CS5560 Knowledge Discovery and Management

Problem Set 6
July 10 (T), 2017

Name: Yalamanchili Soumya
Class ID: 30

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

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

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

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.

Name: Yalamanchili Sowmya
ClassID: 30

(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 re-expressed using Bayes Theorem:

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

There are 2 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) Bernoulli Naive Bayes model

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

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

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

which indicates whether francisco appears in the document with class SFO. If it appears probability = 1 else 0.

Ans: $P(x_{francisco} = true \mid class = SFO) = 1.0$

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

Ans: 0.5

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

Ans: 1.0

(1b) Multinomial NB model

$P(x = francisco \mid class = SFO) = 9/14$

$P(x = london \mid class = SFO) = 1/14$

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

(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. However, it ignore position information, so doesn't distinguish between a cityname occuring at the beginning/end of the itinerary from the one which is occuring in the middle of the

Name: Yalamanchili Sowmya
ClassID: 30

(1d) We can use as a feature the term that occurs in the last position of each document. Non standard feature represented with using non-standard words. The non-standard words are classified to 6 categories using SKIPEZ collection to official, literature, informative, popular, educational and scientific.

$$P(x_{newyork} = true \mid class = JFK) = 1.0$$
$$P(x_{sanfrancisco} = true \mid class = SFO) = 1.0$$
$$P(x_{chicago} = true \mid class = ORD) = 1.0$$

(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 words were seen, and then all categories are tied for the classifie It will rank between classes for which all words we seen similarly to the smoothed classifier (but with possible differences due to the smoothing).

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

formula for laplace (add-1) smoothing for naive Baye

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum (count(w, c)) + 1)}$$

$$= \frac{\text{count}(w_1, c) + 1}{\left(\sum_{w \in V} \text{count}(w, c)\right) + |V|}$$

It will be more likely to choose categories for which some/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.

(a) $\text{precision} = \frac{TP}{TP + FP}$

Here $TP = 3$

$FP = 2$

$= \frac{3}{3+2}$

$= \frac{3}{5}$

$\text{Recall} = \frac{TP}{TP + FN}$

Here $TP = 3$

$FN = 5$

$= \frac{3}{3+5}$

$= \frac{3}{8}$

Therefore precision $= \frac{3}{5}$

recall $= \frac{3}{8}$

Name: Yakkanchili Soumya

ClassID: 30

(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 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 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 is 1/3, the precision for B_q is 0. since B_q didn't return any relevant documents, it is of no utility.