CS5560 Knowledge Discovery and Management

Problem Set 6

July 10 (T), 2017

Name: Syed Moin

Class ID: 28

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

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

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

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.

1A)

a) using Bernoulli's Naive Bayes Model :

i) $P(X_{San\,Francisco} = true \mid class = SFO) = 1.0$

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

iii) $P(X_{San\,Francisco} = true \mid class = JFK) = 1.0$

b) using Multinomial NB model :

i) $P(X = San\,Francisco \mid class = SFO) = \frac{4}{14}$

ii) $P(X = London \mid class = SFO) = \frac{1}{14}$

iii) $P(X = San\,Francisco \mid class = JFK) = \frac{1}{8}$

c) i) For Bernoulli model, its not very accurate, because it ignores frequency information, which is important in this domain.

ii) For Multinomial model, More accurate, because it uses frequency information. However, it ignore position information, so doesn't distinguish between a city name occuring at the beginning/end of the itineary from one occuring in the middle.

d) use as a feature the term occurs in the last position of each document.

II

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 words were seen, and then all categories are tried 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'll be more likely to choose categories for which some/many of the words in the document were unseen.

III given, Relevant document = 3 } 5 retrieved.
irrelevant documents = 2

Total Relevant documents = 8.

i) Precision $= \frac{TP}{TP+FP} = \frac{3}{3+2} = \frac{3}{5}$.

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

b) i) An IR system which always returns no result 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.

ii) There is only 1 document which is relevant

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

$B_q = \{3\}$

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.