UPPSALA
UNIVERSITET

Information Retrieval (5LN712)

Text classification and naïve Bayes

2020-05-25

Ali Basirat

Department of Linguistics and Philology

UPPSALA UNIVERSITET

Today

- The text classification problem
- The naïve Bayesian text classification
- Multinomial model
- The Bernoulli model
- A comparison between the two models
- Summary

UPPSALA UNIVERSITET

The text classification problem

- Queries are generated by probabilistic models from some topic (or document class)
- For example: the terms Taipei and Beijing are generated from a broader document class, China
- In order to classify a document, we should find a class that is more likely to generate the document

UPPSALA UNIVERSITET

The text classification problem

- For a given document space X and a set f of classes C, a document classifier maps each document description in X to a class in C

$$\gamma: X \rightarrow C$$

- The classification function γ is learnt from a training set of labelled documents $(d, c) \in X \times C$

UPPSALA UNIVERSITET

The Naïve Bayesian Approach

- The document-class probabilities can be estimated by the Bayesin equation:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

- The denominator $P(d)$ is a constant for a document and can be ignored

UPPSALA UNIVERSITET

The Naïve Bayesian Approach

- The prior probability $P(c)$ can be estimate from the data and other information about the classes
- We can make a naïve assumption about the independence occurrence of words in a document
- This assumption is not correct but helps us to have an estimation of the document probabilities

UPPSALA UNIVERSITET

The Naïve Bayesian Approach

- If we consider a multinomial probability distribution over positions

$$P(d|c) \approx \prod_{t \in d} P(t|c)$$

- If we consider a Bernoulli distribution over the occurrence of words

$$P(d|c) \approx \prod_{e \in V} P(e|c)$$

UPPSALA UNIVERSITET

Multinomial model

- The probability of a document d being in class c

$$P(c|d) \propto p(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- The best class for a document in NB is the most likely one (maximum a posteriori):

$$c_{map} = \arg\max_{c \in C} \hat{P}(c|d)$$

UPPSALA UNIVERSITET

Multinomial model

- If we expand $\hat{P}(c|d)$, we will have

$$c_{map} = \arg\max_{c \in C} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

- We can also take a logarithm from the right side

$$c_{map} = \arg\max_{c \in C} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

UPPSALA UNIVERSITET

Multinomial model

- We may have zero probabilities for some terms
- It can be resolved by an add-one or Laplace smoothing

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{|c| + |V|}$$

- $|c|$ is the total number of terms in class c
- $|V|$ is the vocabulary size of the collection

UPPSALA UNIVERSITET

Multinomial model

	Doc ID	Words in documents	class
Training set	1	Chinese Beijing Chinese	1
	2	Chinese Chinese Shanghai	1
	3	Chinese Macao	1
	4	Tokyo Japan Chinese	0
Test set	5	Chinese Chinese Chinese Tokyo Japan	?

UPPSALA UNIVERSITET

Multinomial model

TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c / N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

APPLYMULTINOMIALNB($\mathbb{C}, V, prior, condprob, d$)
1 $W \leftarrow$ EXTRACTTOKENSFROMDOC(V, d)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in W$
5 **do** $score[c] \mathrel{+}= \log condprob[t][c]$
6 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

UPPSALA UNIVERSITET

Multinomial model

- The multinomial NB (A) is identical to the multinomial unigram language model (B)
- A: $P(c|d) \propto p(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$
- B: $P(d|q) \propto P(d) \prod_{t \in q} P(t|M_d)$
- The document in A takes the role of a query in B
- The class in A takes the role of a document in B

UPPSALA UNIVERSITET

The Bernoulli model

- A Bernoulli random variable takes one of two values
- A multivariate Bernoulli variable is a vector of Bernoulli variables which is used as an indicator
- When using in text classification, a Bernoulli model only considers the presence of a term in a class
- It ignores the number of occurrences
- It is equivalent to the binary independence model

UPPSALA UNIVERSITET

Bernoulli model

$$P(c|d) \propto P(c)P(d|c)$$

$$P(d = [e_1, \dots, e_M]|c; V) = \prod_{e \in V} P(e|c)$$

$$= \prod_{e \in V_d} P(e|c) \prod_{e \notin V_d} (1 - P(e|c))$$

$$P(c) = \frac{N_c}{N}$$

$$P(d|c) = \frac{N_{ct} + 1}{N_c + 2}$$

- N_c is the number of documents in class c
- N_{ct} is the number of document in class c containing term t
- N is the total number of documents

UPPSALA UNIVERSITET

Bernoulli model

TRAINBERNOULLINB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c / N$
6 **for each** $t \in V$
7 **do** $N_{ct} \leftarrow$ COUNTDOCSINCLASSCONTAININGTERM($\mathbb{D}, c, t$)
8 $condprob[t][c] \leftarrow (N_{ct} + 1)/(N_c + 2)$
9 **return** $V, prior, condprob$

APPLYBERNOULLINB($\mathbb{C}, V, prior, condprob, d$)
1 $V_d \leftarrow$ EXTRACTTERMSFROMDOC(V, d)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in V$
5 **do if** $t \in V_d$
6 **then** $score[c] \mathrel{+}= \log condprob[t][c]$
7 **else** $score[c] \mathrel{+}= \log(1 - condprob[t][c])$
8 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

UPPSALA UNIVERSITET

Bernoulli model

	Doc ID	Words in documents	class
Training set	1	Chinese Beijing Chinese	1
	2	Chinese Chinese Shanghai	1
	3	Chinese Macao	1
	4	Tokyo Japan Chinese	0
Test set	5	Chinese Chinese Chinese Tokyo Japan	?

UPPSALA UNIVERSITET

Multinomial vs Bernoulli model

	Multinomial model	**Bernoulli model**
Event model	Generation of token	Generation of document
Random variable	X=t iff t occurs at given position	$U_t = 1$ iff t occurs in the document
Document representation	$d =< t_1, \dots, t_k, \dots, t_{nd} >, t_k \in V$	$d =< e, \dots, e_i, \dots, e_M >, e_i \in \{0,1\}$
Parameter estimation	$\hat{P}(X = t\|c)$	$\hat{P}(U_i = e\|c)$
Multiple occurrences	Taken into account	Ignored
Length of documents	Can handle longer documents	Works best for short documents
Number of features	Can handle more	Works best with fewer
Estimate for term ‘the’	$\hat{P}\left(X =' the'\middle\|c\right) \approx 0.05$	$\hat{P}(U_{the} = 1\|c) \approx 1.0$

UPPSALA UNIVERSITET

Summary

- Documents can be classified into multiple classes
- A text classifier is a function from document space to a label set
- A naïve Bayes classifier can be used for document classification
- It is similar to a language model information retrieval
- A Bernoulli model can also be used for document classification