1229
Root mean squared error                  0.2782
Relative absolute error                 25.7356 %
Root relative squared error             56.9163 %
Total Number of Instances             1564

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.949     0.146      0.909     0.949     0.929      0.937    0
                 0.854     0.051      0.917     0.854     0.884      0.937    1
Weighted Avg.    0.912     0.108      0.912     0.912     0.911      0.937

=== Confusion Matrix ===

   a   b   <-- classified as
 899  48 |   a = 0
  90 527 |   b = 1
```



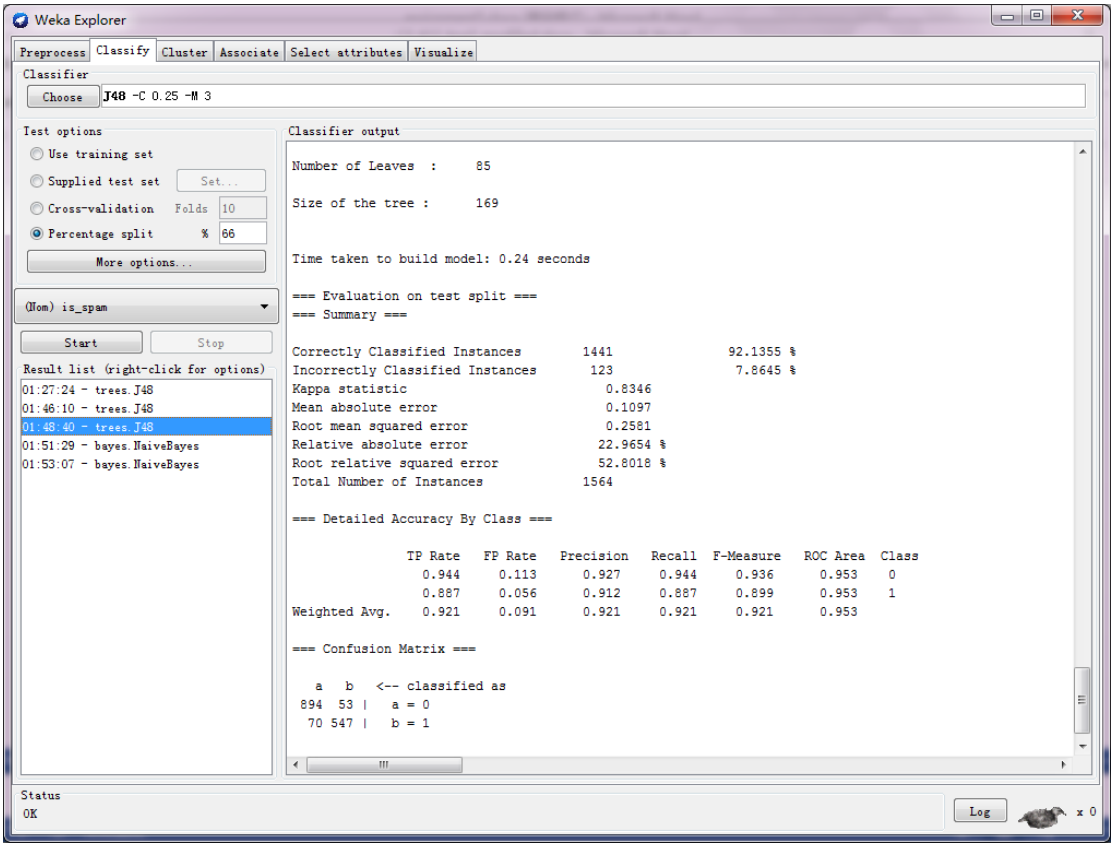


Interesting Rule: Word freq with credit turns out more likely to be classified as spam

3. Play with weka:
   1) **Confidencefactor** is used for pruning. I chose 0.8, larger than default 0.25, then, the tree get much bigger.



2) **MinnumberObj** is to ensure the instances of leaves is bigger than this number. After I change this from 2 to 3, the leaves decreases with bigger instance.



4. Naïve Bayes: