PS-6 Knowledge Discovery & Management

Name Rohith kumar N

CID: 16

① Document Models

Text classifiers often don't use any Kind of deep representation about language : often a document is represented as a bag of words consider a document D, whose class is given by C. In the case of email spam filtering ' there are 2 classes C = S(spam) and c = H(ham) we classify D as the class which has the highest posterior probability $P(c/D)$ which can be represented using Baye's theorem.

$$P(C/_D) = \frac{P(D/_c)\,P(c)}{P(D)} \alpha\ P(D/_c)\,P(c)$$

There are two probabilistic models of documents. Both of which represent documents as a bag of words using the Naive Bayes assumption. Both models represent documents using feature vectors whose components correspond to word types. If we have a vocabulary V, containing NI word types then the feature vector dimension $d = |V|$

Bernouli document model

A document is represented by a feature vector with binary elements taking value 1 if the Corresponding word is present in the document and 0 if the word is not present.

Multinomial document Model

A document is represented by a feature vector with integer elements whose value is the frequency of that word in the document

1a) Bernoulli. Naive Bayes model

The likelihood of a document given a class C_x is given by

$$P(x/C_k) = \prod_{i=1}^{n} P_{Ki}^{x_i} (1-P_{Ki})^{(1-x_i)}$$

So $P(x_{transisco} = true / class = SFo)$ which indicates whether fransisco appears in the document Class SFO. If it appears probability is 1 else 0.

$$P(x_{fransisco} = true / class = SFo) = 1.0$$

$$P(x_{London} = true / class = SFo) = 0.5$$

$$P(x_{fransisco} = true / class = JFk) = 1.0$$

1b) Multinomial Naive Bayes Model

$$p(x = fransisco / class = SFo) = 4/14$$

$$p(x = london / class = SFo) = 1/14$$

$$p(x = fransisco / class = JFk) = 1/8$$

1c)

i) When the Bernoulli Naive Bayes model is applied to the test set after trained on the training set it is not very accurate, because it ignores frequency information, which is important in this domain.

ii) The multinomial model is more accurate, because it uses frequency information. Hawever it ignores position information so doesn't distinguish between a cityname occuring at the beginning/end of the itemary from the one which is occuring in the middle of the model.

1d) We can use as a feature the term that occurs in the last position of each document. Non standard feature represented using non-standard words.

2)

a) It will never choose a category unless all words in a document were seen for that Category for the training set unless there is no category for which all the words were seen and then all categories are tied for the classified. It will rank between classes for whichall words we seen similarly to the smoothed classifier

b) Here it is given that they have doubted the amount of smoothing.

formula for Laplace (add-1) smoothing for Naive Bayes

$$P(w_i|c) = \frac{count(w_i,c)+1}{\sum(count(w,c)+1)} = \frac{count(w,c)+1}{\sum_{w\in V} count(w,c)+|V|}$$

It will be more likely to choose categories for which so many of the words in the document were unseen.

(III) Given that

System returns 3 relevant documents

2 irrelevant documents.

Total 8 relevant documents in the collection.

	Retrieved	Not Retrieved	
Relevant	3 TP	5 FN	8
Non Relevant	2 FP	3 TN	5
	5	8	13

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{3}{3+2} = \frac{3}{5}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+5} = \frac{3}{8}$$

b)

i) An IR System which always returns no results will have high accuracy for most queries, since the corpus usually contains only a few relevant documents. Documents that are truly relevant are the only ones that will be mistakenly classified as non relevant and thus the accuracy is close too1.

Recall and precision are two different measures that can jointly capture the tradeoff between returning more relevant results and returning fewer irrelevant results.

~~Eg:~~

ii) There are ~~of~~ ocurse many correct answers one simple correct answer is

Assume document 1 is the only relevant document

$A_q = \{1,2,3\}$

$B_q = \{3\}$

Both A_q and B_q made 2 mistakes, so they have the same accuracy 80%.

The precision of A_q is $1/3$, the precision for B_q is 0 since B_q didn't return any relevant documents. It is of no utility.