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 2014

Identifying Spam

Spam?

I got your contact information from your countrys 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 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
  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

Bog-of-Words models

Decument: 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	Bernoulli	Multinomial
A	1	1
AM	0	0
ARE	1	1
⋮	⋮	⋮
CAN	0	0
CASINO	1	1
CATEGORY	1	2
⋮	⋮	⋮
THE	1	4
TO	1	3
WINNER	1	1
YOU	1	3
YOUR	1	1

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

Estimating the terms: (non-Bayesian)

Priors: $P(C = c_k) \approx \frac{N_k}{N}$ $\quad (N = \sum_k N_k)$

Likelihoods: assume $P(\mathbf{x} \mid c_k) = \prod_{i=1}^{d} P(x_i \mid c_k)$ (the naive bit)

$$\approx \prod_i \frac{n_{k,i}(x_i)}{N_k}$$

Bayesian class estimation:

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

Generative models for classification

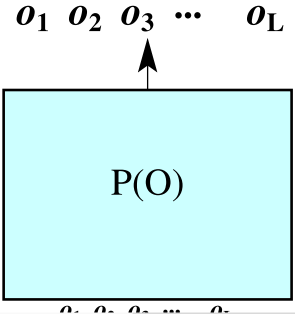
Model for classification

$$P(c_k \,|\, \mathbf{x}) \;=\; \frac{P(\mathbf{x} \,|\, c_k)\, P(c_k)}{P(\mathbf{x})} \;\propto\; P(\mathbf{x} \,|\, c_k)\, P(c_k)$$

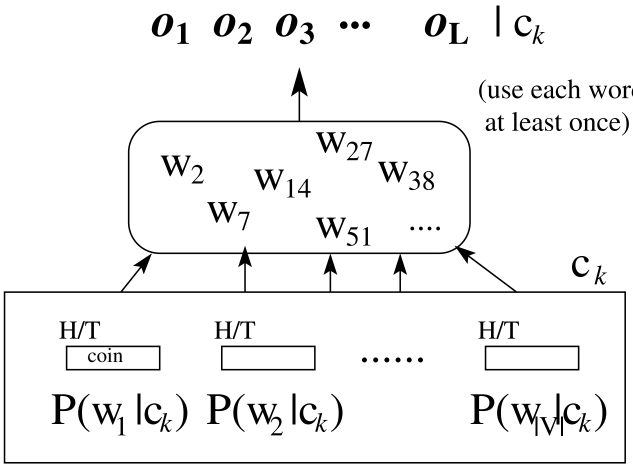
Model for observation $\cdots$ generative model

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

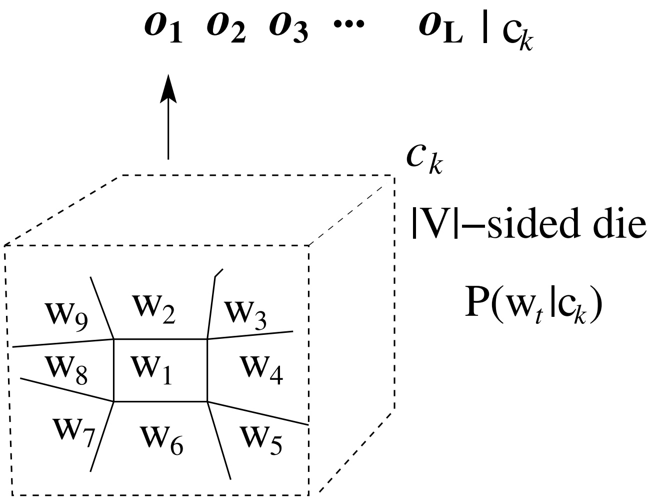
Congratulations to you as we bring to your notice, .



Generative model — Bernoulli document model



Generative model — Multinomial document model



Bernoulli document model

Features: $\mathbf{x} = (x_1, \ldots, x_{|V|})$: length $|V|$ *binary vector* of word occurrences

True generative process:

$\mathbf{x} \leftarrow$ vector of zeros
Human writes email
when tth word used, set $x_t \leftarrow 1$

Model's generative process:

for $t = 1$ to $|V|$:
  Spin biased coin t
  if heads: $x_t \leftarrow 1$ **else**: $x_t \leftarrow 0$

Classification with Bernoulli document model

Training Data:

matrix **B**, document i feature vector: $\mathbf{B}_i$

presence of word t in document i: B_{it}

Parameter estimation:

Priors: $P(c_k) \approx \frac{N_k}{N}$

Likelihoods: $P(w_t \mid c_k) \approx \frac{n_k(w_t)}{N_k}$ (fraction of class k docs with word w_t)

Classify new document D, feature vector: b

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

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

$$P(c_k \mid \mathbf{b}) \propto P(c_k)\, P(\mathbf{b} \mid c_k)$$

Example

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

Vocabulary V**:**

w_1 = goal
w_2 = tutor
w_3 = variance
w_4 = speed
w_5 = drink
w_6 = defence
w_7 = performance
w_8 = field

Example

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

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

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

Estimating priors and likelihoods:

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

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

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

Example (cont.)

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

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 probabilites:

$$P(S\,|\,\mathbf{b}_1) \propto P(S) \prod_{t=1}^{8} [b_{1t} P(w_t\,|\,S) + (1 - b_{1t})(1 - P(w_t\,|\,S))]$$
$$\propto \frac{6}{11}\left(\frac{1}{2} \times \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3}\right) = \frac{5}{891} = 5.6 \times 10^{-3}$$

$$P(I\,|\,\mathbf{b}_1) \propto P(I) \prod_{t=1}^{8} [b_{1t} P(w_t\,|\,I) + (1 - b_{1t})(1 - P(w_t\,|\,I))]$$
$$\propto \frac{5}{11}\left(\frac{1}{5} \times \frac{2}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{1}{5}\right) = \frac{8}{859375} = 9.3 \times 10^{-6}$$

$\Rightarrow$ Classify this document as S.

Multinomial document model

Features: $\mathbf{x} = (x_1, \ldots, x_{|V|})$: length $|V|$ *integer vector* of word counts

True generative process:

$\mathbf{x} \leftarrow$ vector of zeros
human writes email
whenever t th word used, $x_t \leftarrow x_t + 1$

Model's generative process:

$\mathbf{x} \leftarrow$ vector of zeros
for each word in document:
$t \sim$ biased $|V|$-sided die
$x_t \leftarrow x_t + 1$

Classification with multinomial document model

Data:

x_{it} : the count of the number of times w_t occurs in document i

$z_{ik} = 1$ if document i is of class k, 0 otherwise

Parameter estimation:

Priors: $$P(c_k) \approx \frac{N_k}{N}$$

Likelihoods: $$P(w_t \mid c_k) \approx \frac{\sum_{i=1}^{N} x_{it} z_{ik}}{\sum_{t'=1}^{|V|} \sum_{i=1}^{N} x_{it'} z_{ik}}$$

the relative frequency of w_t in documents of class $C = k$ with respect to the total number of words in documents of that class

Classify new document D, feature vector: x:

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

NB: $P(\ )^0 = 1$

$$P(C \mid \mathbf{x}) \propto P(C)\, P(\mathbf{x} \mid c_k)$$

Classification with multinomial document model

Assume a test document D is given as a sequece of words : $(o_1, o_2, \ldots, o_L)$ and $o_i \in V$.

$$P(\mathbf{x} \mid c_k) \propto \prod_{t=1}^{|V|} P(w_t \mid c_k)^{x_t} = \prod_{i=1}^{L} P(o_i \mid c_k)$$

Multinomial distribution

$$\mathbf{x} = (x_1, \ldots, x_{|V|})$$

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

To be more specific,

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

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.

Smoothing

A 'trick' to avoid zero counts:

$$P(w_t \mid C = k) \approx \frac{1 + \sum_{i=1}^{N} x_{it} z_{ik}}{|V| + \sum_{t'=1}^{|V|} \sum_{i=1}^{N} x_{it'} z_{ik}}$$

Add 'the dictionary' to the training data for each class

Known as Laplace's rule of succession. Commonly used.

Laplace's rule of succession can be derived from a Bayesian viewpoint. The imaginary counts can overwhelm the data for large 'vocabularies'. In later courses you may see more sophisticated smoothing methods.

Which document model should we use, Bernoulli or Multinomial?

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

Summary

Our first 'real' application of Naive Bayes

Two models for documents: Bernoulli and Multinomial

As always:

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

Errata for Lecture Note 7:

Section 6 (page 9), remove Item 6:

6. **Non-occurring words:**
Bernoulli: affect the document probabilities.
Multinomial: do not affect the document probabilities.