Inf2b - Learning

Lecture 7: Text Classification using Naive Bayes

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

http://www.inf.ed.ac.uk/teaching/courses/inf2b/
https://piazza.com/ed.ac.uk/spring2020/infr08028
Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2020

Today's Schedule

1. Text classification
2. Bag-of-words models
3. Multinomial document model
4. Bernoulli document model
5. Generative models
6. Zero Probability Problem

Identifying Spam

Spam?

I got your contact information from your country's information directory during my desperate search for someone who can assist me secretly and confidentially in relocating and managing some family fortunes.

Identifying Spam

Spam?

Dear Dr. Steve Renals, The proof for your article, Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition, is ready for your review. Please access your proof via the user ID and password provided below. Kindly log in to the website within 48 HOURS of receiving this message so that we may expedite the publication process.

Identifying Spam

Spam?

Congratulations to you as we bring to your notice, the results of the First Category draws of THE HOLLAND CASINO LOTTO PROMO INT. We are happy to inform you that you have emerged a winner under the First Category, which is part of our promotional draws.

Text Classification using Bayes Theorem

- Document $\mathcal{D}$, with a fixed set of classes $C = \{1, \ldots, K\}$
- Classify $\mathcal{D}$ as the class with the highest posterior probability:

$$
\begin{aligned}
k_{\max} = & \arg\max_k P(C_k \mid \mathcal{D}) = \arg\max_k \frac{P(\mathcal{D} \mid C_k)\, P(C_k)}{P(\mathcal{D})} \\
= & \arg\max_k \; P(\mathcal{D} \mid C_k)\, P(C_k)
\end{aligned}
$$

- How do we represent $\mathcal{D}$?
- How do we estimate $P(\mathcal{D} \mid C_k)$ and $P(C_k)$?

How do we represent $\mathcal{D}$?

- A sequence of words: $\mathcal{D} = (X_1, X_2, \ldots, X_n)$
  computational very expensive, difficult to train

- A set of words (Bag-of-Words)
  - Ignore the position of the word
  - Ignore the order of the word
  - Consider the words in pre-defined vocabulary V $(D = |V|)$

Multinomial document model a document is represented by an integer feature vector, whose elements indicate frequency of corresponding word in the document

$$\mathbf{x} = (x_1, \ldots, x_D) \qquad x_i \in \mathcal{N}_0$$

Bernoulli document model a document is represented by a binary feature vector, whose elements indicate absence or presence of corresponding word in the document

$$\mathbf{b} = (b_1, \ldots, b_D) \qquad b_i \in \{0, 1\}$$

BoW models: Bernoulli vs. Multinomial

Document $\mathcal{D}$: "Congratulations to you as we bring to your notice, the results of the First Category draws of THE HOLLAND CASINO LOTTO PROMO INT. We are happy to inform you that you have emerged a winner under the First Category, which is part of our promotional draws."

Term ($w_t \in V$)	Multinomial ($x_t \in \mathcal{N}_0$) $\mathbf{x} = (x_t)$	Bernoulli ($b_t \in \{0,1\}$) $\mathbf{b} = (b_t)$
bring	1	1
can	0	0
casino	1	1
category	2	1
congratulations	1	1
draws	2	1
first	2	1
lotto	1	1
the	4	1
true	0	0
winner	1	1
you	3	1
$D = 12$	$\mathbf{x} = (1, 0, 1, 2, \ldots, 1, 3)$	$\mathbf{b} = (1, 0, 1, 1, \ldots, 1, 1)$

Notation for document model

- Training documents:

Class	Documents
C_1	$\mathcal{D}_1^{(1)} \ldots \mathcal{D}_i^{(1)} \ldots \mathcal{D}_{N_1}^{(1)}$
$\vdots$	$\vdots$
C_K	$\mathcal{D}_1^{(K)} \ldots \mathcal{D}_i^{(K)} \ldots \mathcal{D}_{N_K}^{(K)}$

- Flattened representation of training data:

Documents	$\mathcal{D}_1$	$\ldots$	$\mathcal{D}_i$	$\ldots$	$\mathcal{D}_N$
Class indicator	z_{1k}	$\ldots$	z_{ik}	$\ldots$	z_{Nk}

where $N = N_1 + \cdots + N_K$,

$$z_{ik} = \begin{cases} 1 & \text{if } \mathcal{D}_i \text{ belongs to class } C_k \\ 0 & \text{otherwise} \end{cases}$$

- Test document : $\mathcal{D}$

Discrete probability distributions - review

Bernoulli distribution

Eg: Tossing a biased coin ($P(H) = p$), the probability of $k = \{0, 1\}$ 0:Tail, 1:Head is

$$P(k) = kp + (1-k)(1-p) = p^k(1-p)^{1-k}$$

Binomial distribution

Eg: Tossing a biased coin n times, the probability of observing Head k times is

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}. \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Multinomial distribution

Eg: Tossing a biased dice n times, the probability of $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$, where x_i is the number of occurrences for face i, is

$$P(\mathbf{x}) = \frac{n!}{x_1! \cdots x_6!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} p_5^{x_5} p_6^{x_6}.$$

Classification with multinomial document model

Assume a test document $\mathcal{D}$ is given as a sequence of words:

$$(o_1,\ o_2,\ \ldots,\ o_n) \qquad o_i \in V = \{w_1, \ldots, w_D\}$$

Feature vector: $\mathbf{x} = (x_1, \ldots, x_D)$ $\cdots$ word frequencies, $\sum_{t=1}^{D} x_t = n$

Document likelihood with multinomial distribution:

$$P(\mathbf{x} \mid C_k) = \frac{n\,!}{\prod_{t=1}^{D} x_t!} \prod_{t=1}^{D} P(w_t \mid C_k)^{x_t} \qquad \text{NB: } P^0 = 1 \ \ (P > 0)$$

For classification, we can omit irrelevant term, so that:

$$P(\mathbf{x} \mid C_k) \propto \prod_{t=1}^{D} P(w_t \mid C_k)^{x_t} \ = \ P(o_1|C_k)\, P(o_2|C_k) \cdots P(o_n|C_k)$$

$$P(C_k \mid \mathbf{x}) \propto P(C_k) \prod_{i=1}^{n} P(o_i \mid C_k)$$

Training of multinomial document model

Features: $\mathbf{x} = (x_1, \ldots, x_D)$: *word frequencies* in a doc.

Training data set

Class	Docs	Feature vectors
C_1	$\mathcal{D}_1^{(1)}$ $\vdots$ $\mathcal{D}_{N_1}^{(1)}$	$\begin{pmatrix} \mathbf{x}_1^{(1)} \\ \vdots \\ \mathbf{x}_{N_1}^{(1)} \end{pmatrix} = \begin{pmatrix} x_{11}^{(1)} & \ldots & x_{1D}^{(1)} \\ \vdots & & \vdots \\ x_{N_1 1}^{(1)} & \ldots & x_{N_1 D}^{(1)} \end{pmatrix}$
	$\hat{P}(C_1) = N_1/N$	$n_1(w_1), \ldots, n_1(w_D)$ $\hat{P}(w_t\|C_1)$: $n_1(w_1)/S_1, \ldots, n_1(w_D)/S_1$
C_k	$\mathcal{D}_1^{(k)}$ $\vdots$ $\mathcal{D}_{N_k}^{(k)}$	$\begin{pmatrix} \mathbf{x}_1^{(k)} \\ \vdots \\ \mathbf{x}_{N_k}^{(k)} \end{pmatrix} = \begin{pmatrix} x_{11}^{(k)} & \ldots & x_{1D}^{(k)} \\ \vdots & & \\ x_{N_k 1}^{(k)} & \ldots & x_{N_1 D}^{(k)} \end{pmatrix}$
	$\hat{P}(C_k) = N_k/N$	$n_k(w_1), \ldots, n_k(w_D)$ $\hat{P}(w_t\|C_k)$: $n_k(w_1)/S_k \ldots, n_k(w_D)/S_k$

$S_k = \sum_{t=1}^{D} n_k(w_t)$

Multinomial doc. model – example

See Note 7!

Classification with Bernoulli document model

A test document $\mathcal{D}$ with feature vector $\boldsymbol{b} = (b_1, \ldots, b_D)$

Document likelihood with (multivariate) Bernoulli distribution:

$$P(\boldsymbol{b} \mid C_k) = \prod_{t=1}^{D} P(b_t|C_k) = \prod_{t=1}^{D} [b_t P(w_t|C_k) + (1-b_t)(1-P(w_t|C_k))]$$

$$= \prod_{t=1}^{D} P(w_t \mid C_k)^{b_t}(1-P(w_t|C_k))^{(1-b_t)}$$

$$\hat{P}(w_t|C_k) = \frac{n_k(w_t)}{N_k}$$

(fraction of class k docs with word w_t)

In Classification,

$$P(C_k \mid \boldsymbol{b}) \propto P(C_k)\, P(\boldsymbol{b} \mid C_k)$$

Training of Bernoulli document model

Features: $\mathbf{b} = (b_1, \ldots, b_D)$: $D = |V|$, i.e. vocabulary
binary vector of word occurrences in a document

Training data set

Class	Docs	Feature vectors
C_1	$\mathcal{D}_1^{(1)}$ $\vdots$ $\mathcal{D}_{N_1}^{(1)}$	$\begin{pmatrix} \mathbf{b}_1^{(1)} \\ \vdots \\ \mathbf{b}_{N_1}^{(1)} \end{pmatrix} = \begin{pmatrix} b_{11}^{(1)} & \ldots & b_{1D}^{(1)} \\ \vdots & & \vdots \\ b_{N_1 1}^{(1)} & \ldots & b_{N_1 D}^{(1)} \end{pmatrix}$
	$\hat{P}(C_1) = N_1/N$	$n_1(w_1), \ldots, n_1(w_D)$ $\hat{P}(w_t \mid C_1)$: $n_1(w_1)/N_1, \ldots, n_1(w_D)/N_1$
C_k	$\mathcal{D}_1^{(k)}$ $\vdots$ $\mathcal{D}_{N_k}^{(k)}$	$\begin{pmatrix} \mathbf{b}_1^{(k)} \\ \vdots \\ \mathbf{b}_{N_k}^{(k)} \end{pmatrix} = \begin{pmatrix} b_{11}^{(k)} & \ldots & b_{1D}^{(k)} \\ \vdots & & \\ b_{N_1 1}^{(k)} & \ldots & b_{N_1 D}^{(k)} \end{pmatrix}$
	$\hat{P}(C_k) = N_k/N$	$n_k(w_1), \ldots, n_k(w_D)$ $\hat{P}(w_t \mid C_k)$: $n_k(w_1)/N_k, \ldots, n_k(w_D)/N_k$

Bernoulli doc. model – example

Classify documents as Sports (S) or Informatics (I)

Vocabulary V:

$$
\begin{aligned}
w_1 &= \textit{goal}\\
w_2 &= \textit{tutor}\\
w_3 &= \textit{variance}\\
w_4 &= \textit{speed}\\
w_5 &= \textit{drink}\\
w_6 &= \textit{defence}\\
w_7 &= \textit{performance}\\
w_8 &= \textit{field}
\end{aligned}
$$

$$D = |V| = 8$$

Bernoulli doc. model – example *(cont.)*

Training data: (rows give documents, columns word presence)

$$\mathbf{B}^{\mathrm{Sport}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\mathbf{B}^{\mathrm{Inf}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Estimating priors and likelihoods:

$$P(S) = 6/11, \quad P(I) = 5/11$$

$$(P(w_t|S)) = \begin{pmatrix} 3/6 & 1/6 & 2/6 & 3/6 & 3/6 & 4/6 & 4/6 & 4/6 \end{pmatrix}$$

$$(P(w_t|I)) = \begin{pmatrix} 1/5 & 3/5 & 3/5 & 1/5 & 1/5 & 1/5 & 3/5 & 1/5 \end{pmatrix}$$

Bernoulli doc. model – example *(cont.)*

Test documents: $\boldsymbol{b}_1 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}$

Priors, Likelihoods: $P(S) = 6/11, \quad P(I) = 5/11$

$$(P(w_t|S)) = \begin{pmatrix} 3/6 & 1/6 & 2/6 & 3/6 & 3/6 & 4/6 & 4/6 & 4/6 \end{pmatrix}$$

$$(P(w_t|I)) = \begin{pmatrix} 1/5 & 3/5 & 3/5 & 1/5 & 1/5 & 1/5 & 3/5 & 1/5 \end{pmatrix}$$

Posterior probabilities:

$$P(S \mid \boldsymbol{b}_1) \propto P(S) \prod_{t=1}^{8} [b_{1t} P(w_t \mid S) + (1 - b_{1t})(1 - P(w_t \mid S))]$$

$$\propto \frac{6}{11} \left(\frac{1}{2} \times \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \right) = \frac{5}{891} = 5.6 \times 10^{-3}$$

$$P(I \mid \boldsymbol{b}_1) \propto P(I) \prod_{t=1}^{8} [b_{1t} P(w_t \mid I) + (1 - b_{1t})(1 - P(w_t \mid I))]$$

$$\propto \frac{5}{11} \left(\frac{1}{5} \times \frac{2}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{1}{5} \right) = \frac{8}{859375} = 9.3 \times 10^{-6}$$

$\Rightarrow$ Classify this document as S.

Summary of the document models

Class	Doc	Multinomial doc. model feature vectors	Bernoulli doc. model feature vectors
C_k	$\begin{matrix}\mathcal{D}_1^{(k)} \\ \vdots \\ \mathcal{D}_{N_k}^{(k)}\end{matrix}$	$\begin{pmatrix}\mathbf{x}_1^{(k)} \\ \vdots \\ \mathbf{x}_{N_k}^{(k)}\end{pmatrix} = \begin{pmatrix} x_{11}^{(k)} & \dots & x_{1D}^{(k)} \\ \vdots & & \vdots \\ x_{N_k 1}^{(k)} & \dots & x_{1D}^{(k)}\end{pmatrix}$	$\begin{pmatrix}\mathbf{b}_1^{(k)} \\ \vdots \\ \mathbf{b}_{N_k}^{(k)}\end{pmatrix} = \begin{pmatrix} b_{11}^{(k)} & \dots & b_{1D}^{(k)} \\ \vdots & & \vdots \\ b_{N_k 1}^{(k)} & \dots & b_{1D}^{(k)}\end{pmatrix}$
$\hat{P}(C_k) = \frac{N_k}{N}$		$n_k(w_1), \dots, n_k(w_D)$	$n_k(w_1), \dots, n_k(w_D)$
$\hat{P}(w_t\|C_k)$:		$\frac{n_k(w_1)}{S_k}, \dots, \frac{n_k(w_D)}{S_k}$	$\frac{n_k(w_1)}{N_k}, \dots, \frac{n_k(w_D)}{N_k}$

$S_k = \sum_{t=1}^{D} n_k(w_t)$

$$P(\mathbf{x} \mid C_k) \propto \prod_{t=1}^{D} P(w_t|C_k)^{x_t} = \prod_{i=1}^{n} P(o_i|C_k)$$

$$P(\mathbf{b} \mid C_k) = \prod_{t=1}^{D} [b_t P(w_t|C_k) + (1-b_t)(1-P(w_t|C_k))]$$

Question

What's the approximate value of:

$$P(\text{“the”} \mid C)$$

(a) in the Bernoulli model

(b) in the multinomial model?

Common words, 'stop words', are often removed from feature vectors.

Generative models

- Models that generate observable data randomly based on a distribution

$$\boxed{P(X)} \longrightarrow x_1, x_2, x_3, \ldots$$

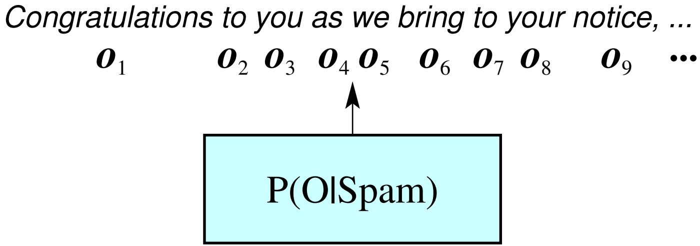
- Examples
  - Coin tossing models

Coin	Generated data sequence
Fair coin ($P(H) = P(T) = 0.5$)	$H, T, T, H, T, H, H, T, \ldots$
Unfair coin ($P(H) = 0.7, P(T) = 0.3$)	$T, H, H, H, H, H, T, H, \ldots$

  - Dice throwing models

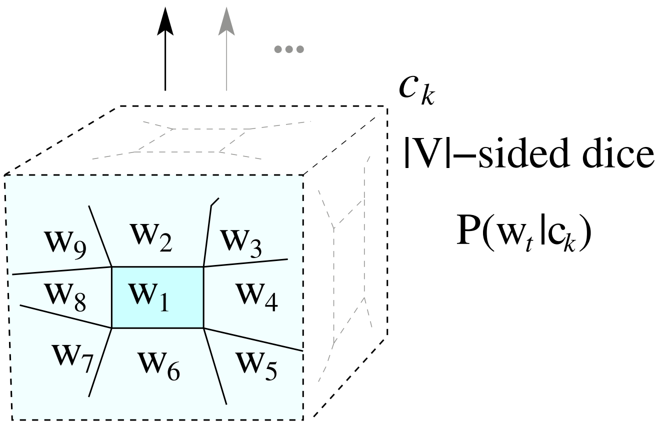
Dice	Generated data sequence
Unbiased dice ($P(X) = 1/6,\ X \in \{1, \ldots, 6\}$)	$2, 4, 3, 5, 3, 6, 5, 5, 4, 6, \ldots$
Biased dice ($P(X)) = (0.1, 0.1, 0.1, 0.1, 0.2, 0.4)$	$6, 6, 5, 5, 6, 1, 2, 6, 6, 6, \ldots$

- Spam mail generator

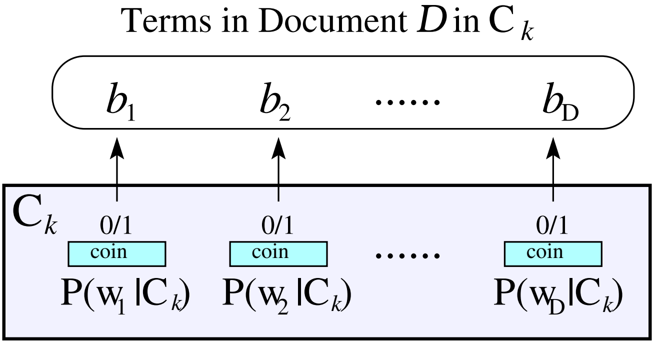


Generative model — Multinomial document model

$\boldsymbol{o}_1 \ \boldsymbol{o}_2 \ \boldsymbol{o}_3 \ \cdots \ \boldsymbol{o}_L \mid c_k$



Generative model — Bernoulli document model



Word relative-frequencies of spam emails

of spam emails: 169

to	0.0395032
the	0.0383633
you	0.0267285
of	0.0257851
and	0.0252349
your	0.0222476
in	0.0200857
i	0.0198892
this	0.0145828
a	0.0138752
my	0.0132463
for	0.0132463
is	0.0112024
3d	0.0108879
with	0.00915845
will	0.00876538
that	0.00849023
as	0.00797925
me	0.00766479
be	0.00703589

from	0.00664282
content	0.00644629
have	0.0059353
bank	0.0059353
usd	0.00581738
on	0.00554223
we	0.00542432
it	0.00518848
are	0.00507056
transfer	0.00479541
our	0.0047561
com	0.00467749
am	0.00467749
account	0.00455957
unlocked	0.00424512
20	0.0041665
email	0.00404858
please	0.00385205
not	0.00377344
all	0.00377344

http	0.00369482
money	0.00345898
by	0.00338037
or	0.00330176
name	0.00322314
funds	0.00322314
was	0.00318384
type	0.00318384
s	0.00318384
0a	0.00314453
if	0.00310522
1	0.00310522
can	0.00306592
payment	0.002948
message	0.002948
address	0.00286938
us	0.00283008
his	0.00279077
contact	0.00279077
has	0.00271216

Generated word sequence example

of kin good your the part of with and atm to new from which projects has the transfer my how 3d and with united in in o beneficiary that died pathak id efforts has to studies have my as can you the 3d you your with transfer will your a your m and the your i is ve country user nokia the this for i value banking an click confirm world i it me my country is 2010 very below i and now until html of position http here of mail following there be while the by for your willing

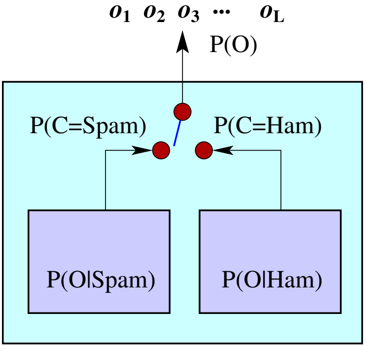
Generative models for classification

Model for classification

$$P(C_k \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid C_k)\, P(C_k)}{P(\mathbf{x})} \propto P(\mathbf{x} \mid C_k)\, P(C_k)$$

Model for observation $\cdots$ generative model

$$P(x) = \sum_{k=1}^{K} P(x|C_k)P(C_k)$$



Smoothing in multinomial document model

- Zero probability problem

$$P(\boldsymbol{x} \mid C_k) \propto \prod_{t=1}^{D} P(w_t \mid C_k)^{x_t} \;=\; 0 \text{ if } \exists j : P(w_j|C_k) = 0$$

$$P(w_t \mid C_k) = \frac{\sum_{i=1}^{N} x_{it} z_{ik}}{\sum_{t'=1}^{|V|} \sum_{i=1}^{N} x_{it'} z_{ik}} = \frac{n_k(w_t)}{\sum_{t'=1}^{D} n_k(w_{t'})}$$

- Smoothing – a 'trick' to avoid zero counts:

$$P(w_t \mid C_k) = \frac{1 + \sum_{i=1}^{N} x_{it} z_{ik}}{|V| + \sum_{t'=1}^{|V|} \sum_{i=1}^{N} x_{it'} z_{ik}} = \frac{1 + n_k(w_t)}{D + \sum_{t'=1}^{D} n_k(w_{t'})}$$

Known as *Laplace's rule of succession* or *add one smoothing*.

Multinomial vs Bernoulli doc. models

	Multinomial	Bernoulli
Generative model	draw a words from a multinomial distribution	draw a document from a multi-dimensional Bernoulli distribution
Document representation	Vector of frequencies	Binary vector
Multiple occurrences	Taken into account	Ignored
Document length	Longer docs OK	Best for short docs
Feature vector dimension	Longer OK	Shorter
Behaviour with "the"	$P(\text{"the"} \mid C_k) \approx 0.05$	$P(\text{"the"} \mid C_k) \approx 1.0$
Non-occurring words in test doc	do not affect likelihood	affect likelihood

Multinomial vs Bernoulli doc. models *(cont.)*

Fig. 1 in A. McCallum and K.Nigam, "A Comparison of Event Models for Naive Bayes Text Classification", AAAI Workshop on Learning for Text Categorization, 1998

Document pre-processing

- Stop-word removal

  Remove pre-defined common words that are not specific or discriminatory to the different classes.

- Stemming

  Reduce different forms of the same word into a single word (base/root form)

- Feature selection

  e.g. choose words based on the mutual information

Exercise 1

Use the Bernoulli model and the Naive Bayes assumption for the following.

Consider the vocabulary $V = \{\texttt{apple}, \texttt{banana}, \texttt{computer}\}$. We have two classes of documents F (fruit) and E (electronics). There are four training documents in class F; they are listed below in terms of the number of occurrences of each word from V in each document:

- `apple`(2); `banana`(1); `computer`(0)
- `apple`(0); `banana`(1); `computer`(0)
- `apple`(3); `banana`(2); `computer`(1)
- `apple`(1); `banana`(0); `computer`(0)

There are also four training documents in class E:

- `apple`(2); `banana`(0); `computer`(0)
- `apple`(0); `banana`(0); `computer`(1)
- `apple`(3); `banana`(1); `computer`(2)
- `apple`(0); `banana`(0); `computer`(1)

Exercise 1 *(cont.)*

1. Write the training data as a matrix for each class, where each row corresponds to a training document.
2. Estimate the prior probabilities from the training data
3. For each class (F and E) and for each word (apple, banana and computer) estimate the likelihood of the word given the class.
4. Consider two test documents:
   - `apple`(1); `banana`(0); `computer`(0)
   - `apple`(1); `banana`(1); `computer`(0)

   For each test document, estimate the posterior probabilities of each class given the document, and hence classify the document.

Exercise 2

Use the Multinomial model and the Naive Bayes assumption for the following.
Consider the vocabulary $V = \{\texttt{fish}, \texttt{chip}, \texttt{circuit}\}$. We have two classes of documents F (food) and E (electronics). There are four training documents in class F; they are listed below;

- `fish chip fish`
- `chip`
- `circuit fish chip`
- `fish fish`

There are also four training documents in class E:

- `circuit circuit`
- `chip circuit`
- `chip chip`
- `circuit`

Exercise 2 *(cont.)*

1. Estimate the parameters of a multinomial model for the two document classes, using add-one smoothing.
2. Consider two test documents:
   - `fish chip`
   - `chip circuit chip circuit fish chip circuit`

   Classify each of the test documents by (approximately) estimating the posterior probability of each class
3. With reference to the test documents in the previous question, explain why a process such as add-one smoothing is used when estimating the parameters of a multinomial model.

Exercise 3

Consider two writers, Baker and Clark, who were twins, and who published four and six children's books, respectively. The following table shows the frequencies of four words, **wizard**, **river**, **star**, and **warp**, with respect to the first page of each book, and the information whether the book was a bestseller or not.

	Words				
Author	wizard	river	star	warp	Bestseller
Baker	1	1	1	0	No
Baker	1	1	0	1	No
Baker	1	1	1	1	yes
Baker	1	1	0	0	No
Clark	0	1	0	1	No
Clark	0	0	2	1	No
Clark	0	2	1	2	Yes
Clark	1	1	1	2	No
Clark	0	1	2	2	Yes
Clark	0	1	2	1	Yes

Two unpublished book drafts, Doc 1 and Doc 2, were found after the death of the writers, but it's not clear which of them wrote the documents.

Exercise 3 *(cont.)*

1. Without having any information about Doc 1 and Doc 2, decide the most probable author of each document in terms of minimum classification error, and justify your decision.
2. The same analysis of word frequencies was carried out for Doc 1 and Doc 2, whose result is shown below. Using the Naive Bayes classification with the multinomial document model without smoothing, find the author of each document.

	wizard	river	start	warp
Doc 1	2	1	1	0
Doc 2	1	1	2	1

3. In addition to modifications to the vocabulary, discuss two possible methods for improving the classification performance.
4. Another document, Doc 3, was found later, and a publisher is considering its publication. Assuming the Naive Bayes classification with the multinomial document model with no smoothing, without identifying the author, predict whether Doc 3 is likely to be a bestseller or not based on the word frequency table for Doc 3 shown below.

	wizard	river	start	warp
Doc 3	0	1	1	2

5. Using the same situations as in part (d) except that we now know the author of Doc 3 was Baker, predict whether Doc 3 is likely to be a bestseller or not.

Summary

- Our first 'real' application of Naive Bayes
- Two BoW models for documents: Multinomial and Bernoulli
- Generative models
- Smoothing (Add-one/Laplace smoothing)
- Good reference:
  C. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, University Press. 2008.
  See Chapter 13 Text classification & Naive Bayes
- **As always:**
  be able to implement, describe, compare and contrast (see Lecture Note)