CS5560 Knowledge Discovery and Management

Problem Set 6

July 10 (T), 2017

Name: SreeLakshmi Nandanamudi
Class ID: 17

References
https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/
https://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html
http://www.nltk.org/book/ch06.html

I. Consider the problem of classifying the origination point of passenger travel itineraries. Suppose we have the following training set of travel itineraries:

Itinerary	Document	Class
1	"smith: new york - chicago - san francisco - new york"	JFK
2	"chen: san francisco - london - paris - san francisco"	SFO
3	"chen: san francisco - tokyo - singapore- san francisco"	SFO
4	"o'brien: chicago - buenos aires - new york - chicago"	ORD

a) Assume that we use a Bernoulli (i.e., binary) Naive Bayes model. Compute the following feature probabilities:

- P(Xfrancisco=true | Class=SFO)
- P(Xlondon=true | Class=SFO)
- P(Xfrancisco=true | Class=JFK)

b) Assume that we use a multinomial NB model instead. Compute the following probabilities:

- P(X=francisco | Class=SFO)
- P(X=london | Class=SFO)
- P(X=francisco | Class=JFK)

c) Consider a standard Naive Bayes classifier trained on the training set and applied to a similar test set. How accurate is this classifier for:

(i) the Bernoulli model, and

(ii) the multinomial model?

d) Construct a non-standard feature representation that is 100% accurate for either model.

II. This problem concerns smoothing Naïve Bayes classifiers. Consider the following formula for Laplace (add-1) smoothing for Naïve Bayes

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

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

a) Suppose we build a Naive Bayes classifier (multinomial or Bernoulli) with no smoothing of the respective P(word | class) probabilities. If a word was unseen in a class, it will thus have a probability of 0. Describe in words the decision procedure of this classifier (emphasizing the effect of the lack of smoothing, and how its decisions will differ from a smoothed Naive Bayes classifier).

b) Suppose we take a smoothed multinomial classifier and double the amount of smoothing (e.g., for a variant of "add 1 smoothing", add 2 to each count, and add to the denominator 2k, where k is the number of samples). What qualitative effect will this have on decisions of the classifier?

III. An IR system returns 3 relevant documents, and 2 irrelevant documents. There are a total of 8 relevant documents in the collection.

a) What is the precision of the system on this search, and what is its recall?

b) Instead of using recall/precision for evaluating IR systems, we could use accuracy of classification. Consider a classifier that classifies documents as being either relevant or non-relevant. The accuracy of a classifier that makes c correct decisions and i incorrect decisions is defined as: c/(c+i).

(i) Why do the recall and precision measures reflect the utility (i.e., quality or usefulness) of an IR system better than accuracy does?

(ii) Suppose that we have a collection of 10 documents, and two different boolean retrieval systems A and B. Give an example of two result sets, Aq and Bq, assumed to have been returned by the system in response to a query q, constructed such that Aq has clearly higher utility and a better score for precision than Bq, but such that Aq and Bq have the same scores on accuracy.

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 two 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 re-expressed using Bayes' Theorem:

$$P(C/D) = \frac{P(D/C)P(C)}{P(D)} \propto P(D/C)P(C)$$

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 $|V|$ word types, then the feature vector dimension $d = |V|$

Bernoulli 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.

a) Assume that we use a Bernoulli (i.e., binary) Naive Bayes model. The following are the feature probabilities :

$$P(X_{francisco} = true \mid class = SFO) = 1.0$$

$$P(X_{london} = true \mid class = SFO) = 0.5$$

$$P(X_{francisco} = true \mid class = JFK) = 1.0$$

b) Assume that we use a multinomial NB model instead. the following are the probabilities :

$P(X = francisco \mid class = SFO) = 4/14$ (assuming no tokenization of punctuation)

$$P(X = London \mid class = SFO) = 1/14$$

$$P(X = francisco \mid class = JFK) = 1/8$$

c) i, The Bernoulli model :

the Bernoulli model is not accurate, because it ignores frequency information, which is important in this domain.

ii, The Multinomial model :

The Multinomial model is more accurate, because it uses frequency information. However it ignores position information, so doesnt distinguish between a city name occuring at the beginning /end of the itinerary from one occurring in the middle.

d) <u>Non-standard feature representation</u> :

Use as a feature the term that occurs in the last position of each document; so that it will be 100% accurate for either model.

$$P\left(X_{\text{Francisco}} = \text{true} \mid \text{Class} = \text{SFO}\right) = \frac{1.0}{\cancel{2}} = \cancel{0.5} = 1.0$$

$$P\left(X_{\text{New york}} = \text{true} \mid \text{class} = \text{JFK}\right) = 1.0$$

$$P\left(X_{\substack{\cancel{\text{francisco}}\\ \text{Chicago}}} = \text{true} \mid \text{class} = \text{ORD}\right) = 1.0$$

<u>II</u>. a) In Naive Bayes classifiers, 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 words were seen and then all categories are tied for the classifier). It will rank between classes for which all words were seen similarly to the smoother classifier (but with possible differences due to the smoothing).

b) In smoothed multinomial classifier, it will be more likely to choose categories for which some/many of the words in the document were unseen.

<u>Laplace</u> <u>smoothing</u> : The Laplace's estimate :

pretend you saw every outcome once more than you actually did :

$$P_{LAP}(x) = \frac{c(x)+1}{\sum_x [c(x)+1]} = \frac{c(x)+1}{N+|X|}$$

Ⓗ Ⓗ Ⓣ

$P_{ML}(X) =$

$P_{LAP}(X) =$

- Can derive this as a MAP estimate with Dirichlet priors.

III.

Given an IR system returns 3 relevant documents, and 2 irrelevant documents and there are a total of 8 relevant documents in the collection.

a) The precision is given by $tp/(tp+fp) = 3/5$

The recall is given by $tp/(tp+fn) = 3/8$

→ In pattern recognition, information retrieval and binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over total relevant instances in the image. Both precision and recall are therefore based on an understanding and measure of relevance.

b) The accuracy of a classifier that makes c correct decisions and i incorrect decisions is defined as : $c/(c+i)$

i, An IR system which always returns no results will have high 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 as non relevant, and thus the accuracy is close to 1. Recall and precision are two different measures that can jointly capture the tradeoff between returning more relevant results and returning fewer irrelevant results.

ii) Given we have a collection of 10 documents and two different boolean retrieval systems A and B. Given an example of two result sets A_q and B_q and A_q has clearly highly utility and a better score for precision than B_q but both have same scores on accuracy.

→ There are of course 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 $A_q = \frac{1}{3}$

The precision fol $B_q = 0$

Since B_q didn't return any relevant documents, it is of no utility.