Inf2b Learning and Data
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/spring2019/infr08009inf2blearning
Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2019
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.
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.
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.
Document $\mathcal{D}$, with a fixed set of classes $C = \{1, \ldots, K\}$

Classify $\mathcal{D}$ as the class with the highest posterior probability:

$$k_{\text{max}} = \arg \max_k P(C_k | \mathcal{D}) = \arg \max_k \frac{P(\mathcal{D} | C_k) P(C_k)}{P(\mathcal{D})}$$

$$= \arg \max_k P(\mathcal{D} | C_k) P(C_k)$$

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

$$x = (x_1, \ldots, x_D) \quad x_i \in \mathbb{N}_0$$

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

$$b = (b_1, \ldots, b_D) \quad b_i \in \{0, 1\}$$
**Document** $D$: “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.”

<table>
<thead>
<tr>
<th>Term ($w_t \in V$)</th>
<th>Multinomial ($x_t \in \mathbb{N}_0$)</th>
<th>Bernoulli ($b_t \in {0, 1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bring</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>can</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>casino</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>category</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>congratulations</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>draws</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>first</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>lotto</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>true</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>winner</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>you</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

$D = 12$ \hspace{1cm} $x = (1, 0, 1, 2, \ldots, 1, 3)$ \hspace{1cm} $b = (1, 0, 1, 1, \ldots, 1, 1)$
Notation for document model

- Training documents:

<table>
<thead>
<tr>
<th>Class</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>( D_1^{(1)} ) ( \ldots ) ( D_i^{(1)} ) ( \ldots ) ( D_{N_1}^{(1)} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( C_K )</td>
<td>( D_1^{(K)} ) ( \ldots ) ( D_i^{(K)} ) ( \ldots ) ( D_{N_K}^{(K)} )</td>
</tr>
</tbody>
</table>

- Flattened representation of training data:

| Documents | \( D_1 \) \( \ldots \) \( D_i \) \( \ldots \) \( D_N \) |
| Class indicator | \( z_{1k} \) \( \ldots \) \( z_{ik} \) \( \ldots \) \( z_{Nk} \) |

where \( N = N_1 + \cdots + N_K \),

\[
z_{ik} = \begin{cases} 
1 & \text{if } D_i \text{ belongs to class } C_k \\
0 & \text{otherwise}
\end{cases}
\]
Assume a test document $D$ is given as a sequence of words:

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

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

Document likelihood with multinomial distribution:

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

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

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

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

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

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

**Multinomial distribution**

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

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

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

**Training data set**

<table>
<thead>
<tr>
<th>Class</th>
<th>Docs</th>
<th>Feature vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>$D^{(1)}_1$</td>
<td>$(\begin{pmatrix} x^{(1)}<em>1 \ \vdots \ x^{(1)}</em>{N_1} \end{pmatrix})$ = $(\begin{pmatrix} x_{11}^{(1)} &amp; \cdots &amp; x_{1D}^{(1)} \ \vdots &amp; \ddots &amp; \vdots \ x_{N_11}^{(1)} &amp; \cdots &amp; x_{N_1D}^{(1)} \end{pmatrix})$</td>
</tr>
<tr>
<td></td>
<td>$D^{(1)}_{N_1}$</td>
<td>$n_1(w_1), \ldots, n_1(w_D)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{P}(C_1) = N_1/N$</td>
<td>$\hat{P}(w_t</td>
</tr>
<tr>
<td>$C_k$</td>
<td>$D^{(k)}_1$</td>
<td>$(\begin{pmatrix} x^{(k)}<em>1 \ \vdots \ x^{(k)}</em>{N_k} \end{pmatrix})$ = $(\begin{pmatrix} x_{11}^{(k)} &amp; \cdots &amp; x_{1D}^{(k)} \ \vdots &amp; \ddots &amp; \vdots \ x_{N_k1}^{(k)} &amp; \cdots &amp; x_{N_kD}^{(k)} \end{pmatrix})$</td>
</tr>
<tr>
<td></td>
<td>$D^{(k)}_{N_k}$</td>
<td>$n_k(w_1), \ldots, n_k(w_D)$</td>
</tr>
<tr>
<td></td>
<td>$\hat{P}(C_k) = N_k/N$</td>
<td>$\hat{P}(w_t</td>
</tr>
<tr>
<td></td>
<td>$S_k = \sum_{t=1}^{D} n_k(w_t)$</td>
<td></td>
</tr>
</tbody>
</table>
See Note 7!
Classification with Bernoulli document model

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

Document likelihood with (multivariate) Bernoulli distribution:

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

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

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

(fraction of class $k$ docs with word $w_t$)

In Classification,

$$ P(C_k | b) \propto P(C_k) P(b | C_k) $$
### 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

<table>
<thead>
<tr>
<th>Class</th>
<th>Docs</th>
<th>Feature vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>( \mathcal{D}^{(1)}_1 )</td>
<td>( \mathbf{b}^{(1)}<em>1 ) = ( \begin{pmatrix} b</em>{11}^{(1)} &amp; \ldots &amp; b_{1D}^{(1)} \end{pmatrix} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( \mathcal{D}^{(1)}_{N_1} )</td>
<td>( \mathbf{b}^{(1)}_{N_1} )</td>
<td>( \begin{pmatrix} b_{N_11}^{(1)} &amp; \ldots &amp; b_{N_1D}^{(1)} \end{pmatrix} )</td>
</tr>
<tr>
<td>( \hat{P}(C_1) = \frac{N_1}{N} )</td>
<td>( \hat{P}(w_t</td>
<td>C_1) : n_1(w_1), \ldots, n_1(w_D) )</td>
</tr>
<tr>
<td>( \mathcal{D}^{(k)}_1 )</td>
<td>( \mathbf{b}^{(k)}<em>1 ) = ( \begin{pmatrix} b</em>{11}^{(k)} &amp; \ldots &amp; b_{1D}^{(k)} \end{pmatrix} )</td>
<td></td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( \mathcal{D}^{(k)}_{N_k} )</td>
<td>( \mathbf{b}^{(k)}_{N_k} )</td>
<td>( \begin{pmatrix} b_{N_k1}^{(k)} &amp; \ldots &amp; b_{N_kD}^{(k)} \end{pmatrix} )</td>
</tr>
<tr>
<td>( \hat{P}(C_k) = \frac{N_k}{N} )</td>
<td>( \hat{P}(w_t</td>
<td>C_k) : n_k(w_1), \ldots, n_k(w_D) )</td>
</tr>
</tbody>
</table>
Classify documents as Sports ($S$) or Informatics ($I$)

**Vocabulary $V$:**

- $w_1 = \text{goal}$
- $w_2 = \text{tutor}$
- $w_3 = \text{variance}$
- $w_4 = \text{speed}$
- $w_5 = \text{drink}$
- $w_6 = \text{defence}$
- $w_7 = \text{performance}$
- $w_8 = \text{field}$

$$D = |V| = 8$$
Bernoulli doc. model – example (cont.)

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

\[
B^{\text{Sport}} = \begin{pmatrix}
1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1
\end{pmatrix}
\]

\[
B^{\text{Inf}} = \begin{pmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0
\end{pmatrix}
\]

**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: \( b_1 = [1 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 1] \)

Priors, Likelihoods: 
\[ P(S) = \frac{6}{11}, \quad P(I) = \frac{5}{11} \]
\[ (P(w_t | S)) = (\frac{3}{6} \ \frac{1}{6} \ \frac{2}{6} \ \frac{3}{6} \ \frac{3}{6} \ \frac{4}{6} \ \frac{4}{6} \ \frac{4}{6}) \]
\[ (P(w_t | I)) = (\frac{1}{5} \ \frac{3}{5} \ \frac{3}{5} \ \frac{1}{5} \ \frac{1}{5} \ \frac{1}{5} \ \frac{3}{5} \ \frac{1}{5}) \]

Posterior probabilities:
\[
P(S | b_1) \propto P(S) \prod_{t=1}^{8} [b_{1t}P(w_t | S) + (1 - b_{1t})(1 - P(w_t | S))] \\
\propto \frac{6}{11} \left( \frac{1}{2} \times \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \right) = \frac{5}{891} = 5.6 \times 10^{-3} \\
\]
\[
P(I | b_1) \propto P(I) \prod_{t=1}^{8} [b_{1t}P(w_t | I) + (1 - b_{1t})(1 - P(w_t | I))] \\
\propto \frac{5}{11} \left( \frac{1}{5} \times \frac{2}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{1}{5} \right) = \frac{8}{859375} = 9.3 \times 10^{-6} \\
\]
\[\Rightarrow\text{ Classify this document as } S.\]
## Summary of the document models

<table>
<thead>
<tr>
<th>Class</th>
<th>Doc</th>
<th>Multinomial doc. model</th>
<th>Bernoulli doc. model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_k$</td>
<td>$D^{(k)}_{1}$</td>
<td>$\mathbf{x}^{(k)}<em>{1}, \ldots, \mathbf{x}^{(k)}</em>{N_k}$</td>
<td>$\mathbf{b}^{(k)}<em>{1}, \ldots, \mathbf{b}^{(k)}</em>{N_k}$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td>$D^{(k)}_{N_k}$</td>
<td>$x^{(k)}<em>{N_k1}, \ldots, x^{(k)}</em>{N_kD}$</td>
<td>$b^{(k)}<em>{N_k1}, \ldots, b^{(k)}</em>{N_kD}$</td>
</tr>
<tr>
<td>$\hat{P}(C_k)$</td>
<td>$\frac{n_k(w_1), \ldots, n_k(w_D)}{N_k} \quad \frac{n_k(w_1), \ldots, n_k(w_D)}{N}$</td>
<td>$\frac{n_k(w_1), \ldots, n_k(w_D)}{N_k} \quad \frac{n_k(w_1), \ldots, n_k(w_D)}{N_k}$</td>
<td></td>
</tr>
<tr>
<td>$\hat{P}(w_t</td>
<td>C_k)$</td>
<td>$\frac{n_k(w_1)}{S_k}, \ldots, \frac{n_k(w_D)}{S_k}$</td>
<td>$\frac{n_k(w_1)}{N_k}, \ldots, \frac{n_k(w_D)}{N_k}$</td>
</tr>
</tbody>
</table>

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

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

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

#### Examples

- **Coin tossing models**
  - Fair coin \((P(H)=P(T)=0.5)\)
    - Generated data sequence: \(H, T, T, H, T, H, H, T, \ldots\)
  - Unfair coin \((P(H)=0.7, P(T)=0.3)\)
    - Generated data sequence: \(T, H, H, H, H, H, T, H, \ldots\)

- **Dice throwing models**
  - Unbiased dice \((P(X) = 1/6, X \in \{1, \ldots, 6\})\)
    - Generated data sequence: \(2, 4, 3, 5, 3, 6, 5, 5, 4, 6, \ldots\)
  - Biased dice \((P(X)) = (0.1, 0.1, 0.1, 0.1, 0.2, 0.4)\)
    - Generated data sequence: \(6, 6, 5, 5, 6, 1, 2, 6, 6, 6, \ldots\)
Generative models (cont.)

- Spam mail generator

\[ \text{P(O|Spam)} \]

\[ o_1 \quad o_2 \quad o_3 \quad o_4 \quad o_5 \quad o_6 \quad o_7 \quad o_8 \quad o_9 \quad \ldots \]
Generative model — Multinomial document model

\[ o_1 \quad o_2 \quad o_3 \quad \ldots \quad o_L \mid c_k \]

\[ c_k \]

\(|V|\)-sided dice

\[ P(w_t \mid c_k) \]
Generative model — Bernoulli document model

Terms in Document $D$ in $C_k$

$C_k$

$P(w_1 | C_k)$ $P(w_2 | C_k)$ $P(w_D | C_k)$

$0/1$ $0/1$ $0/1$

coin coin coin

$b_1$ $b_2$ $\ldots$ $b_D$
Word relative-frequencies of spam emails

# of spam emails: 169

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

\[ P(C_k | x) = \frac{P(x | C_k) P(C_k)}{P(x)} \propto P(x | C_k) P(C_k) \]

Model for observation \( \cdots \) generative model

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

Inf2b Learning and Data: Lecture 7
Text Classification using Naïve Bayes
Smoothing in multinomial document model

- Zero probability problem

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

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

- Smoothing – a ‘trick’ to avoid zero counts:

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

Known as Laplace’s rule of succession or add one smoothing.
## Multinomial vs Bernoulli doc. models

<table>
<thead>
<tr>
<th></th>
<th>Multinomial</th>
<th>Bernoulli</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generative model</strong></td>
<td>draw a words from a multinomial distribution</td>
<td>draw a document from a multi-dimensional Bernoulli distribution</td>
</tr>
<tr>
<td><strong>Document representation</strong></td>
<td>Vector of frequencies</td>
<td>Binary vector</td>
</tr>
<tr>
<td><strong>Multiple occurrences</strong></td>
<td>Taken into account</td>
<td>Ignored</td>
</tr>
<tr>
<td><strong>Document length</strong></td>
<td>Longer docs OK</td>
<td>Best for short docs</td>
</tr>
<tr>
<td><strong>Feature vector dimension</strong></td>
<td>Longer OK</td>
<td>Shorter</td>
</tr>
<tr>
<td><strong>Behaviour with ”the”</strong></td>
<td>$P(”the”</td>
<td>C_k) \approx 0.05$</td>
</tr>
<tr>
<td><strong>Non-occurring words in test doc</strong></td>
<td>do not affect likelihood</td>
<td>affect likelihood</td>
</tr>
</tbody>
</table>
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
Use the Bernoulli model and the Naive Bayes assumption for the following.

Consider the vocabulary $V = \{\text{apple, banana, computer}\}$. 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 $V = \{\text{fish, chip, circuit}\}$. 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.
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, \textit{wizard}, \textit{river}, \textit{star}, and \textit{warp}, with respect to the first page of each book, and the information whether the book was a bestseller or not.

<table>
<thead>
<tr>
<th>Author</th>
<th>wizard</th>
<th>river</th>
<th>star</th>
<th>warp</th>
<th>Bestseller</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Baker</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>Baker</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>Baker</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Clark</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>Clark</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>Clark</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>Yes</td>
</tr>
<tr>
<td>Clark</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>Clark</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
</tr>
<tr>
<td>Clark</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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.
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.
Our first ‘real’ application of Naive Bayes

Two BoW models for documents: Multinomial and Bernoulli

Generative models

Smoothing (Add-one/Laplace smoothing)

Good reference:
See Chapter 13 Text classification & Naive Bayes

As always:
be able to implement, describe, compare and contrast (see Lecture Note)