CS5560 Knowledge Discovery and Management

Problem Set 6

July 10 (T), 2017

Name: Rashmi

Class ID: 29

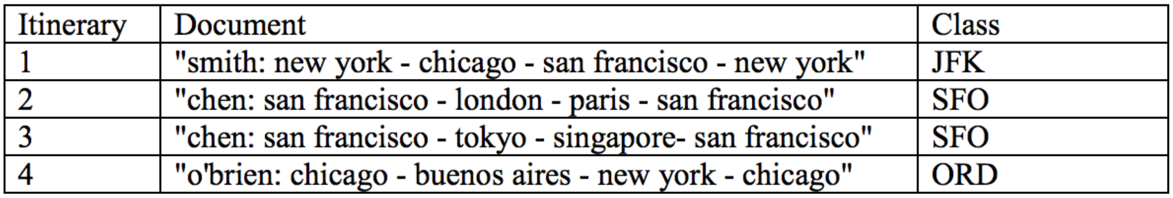
References

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 document is represented as a bag of words. (A bag is like a set that allows repeating elements.) This is an extremely simple representation: it only knows which words are included in the document (and how many times each word occurs), and throws away the word order!

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

An alternative to the multinomial model is the multivariate Bernoulli model or Bernoulli model. It is equivalent to the binary independence model, which generates an indicator for each term of the vocabulary, either 1 indicating presence of the term in the document or 0 indicating absence. The Bernoulli model has the same time complexity as the multinomial model.

- P(Xfrancisco=true | Class=SFO) = 1
- P(Xlondon=true | Class=SFO) =1/2=0.5
- P(Xfrancisco=true | Class=JFK) =1
-

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

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

- P(X=francisco | Class=SFO) =4/14 (no punctuation considered)
- P(X=london | Class=SFO) = 1/14
- P(X=francisco | Class=JFK) =1/8

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

As it ignores frequency of words in documents it is not that perfect.

(ii) the multinomial model?

In comparison it is more accurate as it considers frequency of words but it also does not consider position of words at the same time which is its biggest drawback.

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

We can consider the frequency and even the position of the word. If the word appears before we will give less weightage and if it appears in the last we will give more weightage.

This can be accurate as we are considering both position and frequency.

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

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

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

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).

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 smoothed classifier (but with possible differences due to the smoothing).

b) 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 smoothed classifier (but with possible differences due to the smoothing). 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?

It'll be more likely to choose categories for which some/many of the words in the document were unseen

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?

Precision =TP/TP+FP=3/5
Recall=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) 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?

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 nonrelevant, 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 and returning fewer irrelevant results.

(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.

There are of course many correct answers. One simple correct answer is Assume document 1 is the only relevant document.

Aq = {1,2,3} Bq = {3}

Both Aq and Bq made 2 mistakes, so they have the same accuracy: 80%
The precision of Aq is 1/3, the precision for Bq is 0. Since Bq didn’t return any relevant documents, it is of no utility.