

National University of Computer & Emerging Sciences FAST-Karachi Campus CS4051- Information Retrieval Quiz#5 (Double Weightage)

Dated: May 09, 2024 Marks: 20

Time: 20 min. Question No. 1

Consider the following examples for the task of text classification

Dataset	DocID	Features- Words in documents	Class
			Fruit=Yes/No
Training set	1	Orange, Orange, Lemon	No
	2	Orange, Red, Blue	No
	3	Apricot, Apple, Mango	Yes
	4	Apple, Banana, Orange	Yes
	5	Apple, Orange, Melon	Yes
Test set	6	Orange, Mango, Melon	?
	7	Orange, Red, Lemon	?

a. Using the training data calculate the class prior probabilities?

$$P(Fruit=Yes) = 3/5 = 0.6$$

 $P(Fruit=No) = 2/5 = 0.4$

b. Using Multinomial Naïve Bayes to estimate the probabilities of each term (feature) that you use to classify the given text cases.

P(Orange/Fruit)	1/6	P(Orange/~Fruit)	4/15
P(Mango/Fruit)	1/9	P(Mango/~Fruit)	1/15
P(Melon/Fruit)	1/9	P(Melon/~Fruit)	1/15
P(Red/Fruit)	1/18	P(Red/~Fruit)	2/15
P(Lemon/Fruit)	1/18	P(Lemon/~Fruit)	2/15

c. Predict the class labels for the two instances in test set?

P(d6/Fruit)	0.6 * (1/6) * (1/9) * (1/9) = 0.001
P(d6/~Fruit)	0.4 * (4/15)* (1/15)* (1/15) = 0.0004

Document d6 belongs to class Fruit=Yes.

P(d7/Fruit)	0.6 * (1/6) * (1/18) * (1/18) = 0.0003
P(d7/~Fruit)	0.4 * (4/15)* (2/15)* (2/15) = 0.001

Document d7 belongs to class Fruit=No.

Question No. 2

Consider the following data points in 2D space.

P1(3,1), P2(1,2), P3(1,1), P4(8,1), P5(7,3), P6(6,5) taking P2 and P5 as initial seeds – run vanilla k-means algorithm for two iterations to find the resultant clusters C1 and C2.

First Iteration:

$$C1 = P2$$

$$C2 = P5$$

P1 is close to P2 than P5, similarly P3 is close to P2 than P5 P4 is close to P5 than P2. Hence

First Iteration:

$$C2=(7,3)$$

P1 is close to C1 than C2, similarly P3 is close to C1 than C2

P4 is close to P5 than P2. Hence

$$C1=\{ P1, P2, P3 \} = \{5/3, 4/3 \}$$

 $C2=\{ P4, P5, P6 \} = (7, 3)$