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

Jan-Mar 2015
Today’s Schedule

1. Text classification
2. Bag-of-words models
3. Bernoulli document model
4. Multinomial 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 $D$, with a fixed set of classes $C = \{c_1, \ldots, c_K\}$

Classify $D$ as the class with the highest posterior probability:

$$P(c_k|D) = \frac{P(D|c_k)P(c_k)}{P(D)} \propto P(D|c_k)P(c_k)$$

- How do we represent $D$?
- How do we estimate $P(D|c_k)$ and $P(c_k)$?
How do we represent $D$?

- A sequence of words: $D = (X_1, X_2, \ldots, X_L)$
  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

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

**Multinomial document model** a document is represented by an integer feature vector, whose elements indicate frequency of corresponding word in the document
BoW models: Bernoulli vs. Multinomial

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>Bernoulli (b_t \in {0, 1})</th>
<th>Multinomial (x_t \in \mathbb{N}_0)</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>1</td>
<td>2</td>
</tr>
<tr>
<td>congratulations</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>draws</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>first</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>lotto</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
<td>4</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>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Training of Bernoulli document model

- **Features:** $b = (b_1, \ldots, b_d)$ : $d = |V|$, i.e. vocabulary *binary vector* of word occurrences in a document

- **Training data set**

<table>
<thead>
<tr>
<th>Class</th>
<th>doc</th>
<th>feature vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>$D^{(1)}_1$</td>
<td>$b^{(1)}<em>1 = (b^{(1)}</em>{11}, \ldots, b^{(1)}_{1d})$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td>$D^{(1)}_{N_1}$</td>
<td>$b^{(1)}<em>{N_1} = (b^{(1)}</em>{N_11}, \ldots, b^{(1)}_{N_1d})$</td>
</tr>
<tr>
<td>$N_1$</td>
<td>$P(C_1) = N_1 / N$</td>
<td>$P(w_t</td>
</tr>
<tr>
<td>$C_2$</td>
<td>$D^{(2)}_1$</td>
<td>$b^{(2)}<em>1 = (b^{(2)}</em>{11}, \ldots, b^{(2)}_{1d})$</td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td>$D^{(2)}_{N_2}$</td>
<td>$b^{(2)}<em>{N_2} = (b^{(2)}</em>{N_21}, \ldots, b^{(2)}_{N_2d})$</td>
</tr>
<tr>
<td>$N_2$</td>
<td>$P(C_2) = N_2 / N$</td>
<td>$P(w_t</td>
</tr>
</tbody>
</table>
Classification with Bernoulli document model

• Training (Parameter estimation)
  
  Prior: \( P(c_k) \approx \frac{N_k}{N} \)
  
  Likelihoods: \( P(w_t \mid c_k) \approx \frac{n_k(w_t)}{N_k} \) (fraction of class \( k \) docs with word \( w_t \), \( t = 1, \ldots, d \))

• Classify new document \( D \), feature vector: \( b = (b_1, \ldots, b_d) \)

\[
P(c_k \mid b) \propto P(c_k) P(b \mid c_k)
\]

\[
P(b \mid c_k) = \prod_{t=1}^{d} P(b_t \mid c_k) = \prod_{t=1}^{d} [b_t P(w_t \mid c_k) + (1-b_t)(1-P(w_t \mid c_k))]
\]

\[
= \prod_{t=1}^{d} P(w_t \mid c_k)^{b_t} (1-P(w_t \mid c_k))^{1-b_t}
\]
Classify documents as Sports (S) or Informatics (I)

**Vocabulary** \( V \):

\[
\begin{align*}
w_1 &= \text{goal} \\
w_2 &= \text{tutor} \\
w_3 &= \text{variance} \\
w_4 &= \text{speed} \\
w_5 &= \text{drink} \\
w_6 &= \text{defence} \\
w_7 &= \text{performance} \\
w_8 &= \text{field}
\end{align*}
\]

\[ d = |V| = 8 \]
### Bernoulli doc. model – example (cont.)

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

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

**Estimating priors and likelihoods:**

\[
P(S) = \frac{6}{11}, \quad P(I) = \frac{5}{11}
\]

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

\[
(P(w_t|I)) = \begin{pmatrix}
1/5 & 3/5 & 3/5 & 1/5 & 1/5 & 1/5 & 3/5 & 1/5
\end{pmatrix}
\]
Test documents: $b_1 = [1, 0, 0, 1, 1, 1, 0, 1]$

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

$(P(w_t|S)) = (3/6 \ 1/6 \ 2/6 \ 3/6 \ 3/6 \ 4/6 \ 4/6 \ 4/6)$

$(P(w_t|I)) = (1/5 \ 3/5 \ 3/5 \ 1/5 \ 1/5 \ 1/5 \ 3/5 \ 1/5)$

Posterior probabilities:

$P(S|b_1) \propto P(S) \prod_{t=1}^{8} [b_{1t}P(w_t|S) + (1-b_{1t})(1-P(w_t|S))]$

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

$P(I|b_1) \propto P(I) \prod_{t=1}^{8} [b_{1t}P(w_t|I) + (1-b_{1t})(1-P(w_t|I))]$

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

⇒ Classify this document as $S$. 
Text classification

Bag-of-words models

Bernoulli document model

Multinomial document model

Generative models

Zero Probability Problem
### Training of multinomial document model

- **Features:** \( x = (x_1, \ldots, x_d) : \textit{word frequencies} \) in a doc.
- **Training data set**

<table>
<thead>
<tr>
<th>Class</th>
<th>doc</th>
<th>feature vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>( D^{(1)}_1 )</td>
<td>( \mathbf{x}^{(1)}<em>1 = (x^{(1)}</em>{11}, \ldots, b^{(1)}_{1d}) )</td>
</tr>
<tr>
<td></td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td></td>
<td>( D^{(1)}_{N_1} )</td>
<td>( \mathbf{x}^{(1)}<em>{N_1} = (x^{(1)}</em>{N_11}, \ldots, x^{(1)}_{N_1d}) )</td>
</tr>
<tr>
<td>( N_1 )</td>
<td>( P(C_1) = N_1 / N )</td>
<td>( s_1(w_1), \ldots, s_1(w_d) )</td>
</tr>
<tr>
<td></td>
<td>( P(w_t</td>
<td>C_1) : s_1(w_1)/S_1, \ldots, s_1(w_d)/S_1 )</td>
</tr>
</tbody>
</table>

| \( C_2 \) | \( D^{(2)}_1 \) | \( \mathbf{x}^{(2)}_1 = (x^{(2)}_{11}, \ldots, x^{(2)}_{1d}) \) |
|       | \( \vdots \) | \( \vdots \) |
|       | \( D^{(2)}_{N_2} \) | \( \mathbf{x}^{(2)}_{N_2} = (x^{(2)}_{N_21}, \ldots, x^{(2)}_{N_2d}) \) |
| \( N_2 \) | \( P(C_2) = N_2 / N \) | \( s_2(w_1), \ldots, s_2(w_d) \) |
|       | \( P(w_t|C_2) : s_2(w_1)/S_2, \ldots, s_2(w_d)/S_2 \) |

### Calculation of \( S_i \)

\[
S_i = \sum_{t=1}^{d} s_i(w_t)
\]
Data:
\[ x_{it} : \text{the count of the number of times } w_t \text{ occurs in document } i \]
\[ z_{ik} = 1 \text{ if document } i \text{ is of class } k, 0 \text{ otherwise} \]

Training (Parameter estimation):

**Priors:**
\[ P(c_k) \approx \frac{N_k}{N} \]

**Likelihoods:**
\[ P(w_t | c_k) \approx \frac{\sum_{i=1}^{N} x_{it} z_{ik}}{\sum_{t'=1}^{d} \sum_{i=1}^{N} x_{it'} z_{ik}} = \frac{s_k(w_t)}{\sum_{t'=1}^{d} s_k(w_{t'})} \]

Classify new document \( D \), feature vector: \( x = (x_1, \ldots, x_d) \):

\[ P(c_k | x) \propto P(c_k) P(x | c_k) \]

\[ P(x | c_k) \propto \prod_{t=1}^{d} P(w_t | c_k)^{x_t} \]

NB: \( P( )^0 = 1 \)
Assume a test document $D$ is given as a sequence of words:

$$(o_1, o_2, \ldots, o_L)$$

$L$: the length of sequence, $o_i \in V$

Feature vector: $x = (x_1, \ldots, x_d)$ \(\cdots\) word frequencies

$$P(x | c_k) \propto \prod_{t=1}^{d} P(w_t | c_k)^{x_t}$$

$$\propto P(o_1 | c_k) P(o_2 | c_k) \cdots P(o_L | c_k)$$

$$\propto \prod_{i=1}^{L} P(o_i | c_k)$$

See Q1 in Tutorial 6 for an example.
Multinomial distribution

\[ x = (x_1, \ldots, x_{|V|}) \]

\[ P(x \mid c_k) \propto \prod_{t=1}^{|V|} P(w_t \mid c_k)^{x_t} \]

To be more specific,

\[ P(x \mid c_k) = \frac{n!}{\prod_{t=1}^{|V|} x_t!} \prod_{t=1}^{|V|} P(w_t \mid c_k)^{x_t} \]

where \( n = \sum_{t=1}^{|V|} x_t \), i.e. the total number of words in the document.
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.
1. Text classification

2. Bag-of-words models

3. Bernoulli document model

4. Multinomial document model

5. Generative models

6. Zero Probability Problem
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

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

*Congratulations to you as we bring to your notice, ...*
Generative model — Bernoulli document model

Terms in Document D in $C_k$

$\begin{array}{cccc}
  b_1 & b_2 & \ldots & b_d \\
\end{array}$

$C_k$

$\begin{array}{ccc}
  \text{coin} & \text{coin} & \ldots & \text{coin} \\
  0/1 & 0/1 & \ldots & 0/1 \\
\end{array}$

$\begin{array}{ccc}
  P(w_1 | c_k) & P(w_2 | c_k) & \ldots & P(w_d | c_k) \\
\end{array}$
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

\[ P(w_t \mid c_k) \]
Word relative-frequencies of spam emails

# of spam emails: 169

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>to</td>
<td>0.0395032</td>
<td>from</td>
<td>0.00664282</td>
<td>http</td>
<td>0.00369482</td>
</tr>
<tr>
<td>the</td>
<td>0.0383633</td>
<td>content</td>
<td>0.00644629</td>
<td>money</td>
<td>0.00345898</td>
</tr>
<tr>
<td>you</td>
<td>0.0267285</td>
<td>have</td>
<td>0.0059353</td>
<td>by</td>
<td>0.00338037</td>
</tr>
<tr>
<td>of</td>
<td>0.0257851</td>
<td>bank</td>
<td>0.0059353</td>
<td>or</td>
<td>0.00330176</td>
</tr>
<tr>
<td>and</td>
<td>0.0252349</td>
<td>usd</td>
<td>0.00581738</td>
<td>name</td>
<td>0.00322314</td>
</tr>
<tr>
<td>your</td>
<td>0.0222476</td>
<td>on</td>
<td>0.00554223</td>
<td>funds</td>
<td>0.00322314</td>
</tr>
<tr>
<td>in</td>
<td>0.0200857</td>
<td>we</td>
<td>0.00542432</td>
<td>was</td>
<td>0.00318384</td>
</tr>
<tr>
<td>i</td>
<td>0.0198892</td>
<td>it</td>
<td>0.00518848</td>
<td>type</td>
<td>0.00318384</td>
</tr>
<tr>
<td>this</td>
<td>0.0145828</td>
<td>are</td>
<td>0.00507056</td>
<td>s</td>
<td>0.00318384</td>
</tr>
<tr>
<td>a</td>
<td>0.0138752</td>
<td>transfer</td>
<td>0.00479541</td>
<td>0a</td>
<td>0.00314453</td>
</tr>
<tr>
<td>my</td>
<td>0.0132463</td>
<td>our</td>
<td>0.0047561</td>
<td>if</td>
<td>0.00310522</td>
</tr>
<tr>
<td>for</td>
<td>0.0132463</td>
<td>com</td>
<td>0.00467749</td>
<td>1</td>
<td>0.00310522</td>
</tr>
<tr>
<td>is</td>
<td>0.0112024</td>
<td>am</td>
<td>0.00467749</td>
<td>can</td>
<td>0.00306592</td>
</tr>
<tr>
<td>3d</td>
<td>0.0108879</td>
<td>account</td>
<td>0.00455957</td>
<td>payment</td>
<td>0.002948</td>
</tr>
<tr>
<td>with</td>
<td>0.00915845</td>
<td>unlocked</td>
<td>0.00424512</td>
<td>message</td>
<td>0.002948</td>
</tr>
<tr>
<td>will</td>
<td>0.00876538</td>
<td>20</td>
<td>0.0041665</td>
<td>address</td>
<td>0.00286938</td>
</tr>
<tr>
<td>that</td>
<td>0.00849023</td>
<td>email</td>
<td>0.00404858</td>
<td>us</td>
<td>0.00283008</td>
</tr>
<tr>
<td>as</td>
<td>0.00797925</td>
<td>please</td>
<td>0.00385205</td>
<td>his</td>
<td>0.00279077</td>
</tr>
<tr>
<td>me</td>
<td>0.00766479</td>
<td>not</td>
<td>0.00377344</td>
<td>contact</td>
<td>0.00279077</td>
</tr>
<tr>
<td>be</td>
<td>0.00703589</td>
<td>all</td>
<td>0.00377344</td>
<td>has</td>
<td>0.00271216</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 • • • generative model

\[
P(x) = \sum_{k=1}^{K} P(x | c_k) P(c_k)
\]
Smoothing in multinomial document model

- Zero probability problem

\[ P(\mathbf{x} \mid c_k) \propto \prod_{t=1}^{d} P(w_t \mid c_k)^{x_t} \]

\[ P(w_t \mid c_k) \approx \frac{\sum_{i=1}^{N} x_{it}z_{ik}}{\sum_{|V|} \sum_{i=1}^{N} x_{it'}z_{ik}} = \frac{s_k(w_t)}{\sum_{t' = 1}^{d} s_k(w_{t'})} \]

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

\[ P(w_t \mid c_k) \approx \frac{1 + \sum_{i=1}^{N} x_{it}z_{ik}}{|V| + \sum_{t' = 1}^{d} \sum_{i=1}^{N} x_{it'}z_{ik}} = \frac{1 + s_k(w_t)}{d + \sum_{t' = 1}^{d} s_k(w_{t'})} \]

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

- See Q2 in Tutorial 6.
## Multinomial vs Bernoulli doc. models

<table>
<thead>
<tr>
<th></th>
<th>Bernoulli</th>
<th>Multinomial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generative model</strong></td>
<td>draw a document from a multidimensional Bernoulli distribution</td>
<td>draw a words from a multinomial distribution</td>
</tr>
<tr>
<td><strong>Document representation</strong></td>
<td>Binary vector</td>
<td>Vector of frequencies</td>
</tr>
<tr>
<td><strong>Multiple occurrences</strong></td>
<td>Ignored</td>
<td>Taken into account</td>
</tr>
<tr>
<td><strong>Document length</strong></td>
<td>Best for short docs</td>
<td>Longer docs OK</td>
</tr>
<tr>
<td><strong>Feature vector dimension</strong></td>
<td>Shorter</td>
<td>Longer OK</td>
</tr>
<tr>
<td><strong>Behaviour with ”the”</strong></td>
<td>$P(&quot;the&quot;</td>
<td>c_k) \approx 1.0$</td>
</tr>
<tr>
<td><strong>Non-occurring words</strong></td>
<td>affect likelihood</td>
<td>do not 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
Our first ‘real’ application of Naive Bayes

Two BoW models for documents: Bernoulli and Multinomial

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)