3333333333333333333333333333333333333333333333333333333

National University of Computer & Emerging Sciences
FAST-Karachi Campus
Information Retrieval (CS317)
Quiz#3

Dated: May 03, 2018 Marks: 30
Time: 30 min.
Std-ID: ____SOL__________

Question NO. 1

Answer the following questions with a brief description about the concept associated with the question.

1. Define the term concept drift?

   Concept drift – can be defined as a gradual change over time of the concept underlying a class. Example: <cell Phone> class do have a lot of changing features, working and functionalities over time.

2. Explain how the independent assumption in Naïve Bayes is not true for text. Give an example to support your answer.

   This is very true that independence assumption, is overly simplified and far from reality for natural language text. The conditional independence assumption states that features are independent of each other given the class, but in NLP terms pair are often more dependent than others Like Hong Kong, while estimating the class probabilities for a given instance even though the probability estimates of NB are of low quality, its classification decisions are surprisingly good. Correct estimation implies accurate prediction, but accurate prediction does not imply correct estimation. NB classifiers estimate badly, but often classify well.

3. What is clustering hypothesis?

   Cluster Hypothesis - In information retrieval, it states that documents that are clustered together "behave similarly with respect to relevance to information needs"- If a document (d) is relevant to a query, the documents that falls in the same cluster as d, will also be relevant to the query.

4. How mutual information can be regarded as feature selection? Explain.

   Mutual Information (MI) measures how much information the presence/absence of a term contributes to making the correct classification decision on class C, mathematically I(U,C), where U is a feature and C is the class. This can obviously be a feature selection criterion for classification task.

Question NO. 2

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 test 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. 3

a. Consider the documents

D1: He moved from London, Ontario, to London, England.

D2: He moved from London, England, to London, Ontario.

D3: He moved from England to London, Ontario.

Which of the documents given above have identical and different bag of words representations for (i) the Bernoulli model (ii) the multinomial model? If there are differences, describe them.

(i) For the Bernoulli model, the 3 documents are identical.

(ii) For the multinomial model, documents 1 and 2 are identical and they are different from document 3, because the term London occurs twice in documents 1 and 2, but occurs once in document 3.

b. What are the some of the drawbacks of K-Mean clustering?

1) It requires number of clusters as an input (k), sometime it is not easy to guess this number for a dataset.
2) It is very sensitive for initial seeds.
3) Outliers may unnecessarily increase conversion for it.