# CS5560 Knowledge Discovery and Management

Problem Set 6

July 10 (T), 2017

Name: SUJITHA PUTHANA

Class ID: 24

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.

1 a) **Bernouil Model** :- It is equivalent to the binary independance model, which generates an indicator for each term in the vocabulary either "1" indicating presence of the term in the document or "0" indicating absence.

→ P (x fransisco = true | class SFO)

- Total no. of class of SFO = 2
- Presence of term "fransico" in no. of doc = 2

$\Rightarrow 2/2 = \boxed{1}$

→ P (x london = true | class = SFO)

- Total SFO classes = 2
- Presence of "london" in doc = 1

$\Rightarrow 1/2 = \boxed{0.5}$

→ P (x fransico = true | class JFK)

- Total JFK classes = 1
- Presence of "fransico" in doc = 1

$\Rightarrow 1/1 = \boxed{0.}$

b) **Multinomial Model** :- It generates one term from the vocabulary in each position of the document, where we assume a generative model

| | Multinomial Model | Bernoulli model |
|---|---|---|
| Event model | Generation of token | Generation of document |
| Random Variable | $X = t$ iff $t$ occurs at given pos | $U_t = 1$ iff $t$ occurs in doc |
| document representation | $d = \{t_1, \ldots, t_k, \ldots, t_{n_d}\}, t_k \in V$ | $d = \{e_1, \ldots e_i, \ldots e_M\}$ $e_i \in \{0,1\}$ |
| Parameter estimation | $\hat{P}(X = t \mid c)$ | $\hat{P}(U_i = e \mid c)$ |
| decision rule: maximize | $\hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(X = t_k \mid c)$ | $\hat{P}(c) \prod_{t \in V} \hat{P}(U_i = e_i \mid c)$ |
| Multiple Occurences | taken into account | ignored |
| Length of docs | can handle longer docs | Work best for short docs |
| # feature | can handle more | Works best with fewer |
| estimate for term the | $\hat{P}(X = \text{the} \mid c) \approx 0.5$ | $\hat{P}(U_{the} = 1 \mid c) \approx 1.0$ |

→ $P(X = \text{fransisco} \mid \text{Class} = SFO) =$

- Occurance of Fransicso in class SFO = 4
- Word Count in class SFO = 14

⇒ $4/14$

→ $P(X = \text{london} \mid \text{Class} = SFO)$

- Occurance of Word "london" = 1
- Word Count in class SFO = 14

⇒ $1/14$

→ $P(X = \text{fransico} \mid \text{Class} = JFK)$

- Occurance of Word fransico in class JFK = 1
- Word Count in class JFK = 8

⇒ $1/8$

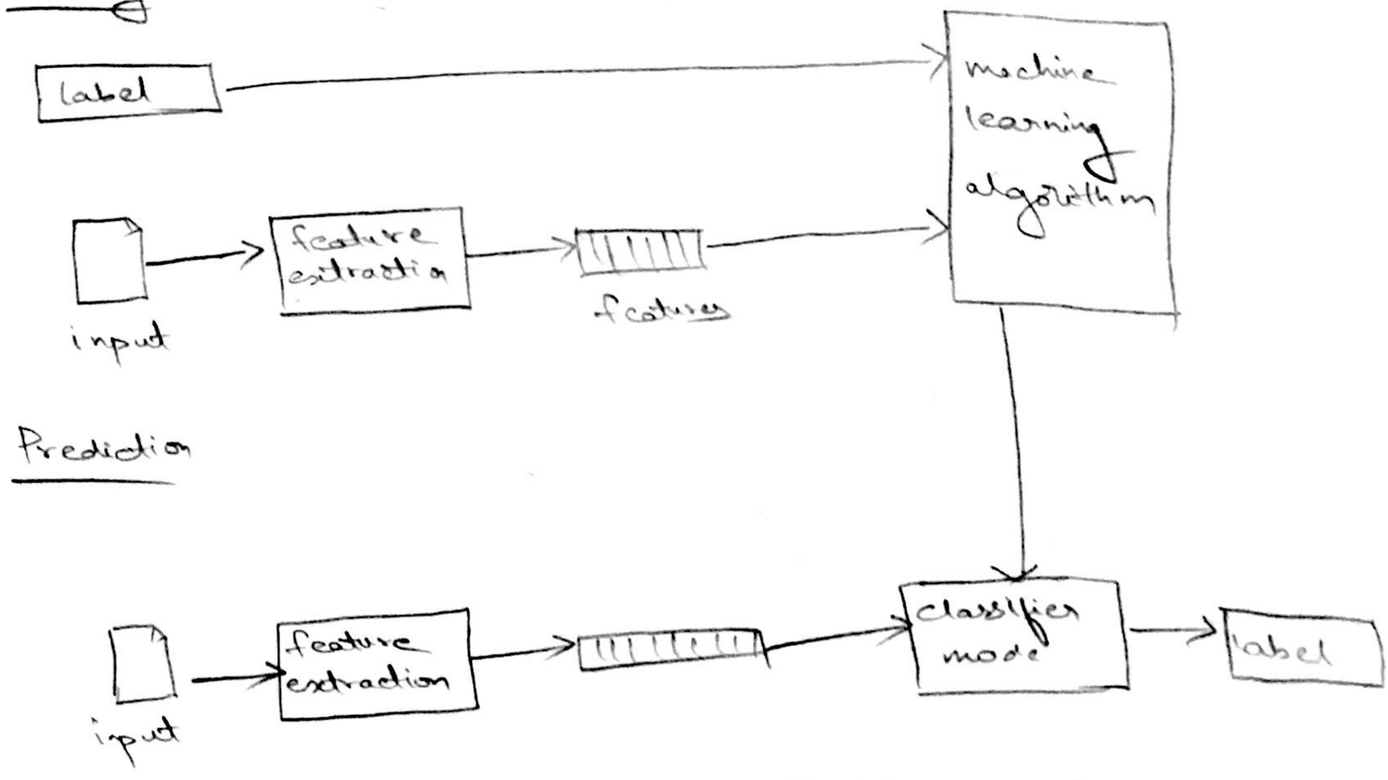
c) → Multinomial model takes into accout multiple occurence of the word and can handle more docus and features so multinomial model is more accurate, because it uses frequency information

→ Bernoulli Model in contrast does not consider the frequent occurence of words, so it is not much accurate, as it models absence of terms explicitly.

2) Classification :-

Classification is the task of choosing the correct class label for a given input

Training




→ A classifier is called supervised if it is built based on training corpus containing the correct label for each input.

a) → In the naive bayes classifier when occurance of word is "o" ⇒ probability of occurance is 0

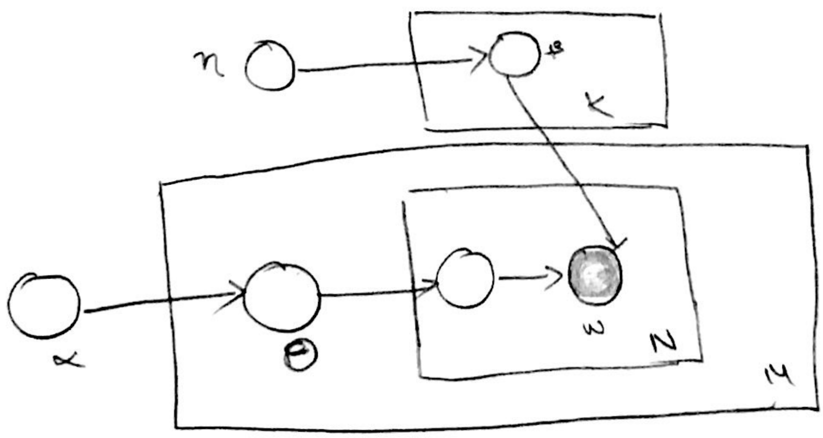
→ If p(word | class) = 0, then we can never choose a category.

→ Classification are based on training set, so we can rank for classes for which all words were seen similarity to the smoothed classifier.

b) Smoothing Multinomial Model :-

→ Large vocabulary size is characteristic of document corpora often problems with sparsity.

→ Maximum likelihood estimates of the multinomial parameters β assign zero probability to new words, and thus zero probability to new document.




Retrieval formula using general smoothing scheme

$$\log p(q|d) = \sum_{w \in V, c(w|q) > 0} c(w|q) \log p(w|d)$$

$$p(w|d) = \begin{cases} P_{seen}(w|d) & \text{if } w \text{ is seen in } d \\ \alpha_d \, p(w|c) & \text{otherwise} \end{cases}$$

→ $P(word \mid count)$

Let initial word = w
count = c

$\Rightarrow 1/K$

→ Increas in smoothing

$\Rightarrow 1/2K$

→ It'll be more likely to choose categories for which many of the words in the document

III) Total positive (tp) = 5

Total (tp+fp) = 8

| | ↓ Relevant | ↓ Irrelevant |
|---|---|---|
| Retrieved → | True positive<br>3<br>relevant, retrieved | false positive<br>2<br>irrelevant, retrieved |
| Not retrieved → | False negative<br>5<br>relevant, Not retrieved | True Negative<br>irrelevant, not retrieved |

$$\text{Precision} = \frac{tp}{(tp + fp)} = \frac{3}{3+2} = \underline{\frac{3}{5}}$$

$$\text{Recall} = \frac{tp}{(tp + fn)} = \frac{3}{8} = \underline{\frac{3}{8}}$$

Scanned by CamScanner

b) (i)

-> Precision, indicates how many of the items that we identified were relevant

-> Recall, indicates how many of the relevant items that we identified.

| | Relevant | Irrelevant |
|---|---|---|
| Retrieved | TP | FP |
| Not Retrieved | FN | TN |

-> For the information retrieval system, returns no results will have high accuracy for most queries, since the corpus usually contains only a few relevant documents

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

ii) Precision (Aq) > Precision (Bq)

For Aq

| | Relevant | Irrelevant |
|---|---|---|
| Retrieved | 1 | 2 |
| Not retrieved | 1 | |

=> Precision (Aq) = $1/3$

For Bq

| | Relevant | Irrelevant |
|---|---|---|
| Retrieved | 0 | 2 |
| Not Retrieved | 2 | |

=> Precision (Bq) = 0

→ In the above case both $A_q$ & $B_q$ made 2 mistakes ⟹ accuracy = 80%

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