Today's Schedule

- Text classification
- Bag-of-words models
- Multinomial document model
- Bernoulli document model
- Generative models
- Zero Probability Problem

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.

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_i ∈ V)</th>
<th>Multinomial (x_i ∈ N_0)</th>
<th>Bernoulli (b_i ∈ {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>0</td>
<td>1</td>
</tr>
<tr>
<td>category</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>congratulations</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>draws</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>first</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>lotto</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>true</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>you</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

D = 12

x = (1, 0, 1, 1, ... , 1)
b = (0, 0, 1, 1, ... , 1)

Notation for document model

- Training documents:
  - Class Documents
    - C_1 : D_1^{(1)} ... D_n^{(1)}
    - ... ...
    - C_K : D_1^{(K)} ... D_n^{(K)}
  - Flattened representation of training data:
    - Documents : D_1 ... D_n
    - Class indicator : z_1 ... z_n
      - z_k = 1 if D_i belongs to class C_k
      - 0 otherwise
  - Test document : D

How do we represent D?

- A sequence of words: D = (X_1, X_2, ... , X_n)
- 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, ... , x_k) x_k ∈ 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, ... , b_k) b_k ∈ {0, 1}
### Classification with multinomial document model

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: $x = (x_1, \ldots, x_D)$  
$\quad \sum_{i=1}^{D} x_i = n$

Document likelihood with multinomial distribution:

$$P(x | C_k) = \prod_{i=1}^{D} P(w_i | C_k)^{x_i}$$

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

$$P(C_k | x) \propto \sum_{C_k} P(x | C_k)$$

### Discrete probability distributions - review

#### Bernoulli distribution

Eg: Tossing a biased coin ($P(H) = p$), the probability of

$k \in \{0, 1\}$: Tail is:

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

#### Binomial distribution

Eg: Tossing a biased coin $n$ times, the probability of

$\quad k = \binom{n}{k}$  
$\quad \text{of observing Head is times}:

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

#### Multinomial distribution

Eg: Tossing a biased dice $n$ times, the probability of

$\quad k = \binom{n}{k}$  
$\quad \text{of getting face}:

$$P(k) = \binom{n}{k} p_1^{k_1} p_2^{k_2} p_3^{k_3} p_4^{k_4}$$

### Training of multinomial document model

Features: $b = (b_1, \ldots, b_D)$  
$\quad D = |V|$, i.e. vocabulary

$$\text{binary vector of word occurrences in a document}$$

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

#### Classification with Bernoulli document model

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

$$\text{Document likelihood with (multivariate) Bernoulli distribution:}$$

$$P(b, C_k) = \prod_{i=1}^{D} \left( \binom{b_i}{1} P(w_i | C_k) + (1-b_i)(1-P(w_i | C_k)) \right)$$

$$P(w_i | C_k) = \frac{n_i}{N_k}$$

$\quad \text{of class } k \text{ docs with word } w_i$

$$P(C_k | b) \propto \prod_{i=1}^{D} P(w_i | C_k)^{b_i}$$

### Training of Bernoulli document model

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

$$\text{Training data set}$$

$$P(C_k) = \frac{n_k}{N}$$

$$P(w_i | C_k) = \frac{n_i}{N_k}$$

### Bernoulli doc. model – example

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

#### Vocabulary $V$:

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

$\quad D = |V| = 8$

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

Training data:

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1
\end{bmatrix}$$

#### Prior:

$$P(S) = 6/11$$
$$P(I) = 5/11$$

$$P(w_i | S) = \{3/6, 1/6, 2/6, 3/6, 3/6, 4/6, 4/6, 4/6\}$$

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

Evaluating priors and likelihoods:

$$P(S) = 6/11$$
$$P(I) = 5/11$$

$$P(w_i | S) = \{3/6, 1/6, 2/6, 3/6, 3/6, 4/6, 4/6, 4/6\}$$

$$P(w_i | I) = \{1/5, 2/5, 3/5, 1/5, 1/5, 2/5, 3/5, 1/5\}$$

Priors, Likelihoods:

Eg: Tossing a biased dice $n$ times, the probability of

$$P(k) = \binom{n}{k} p_1^{k_1} p_2^{k_2} p_3^{k_3} p_4^{k_4}$$

$$P(C_k | b) \propto \prod_{i=1}^{D} P(w_i | C_k)^{b_i}$$

#### Posterior probabilities:

$$P(S | b_1) \times P(S)$$

$$P(I | b_1) \times P(I)$$

$$P$$

$$\text{Classify this document as } S.$$

$$\text{Test documents: } b_1 = [1 0 0 1 1 1 0 0]$$

$$\text{Observations:}$$

$$b_1 = [1 0 0 1 1 1 0 0]$$

$$\text{Priors, Likelihoods:}$$

$$P(S) = 6/11$$

$$P(w_i | S) = \{3/6, 1/6, 2/6, 3/6, 3/6, 4/6, 4/6, 4/6\}$$

$$P(w_i | I) = \{1/5, 2/5, 3/5, 1/5, 1/5, 2/5, 3/5, 1/5\}$$

$$P(C_k | b) \propto \prod_{i=1}^{D} P(w_i | C_k)^{b_i}$$

$$P$$

$$\text{Classify this document as } S.$$
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.

Word relative-frequencies of spam emails

<table>
<thead>
<tr>
<th># of spam emails: 169</th>
</tr>
</thead>
<tbody>
<tr>
<td>to 0.0383633</td>
</tr>
<tr>
<td>the 0.0383633</td>
</tr>
<tr>
<td>0.0267285</td>
</tr>
<tr>
<td>of 0.0257851</td>
</tr>
<tr>
<td>0.0253249</td>
</tr>
<tr>
<td>you 0.0224767</td>
</tr>
<tr>
<td>0.0093536</td>
</tr>
<tr>
<td>bank 0.0093536</td>
</tr>
<tr>
<td>0.0081738</td>
</tr>
<tr>
<td>usd 0.006423</td>
</tr>
<tr>
<td>0.005423</td>
</tr>
<tr>
<td>in 0.0041848</td>
</tr>
<tr>
<td>this 0.0041848</td>
</tr>
<tr>
<td>0.0030956</td>
</tr>
<tr>
<td>a 0.0030956</td>
</tr>
<tr>
<td>my 0.0030956</td>
</tr>
<tr>
<td>for 0.0030956</td>
</tr>
<tr>
<td>is 0.0030956</td>
</tr>
<tr>
<td>0.0030956</td>
</tr>
<tr>
<td>3d 0.0030956</td>
</tr>
<tr>
<td>with 0.0030956</td>
</tr>
<tr>
<td>0.0030956</td>
</tr>
<tr>
<td>that 0.0030956</td>
</tr>
<tr>
<td>0.0030956</td>
</tr>
<tr>
<td>0.0030956</td>
</tr>
<tr>
<td>0.0030956</td>
</tr>
</tbody>
</table>

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 a 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 country is 2010 very below i and now until html of position http here of mail following there but the by for your willing

Model for classification:
\[ P(C_k \mid x) = \frac{P(x \mid C_k) P(C_k)}{P(x)} \]

Model for observation - generative model
\[ P(x) = \sum_{k=1}^{K} P(x \mid C_k) P(C_k) \]

Models that generate observable data randomly based on a distribution

Examples
- Coin tossing models
- Dice tossing models

Unbiased dice \( P(X = \frac{1}{6}, X \in \{1, \ldots, 6\}) = \frac{1}{6} \)
Biased dice \( P(X) = (0.1, 0.1, 0.1, 0.1, 0.2, 0.4) \)
Text Classification using Naive Bayes

Smoothing in multinomial document model

- Zero probability problem
  \[ P(x | C_k) \propto \prod_{i=1}^{D} P(w_i | C_k)^{n_i} \]
  \[ P(w_i | C_k) = \frac{n_i(w_i)}{\sum_{j=1}^{V} n_j(w_i)} = \frac{n_k(w_i)}{\sum_{j=1}^{V} n_j(w_i)} \]

- Smoothing - a 'trick' to avoid zero counts:
  \[ P(w_i | C_k) = \frac{1 + \sum_{j=1}^{V} n_j(w_i)}{D + \sum_{j=1}^{V} n_j(w_i)} = \frac{1 + n_k(w_i)}{D + \sum_{k=1}^{K} n_k(w_i)} \]
  Known as Laplace’s rule of succession or add one smoothing.

Multinomial vs Bernoulli doc. models

<table>
<thead>
<tr>
<th>Generative model</th>
<th>Multinomial</th>
<th>Bernoulli</th>
</tr>
</thead>
<tbody>
<tr>
<td>draw a words from a multinomial distribution</td>
<td>draw a document from a multi-dimensional Bernoulli distribution</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document representation</th>
<th>Vector of frequencies</th>
<th>Binary vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple occurrences</td>
<td>Taken into account</td>
<td>Ignored</td>
</tr>
<tr>
<td>Document length</td>
<td>Longer docs OK</td>
<td>Best for short docs</td>
</tr>
<tr>
<td>Feature vector dimension</td>
<td>Longer OK</td>
<td>Shorter</td>
</tr>
<tr>
<td>Behaviour with &quot;the&quot;</td>
<td>P(&quot;the&quot;</td>
<td>C_k) \approx 0.05</td>
</tr>
<tr>
<td>Non-occurring words in test doc</td>
<td>do not affect likelihood</td>
<td>affect likelihood</td>
</tr>
</tbody>
</table>

Exercise 1

1. Write the training data as a matrix for each class, where each row corresponds to a training document.

   Consider two test documents:
   No wizard river star warp
   No 1 1 1 2
   0 0 2 1
   35
   32
   0 2 1 2

   Estimate the parameters of a multinomial model for the two classes, and hence classify the document.

   Consider two test documents:
   Yes 1 1 0 0
   Yes 0 1 2 2
   33
   36
   Yes
   No
   1 1 0 0

   Estimate the prior probabilities from the training data.

   With reference to the test documents in the previous question, estimate the posterior probability 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 \( V = \{ \text{apple, banana, computer} \} \). 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:
   - \( \text{apple(2); banana(1); computer(0)} \)
   - \( \text{apple(0); banana(2); computer(0)} \)
   - \( \text{apple(0); banana(0); computer(0)} \)
   - \( \text{apple(1); banana(0); computer(0)} \)

   There are also four training documents in class \( E \):
   - \( \text{apple(2); banana(0); computer(0)} \)
   - \( \text{apple(0); banana(0); computer(1)} \)
   - \( \text{apple(3); banana(1); computer(2)} \)
   - \( \text{apple(0); banana(0); computer(1)} \)

   Exercise 3

   Consider two writers, Baker and Clark, who were twins, and who published four and six childrens books, respectively. The following table shows the frequencies of four words, \( \text{wizard, river, star, and warp} \), with respect to the first page of each book, and the information whether the book was a bestseller or not.

<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>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Baker</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>Baker</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Clark</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</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>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>Clark</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>No</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 its 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.

<table>
<thead>
<tr>
<th>Author</th>
<th>River</th>
<th>Start</th>
<th>Warp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc 2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Author</th>
<th>River</th>
<th>Start</th>
<th>Warp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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