CSL-554
Text and Web Intelligence Analytics

ASSIGNMENT-2

Submission Instructions:

1. **The deadline to submit assignment-2 is 15 September 2018 by 1:00 pm (no late submissions will be allowed).**
2. **The weightage for assignment-2 is 20% in assignment/ internal assessment component.**
3. **Each question in Assignment-2 is of 20 marks.**

Question 1: Imagine that you are given the following set of training examples. Each feature can take on one of three nominal values: a, b, or c

F1	F2	F3	Category
a	c	a	+
c	a	c	+
a	a	c	−
b	c	a	−
c	c	b	−

a.) How would a Naive Bayes system classify the following test example? (Be sure to show your working.)

F1 = a, F2 = c , F3 = b

Solution:

Using Laplace correction:

$$P(F1 = a|Class = +)P(F2 = c|Class = +)P(F3 = b|Class = +)P(Class = +) =$$

2/4 * 2/4 * 1/4 = 1/16.

$$P(F1 = a|Class = -)P(F2 = c|Class = -)P(F3 = b|Class = -)P(Class = -) =$$

2/5 * 3/5 * 1/5= 6/125.

P(+| instance) = 1/16 * 2/5 = 0.025
P(-| instance) = 6/125 * 3/5 = 0.028

Therefore, the instance will be classified as - .

b.) Describe how a 3-nearest-neighbor algorithm would classify the test example given above. Use hamming distance.

Solution:

Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different.

Hamming distance of given instance to other points is 1,2,2,1,3.

Hence, 3 Nearest neighbours are either instances 1,2 and 4 or 1,3 and 4.

Thus, majority class will be:

+ if instances 1,2 and 4 are considered as nearest neighbours else -.

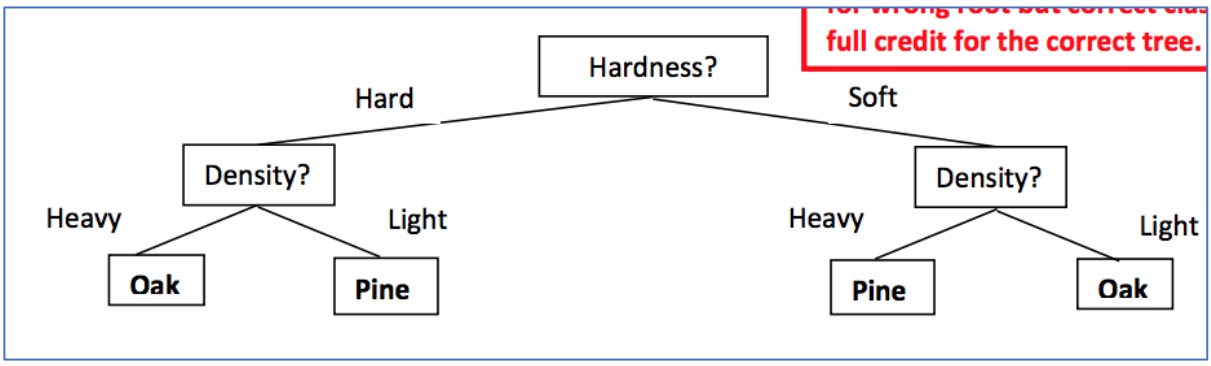
Question 2: You are a robot in a lumber yard, and must learn to discriminate Oak wood from Pine wood. You choose to learn a Decision Tree classifier. You are given the following examples.

Example	Density	Grain	Hardness	Class
Example #1	Heavy	Small	Hard	Oak
Example #2	Heavy	Large	Hard	Oak
Example #3	Heavy	Small	Hard	Oak
Example #4	Light	Large	Soft	Oak
Example #5	Light	Large	Hard	Pine
Example #6	Heavy	Small	Soft	Pine
Example #7	Heavy	Large	Soft	Pine
Example #8	Heavy	Small	Soft	Pine

Draw the decision tree that would be constructed by recursively applying information gain to select roots of sub-trees. Classify these new examples as Oak or Pine using your decision tree.

a.) What class is [Density=Light, Grain=Small, Hardness=Hard]?
b.) What class is [Density=Light, Grain=Small, Hardness=Soft]?

Solution:



What class is [Density=Light, Grain=Small, Hardness=Hard]? Pine
What class is [Density=Light, Grain=Small, Hardness=Soft]? Oak

Question 3: Assume a biword index. Give an example of a document which will be returned for a query of New York University but is actually a false positive which should not be returned.

Solution:

New York City has Old York University
Located in Manhattan

Question 4: Imagine that you are given the following set of training examples.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

a.) How would a Naive Bayes system classify the following test example? (Be sure to show your work.)

age <=30, Income = medium, Student = yes, Credit_rating = Fair

Solution: The test instance will be classified as Buys_Computer = Yes.

Question 5: Imagine that you are given the following set of data points. Using k-means clustering algorithm, partition the data points into 3 clusters (Be sure to show your working).

Solution:

Cluster	Data Point
A	(2, –1)
A	(–1, 2)
A	(–2, 1)
A	(1, 2)
B	(4, 0)
B	(4, –1)
B	(0, –2)
B	(0, –5)
C	(–1, 0)
C	(3, 8)
C	(–2, 0)
C	(0, 0)