CAP 6777 Web Mining – Khoa Hoang

2016 Summer

Text Clustering and Classification [15 points, Due: 06/16]

Part I: Questions and Answers [0.5 pt/each]

1. *What is Entropy? Please show the formula, and explain how to use Entropy to quantify the randomness of the system.*

⇨ Entropy is the formula to calculate the average uncertainty, impurity of information when we're not sure the outcome of an information source. The higher the entropy, the more the information contents

⇨ Entropy = $\sum_i - p_i \log_2 p_i$

p_i *is the probability of class i*

2. *What is Bayes Rule? Please explain how to use Bayes rules for classification (or decision making).*

Given a document ***d*** and a class ***c***, the formula to calculate the probability of class ***c*** in document ***d*** is called Bayer rule

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Bayer rule helps to compare the assumption probability with the observed (input) data so that it can make the final decision in classification.

3. *What is the conditional independence assumption of the Naïve Bayes learning*? => It is an assumption that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i|c_j)$.

Please explain the relationship between Naïve Bayes classifier and the Bayes Rule. => the Bayes rule is the based formula to calculate the estimation in Naïve Bayes classifier.

4. *What is "Information Gain"? please explain how to use "Information Gain" to construct a decision tree.* => Info Gain is a measurement that is used to measure the expected reduction in entropy given the value of some attribute. The question #2 is the example of how info gain help to figure out what attribute should be the root node of a tree and other leaf-node of a tree.

5. *What is the bias of the "Information Gain"? please show one solution to reduce the bias of the information gain*. => the bias of the Info gain is the problem of having maximized info gain value, attributes having a large number of values. One solution is using split information value.

6. *Please list the major steps of using binominal Naïve Bayes learning for text classification.*

⇨ The major step:

- Extract X_w from dictionary

- Check X_w = true in document d if w appears in dictionary.

  $\hat{P}(X_w = t \mid c_j) =$ fraction of documents of topic c_j in which word w appears
- Get P and repeat again.

Part II: Decision Tree and Naïve Bayes Classification [**9 pts**]

Question 2 [2.5 pts]. Given the following 5 instances each with four attributes (Outlook, Temperature, Humidity, and Wind) and one class label:

- Please manually construct a decision tree by using **Information Gain Ratio** as the attribute selection criteria (list the major steps of the tree constructions, and report the final decision tree) [2 pts]

ID	Outlook	Temperature	Humidity	Wind	Class
1	Sunny	Hot	High	Weak	No
2	Sunny	Cool	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Sunny	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes

The information gain of each attribute:

The entropy of the system is Entropy(S) = $-(2/5)x\log_2(2/5) - (3/5)x\log_2(3/5) = 0.971$

- Outlook(sunny, overcast, rain)
  - $S_{sunny} = \{1,2,4\}$ => $Entropy(S_{sunny}) = -(2/3)\log_2(2/3)-(1/3)\log_2(1/3) = 0.917$
  - $S_{overcast} = \{3\}$ => $Entropy(S_{overcast}) = -1\log_2 1=0$
  - $S_{rain} = \{5\}$ => $Entropy(S_{rain}) = -1\log_2 1=0$

  $=> (|S_{sunny}|/S)\, Entropy(S_{sunny}) + (|S_{overcast}|/S) Entropy(S_{overcast}) + (|S_{rain}|/S) Entropy(S_{rain})$

  $= (3/5)x0.917 + (1/5)x0 + (1/5)x0 = 0.5502$

  => **Gain(S,Outlook) =0.971-0.5502=0.4208**

- Temperature (hot, cool, mild)
  - $S_{hot} = \{1,3\}$ => $Entropy(S_{hot}) = -0.5\log_2 0.5-0.5\log_2 0.5=1$
  - $S_{cool} = \{2,5\}$ => $Entropy(S_{cool}) = -0.5\log_2 0.5-0.5\log_2 0.5=1$
  - $S_{mild} = \{4\}$ => $Entropy(S_{mild}) = -1\log_2 1=0$

  $=> (|S_{hot}|/S)\, Entropy(S_{hot}) + (|S_{mild}|/S) Entropy(S_{mild}) + (|S_{cool}|/S) Entropy(S_{cool})$

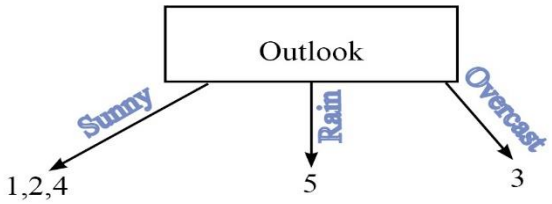
= (2/5)x1 +(2/5)x0 +(1/5)x1 = 3/5 =0.6

=>**Gain(S,Temperature)=0.971-0.6=0.371**

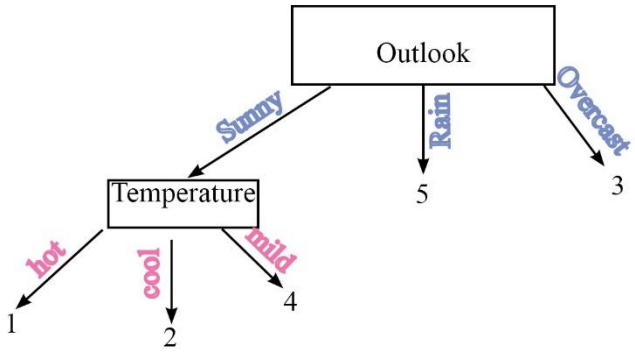
- Humidity (high, normal)
  =>**Gain(S,Humidity)=0.971-0.8=0.171**

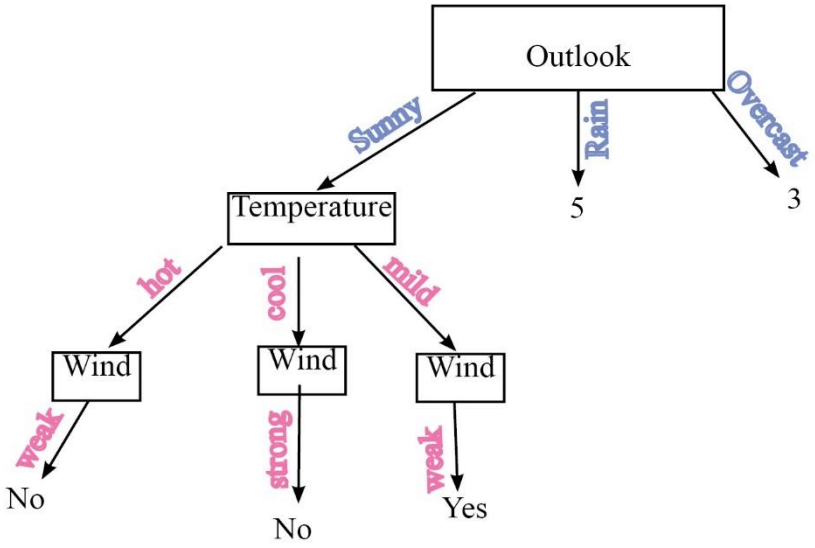
- Wind (weak, strong)
  =>**Gain(S,Wind)=0.971-0.649=0.322**

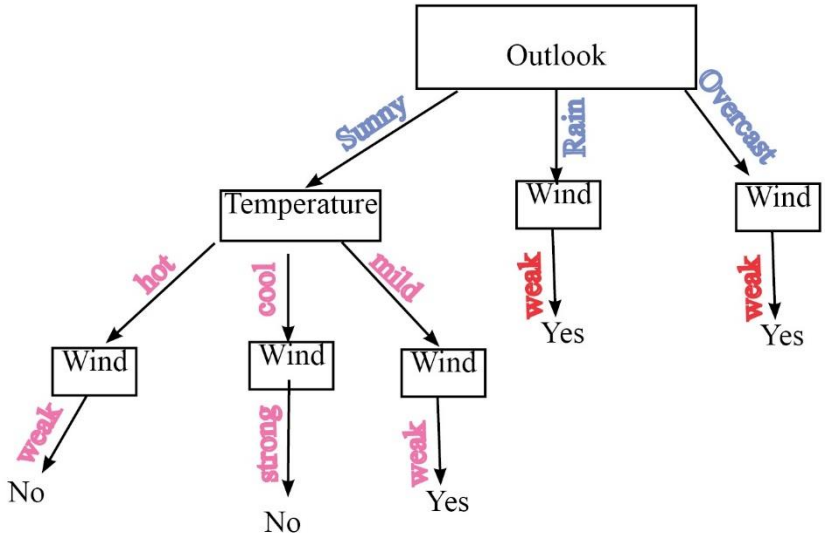
⇨ Choose Outlook is the root node since its info gain is highest (0.4208)



Among 1,2,4 choose Temperate attribute since its highest info gain







- *It is known that C4.5 uses Information Gain Ratio measure for decision tree construction, does C4.5 search all hypothesis to find the best tree (explain why or why not, 0.25 pt)?*

⇨ First of all, we have to know what information gain ratio is. The information gain ratio is the ratio between the information gain and the intrinsic value. In order to have info gain, the system has to calculate the entropy of the system, thus it searches all the nodes => it searches all hypothesis to find the best tree

- The inductive Bias of C4.5 is greedy

Question 3 [2.5 pts]: Given the following toy dataset with 15 Instances

- Please manually construct a Naïve Bayes Classifier (list the major steps, including the values of the priori probability [1.0 pt] and the conditional probabilities [1.0 pt]. Please use *m*-estimate to calculate the conditional probabilities (*m*=1, and *p* equals to 1 divided by the number of attribute values for each attribute).

ID	Outlook	Temperature	Humidity	Wind	Class
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Mild	Normal	Weak	No
14	Rain	Hot	High	Strong	Yes
15	Rain	Mild	High	Strong	No

Priori probability:

P(n) = 6/15
P(y) = 9/15

outlook	
P(sunny|y) = 2/9	**P(sunny|n) = 3/6**

P(overcast|y) = 3/9	**P(overcast|n) = 1/6**
P(rain|y) = 4/9	**P(rain|n) = 2/6**
temperature	
P(hot|y) = 2/9	**P(hot|n) = 2/6**
P(mild|y) = 4/9	**P(mild|n) = 3/6**
P(cool|y) = 3/9	**P(cool|n) = 1/6**
humidity	
P(high|y) = 4/9	**P(high|n) = 4/6**
P(normal|y) = 5/9	**P(normal|n) = 2/6**
windy	
P(strong|y) = 4/9	**P(strong|n) = 3/6**
P(weak|y) = 5/9	**P(weak|n) = 3/6**

- *Please use your Naïve Bayes classifier to determine whether a person should play tennis or not, under conditions that "Outlook=Overcast & Temperature=Hot & Humidity =Normal& Wind=Weak". [0.5 pt]*

$$p(y)\,p(overcast \mid y)\,p(hot \mid y)\,p(normal \mid y)\,p(weak \mid y) =$$
$$= (9/15) * (3/9) * (2/9) * (5/9) * (5/9) = 0.00152$$
$$p(n)\,p(overcast \mid n)\,p(hot \mid n)\,p(normal \mid n)\,p(weak \mid n) =$$
$$= (6/15) * (1/6) * (2/6) * (2/6) * (3/6) = 0.0037$$

⇨ Likelihood of No > Likelihood of Yes =>should not play tennis.

In case m=1 (*m*=1, and *p* equals to 1 divided by the number of attribute values for each attribute).

- ⇨ Use this formula:

$$\hat{P}(x_{i,k} \mid c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m} = \frac{N(X_i = x_{i,k}, C = c_j) + 1/(15)}{N(C = c_j) + 1}$$

outlook	
P(sunny|y) = [(2+1/15)/(9+1)]=31/150	**P(sunny|n) = 46/105**
P(overcast|y) = 31/150	**P(overcast|n) = 16/105**
P(rain|y) = 61/150	**P(rain|n) = 31/105**
temperature	
P(hot|y) = 31/150	**P(hot|n) = 31/105**
P(mild|y) = 61/150	**P(mild|n) = 46/105**
P(cool|y) = 31/150	**P(cool|n) = 16/105**
humidity	
P(high|y) = 61/150	**P(high|n) = 61/105**
P(normal|y) = 76/150	**P(normal|n) = 31/105**
windy	
P(strong|y) = 61/150	**P(strong|n) = 46/105**
P(weak|y) = 76/150	**P(weak|n) = 46/105**

Question 4 [2 pts]: A patient takes a lab test and the result comes back positive. Assume the test returns a correct positive result in only 95% of the cases in which the disease is actually present, and a correct negative result in only 95% of the cases in which the disease is not present. Assume further that 0.001 of the entire population have this cancer. Please use Bayes Rule to derive the probability of the patient having the cancer given that his/her lab test is positive (list the major steps). [2 pt]

$$P(cancer) = .001, P(\neg cancer) = .999$$
$$P(+ \mid cancer) = .95, P(- \mid cancer) = .05$$
$$P(+ \mid \neg cancer) = .05, P(- \mid \neg cancer) = .95$$
$$P(cancer \mid +) = \frac{P(+ \mid cancer)P(cancer)}{P(+)} = \frac{0.00095}{P(+)}$$
$$P(\neg cancer \mid +) = \frac{P(+ \mid \neg cancer)P(\neg cancer)}{P(+)} = \frac{0.04995}{P(+)}$$

⇨ $P(\neg cancer \mid +) > P(cancer \mid +)$

The chance not having cancer of that patient is 53 times larger than the chance of having cancer.

Question 5 [2 pts]: *Researchers want to study whether there is a dependency between gender and the voting preference (i.e., whether gender and the voting preference is independent of each other or not), so they randomly collect public opinion poll from 1000 individuals/voters (among them 400 are male and 600 are female). For all 1000 voters, 450 of them are Republican, 450 of them are Democrat, and 100 of them are Independent. The observed numbers for each group is detailed in the following table, please use Chi-Square Test (χ^2) to explain whether there is a dependence between gender and the voting preference.*

- *Please calculate the expected numbers of Male-Republican, Male-Democrat, Male-Intendent, Female-Republican, Female Democrat, Female-Independent [1 pt]*
- *Please calculate the Chi-Square value, and the corresponding p-value [0.5 pt]*
- *Please explain whether there is a dependence between gender and the voting preference [0.5 pt]*

		Voting Preference			
		Republican	Democrat	Independent	Total
Voters	Male Voters	200 180	150 180	50 40	400 (25%)
	Female Voters	250 270	300 270	50 60	600 (75%)

	Total	450	450	100	1000

The expected male voters who favor Republican is (450x400)/1000= 180

The expected male voters who favor Democrat is (450x400)/1000= 180
Similarly we have the expected values of other elements in purple color.

$$\chi^2(j,a) = \sum (O-E)^2/E = (200-180)^2/180 + (150-180)^2/180 + (50-40)^2/40 +$$
$$+ (250-270)^2/270 + (300-270)^2/270 + (50-60)^2/60 = 16.203$$

- The null hypothesis is *there is an independence between gender and the voting preference.*

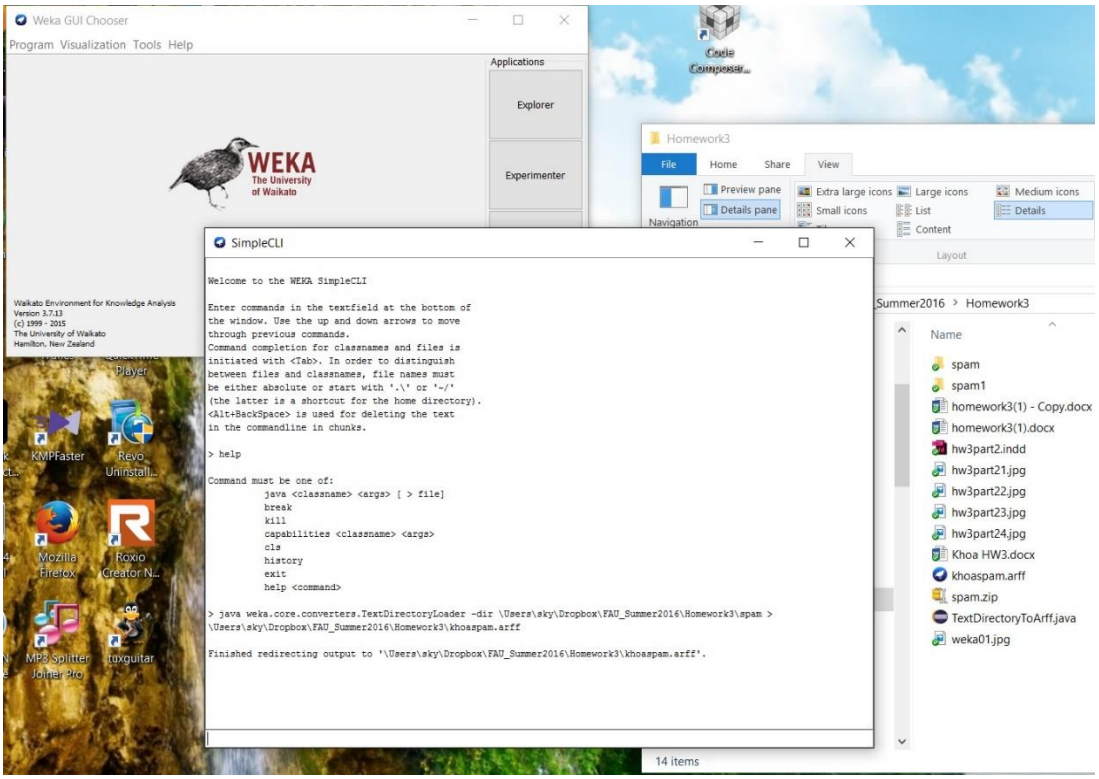
DF=(# rows-1)*(#columns-1) = 2

⇨ Looking at the table of CHI-square, with Df=2 and the threshold p = 0.05 => critical value is 5.99 <X^2 = 16.203. => reject the null hypothesis => the gender and the voting preference are dependent

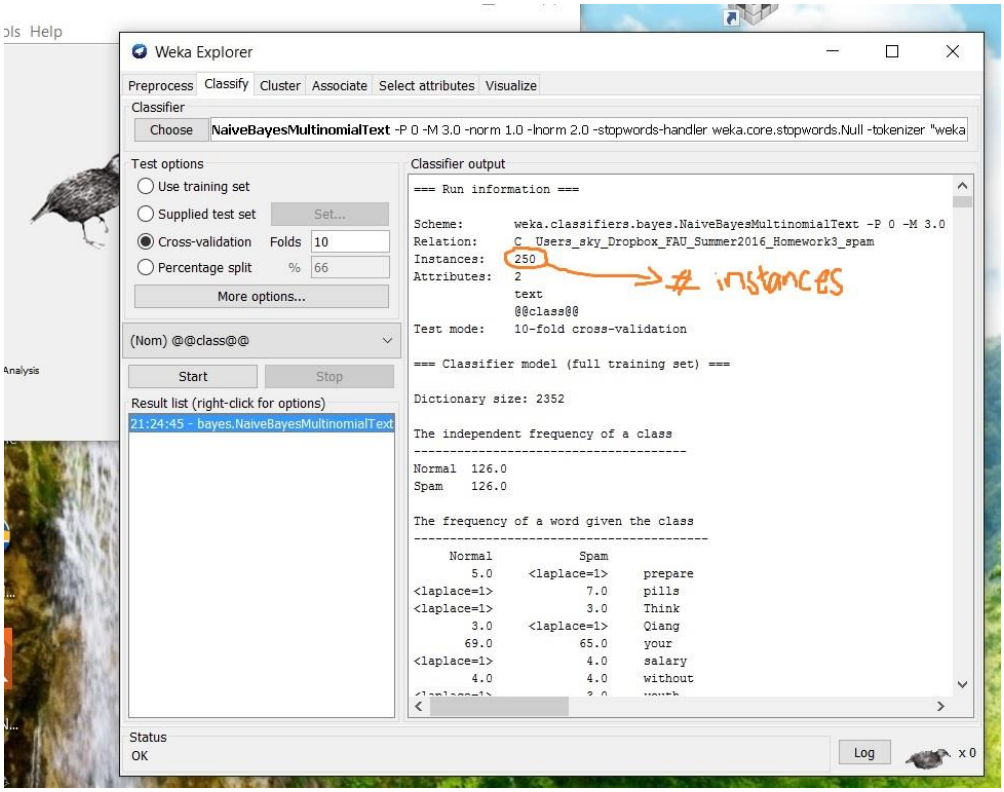
Part III: Hands-on Spam Filtering Practice [**3 pts**]
Report:

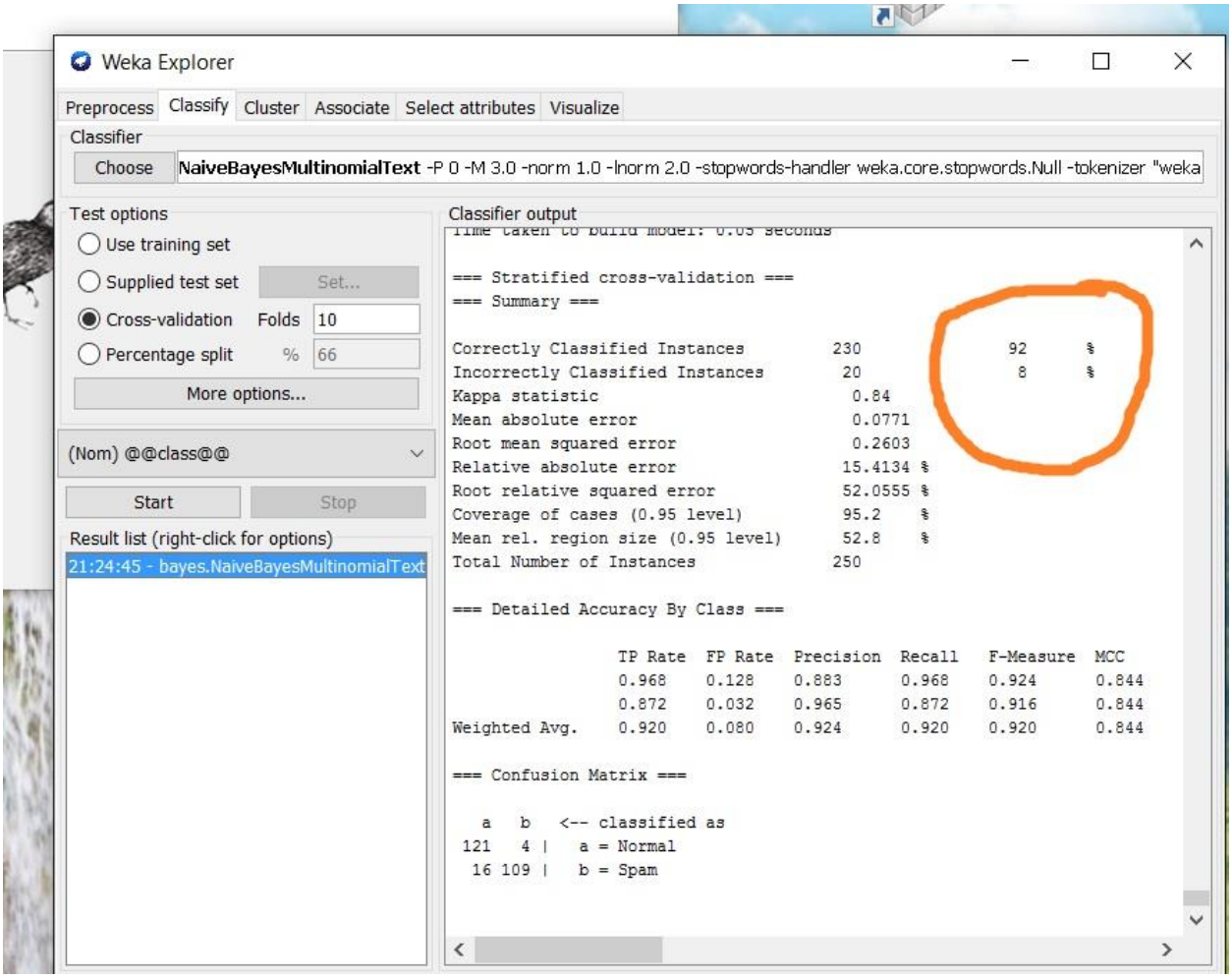
1. *Please download and install WEKA (http://www.cs.waikato.ac.nz/ml/weka/), and show a screenshot that WEKA is running on your computer. Please also submit the converted arff file [1 pt].*

⇨ See the attachment file ***khoaspam.arff*** for the converted file.



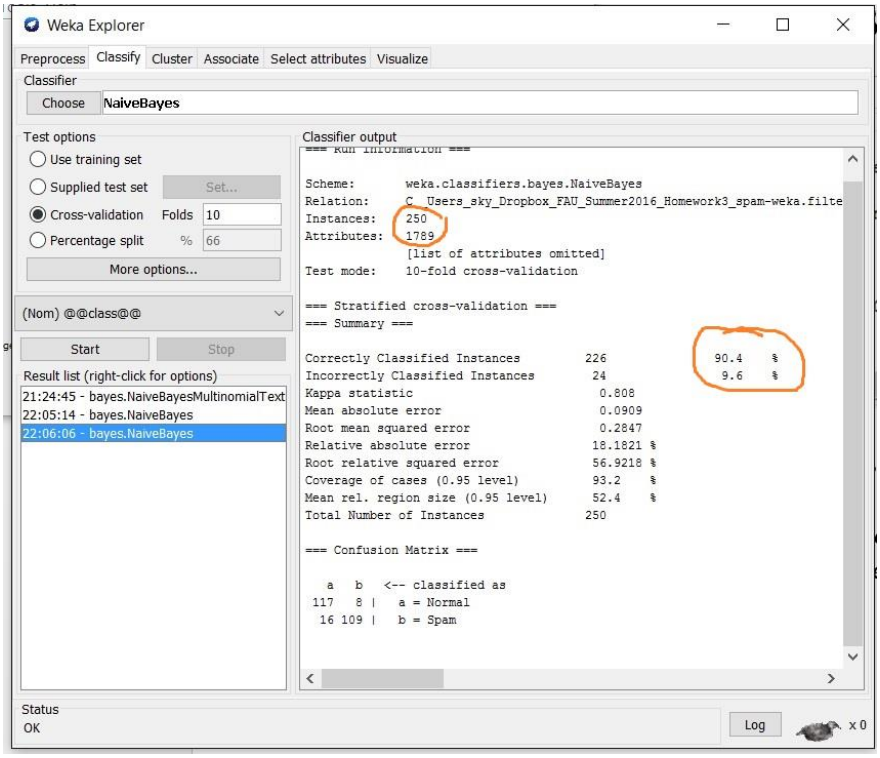
2. *Please report the classification accuracy of your Naïve Bayes classifier, using 10-fold cross validation [capture a screenshot of the WEKA results]* [1 pt]

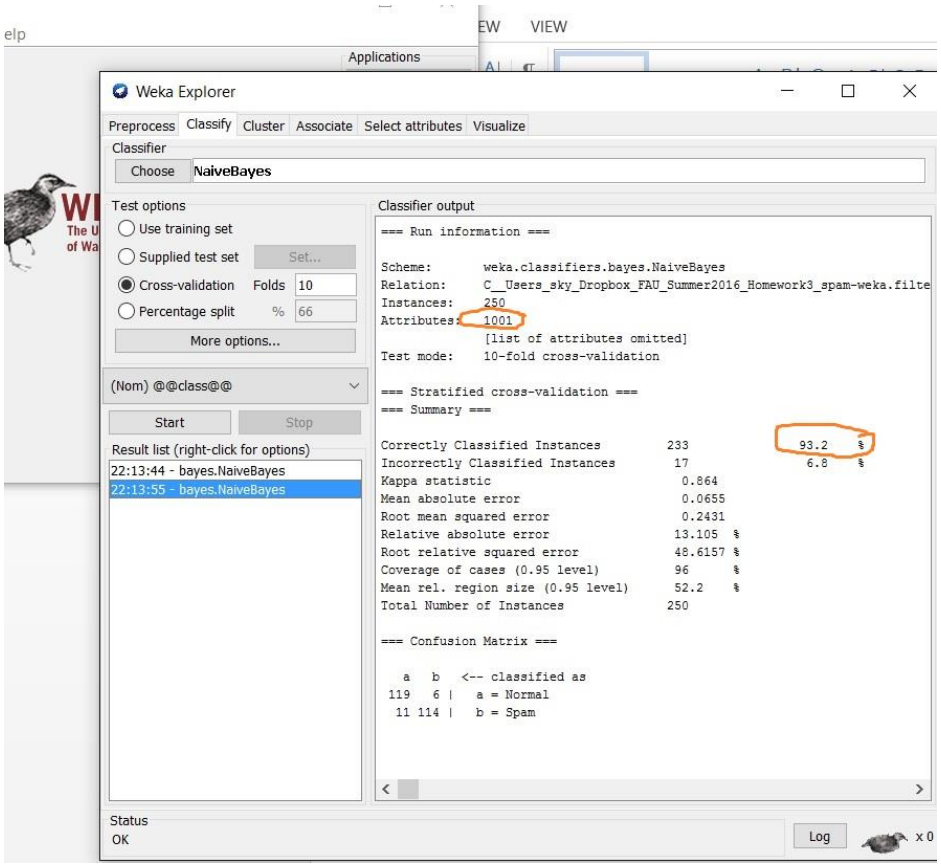




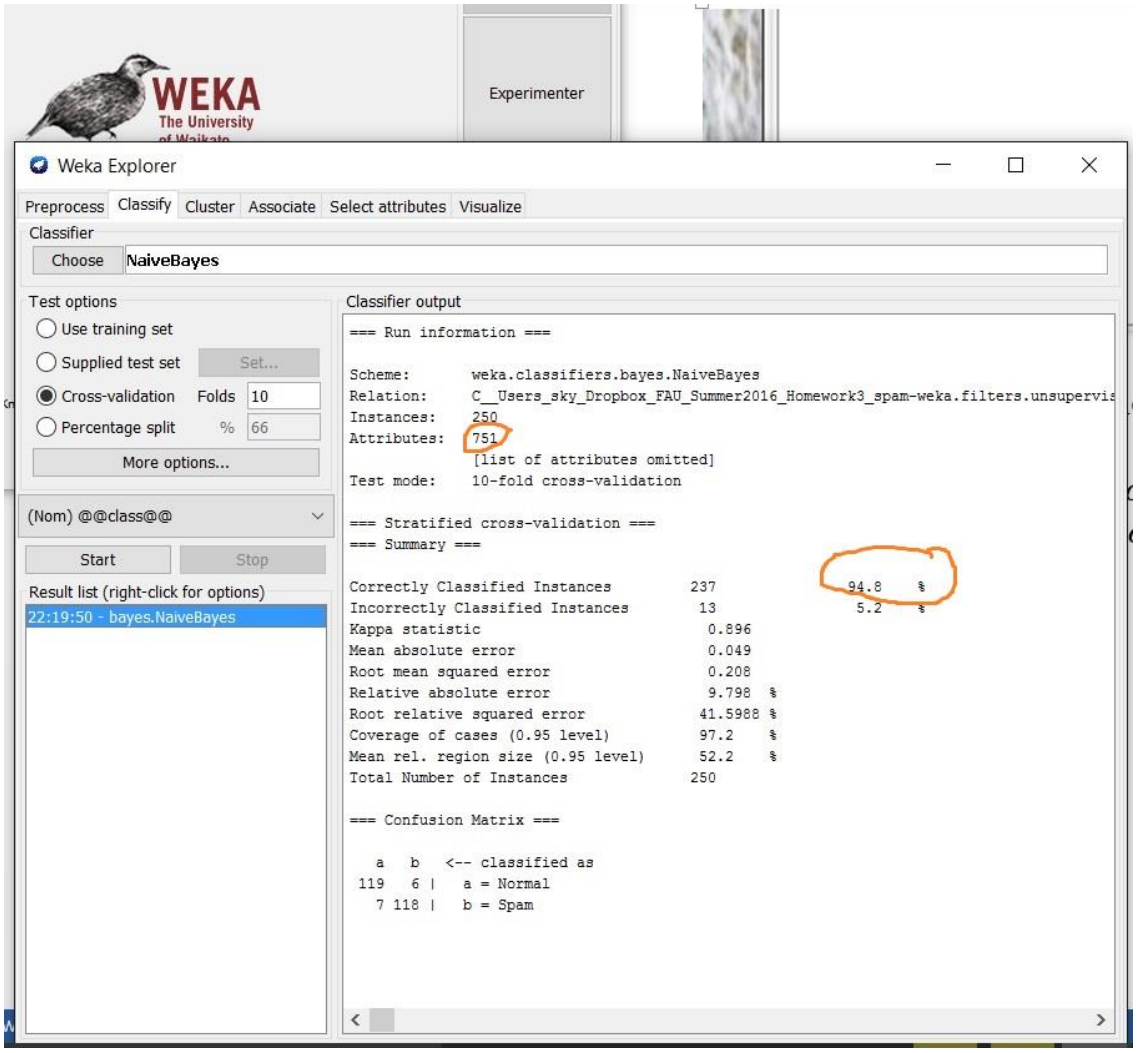
⇨ The correctly classified instances is 92%.

3. *Please use Information Gain to select 1000, 750, 500, 250, 100 features/keywords, respectively, and report the Naïve Bayes classification results for each classifiers. [1 pt]*





When choose 1000 features (attributes), the accurate of classification is 93.2%



- When choose 750 features => The accurate of classification is 94.8%

- 500 features => the accurate is 95.6%

- 250 features => the accurate is 94.4%

- 100 features => the accurate is 94.4%

=> With the number of feature/attribute 500, the accurate of the system is the highest.