Information Retrieval

Weike Pan

The slides are **adapted from those provided by Prof. Hinrich Schütze** at University of Munich (http://www.cis.lmu.de/~hs/teach/14s/ir/).

Chapter 13 Text classification & Naive Bayes

- 13.1 The text classification problem
- 13.2 Naive Bayes text classification
- 13.3 The Bernoulli model
- 13.4 Properties of Naive Bayes
- 13.5 Feature selection
- 13.6 Evaluation of text classification
- 13.7 References and further reading

Outline

- 13.1 The text classification problem
- 13.2 Naive Bayes text classification
- 13.3 The Bernoulli model
- 13.4 Properties of Naive Bayes
- 13.5 Feature selection
- 13.6 Evaluation of text classification
- 13.7 References and further reading

13.1 The text classification problem

Given:

- A document space $\mathbb{X}$
  - Documents are represented in this space – typically some type of high-dimensional space.
- A fixed set of classes $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$
  - The classes are human-defined for the needs of an application (e.g., spam vs. ham).
- A training set $\mathbb{D}$ of labeled documents. Each labeled document $\langle d, c \rangle \in \mathbb{X} \times \mathbb{C}$

Using a learning method or learning algorithm, we then wish to learn a classifier γ that maps documents to classes:

$$\gamma : \mathbb{X} \to \mathbb{C}$$

13.1 The text classification problem

- Examples of text classification:
  - Language identification (English vs. other languages)
  - The automatic detection of spam pages (spam vs. ham)
  - Sentiment detection: is a movie or product review positive or negative (positive vs. negative)
  - Topic-specific or vertical search – restrict search to a “vertical” like “related to health” (relevant to vertical vs. not)

Outline

- 13.1 The text classification problem
- 13.2 Naive Bayes text classification
- 13.3 The Bernoulli model
- 13.4 Properties of Naive Bayes
- 13.5 Feature selection
- 13.6 Evaluation of text classification
- 13.7 References and further reading

13.2 Naive Bayes text classification

- The probability of a document *d* being in a class *c*:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

prior probability

conditional probability w.r.t. the term t_k of the *k*th token and the class c

the number of tokens in the document

- The best class is the most likely or maximum a posteriori (MAP) class:

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

- Sum log probabilities instead of multiplying probabilities (avoid underflow)

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)}$$

Add one to each count to avoid zeros

13.2 Naive Bayes text classification

- Training (训练阶段)

```
TrainMultinomialNB(ℂ, 𝔻)
V ← ExtractVocabulary(𝔻)
N ← CountDocs(𝔻)
for each c ∈ ℂ
do N_c ← CountDocsInClass(𝔻, c)
    prior[c] ← N_c / N
    text_c ← ConcatenateTextOfAllDocsInClass(𝔻, c)
    for each t ∈ V
    do T_ct ← CountTokensOfTerm(text_c, t)
    for each t ∈ V
    do condprob[t][c] ← (T_ct + 1) / Σ_t'(T_ct' + 1)
return V, prior, condprob
```

13.2 Naive Bayes text classification

- Test (测试阶段)

ApplyMultinomialNB($\mathbb{C}$, V, *prior*, *condprob*, d)

$W \leftarrow$ ExtractTokensFromDoc(V, d)

for each $c \in \mathbb{C}$

do $score[c] \leftarrow \log prior[c]$

  for each $t \in W$ → t: token

  do $score[c] += \log condprob[t][c]$

return $\arg\max_{c \in \mathbb{C}} score[c]$

Outline

- 13.1 The text classification problem
- 13.2 Naive Bayes text classification
- 13.3 The Bernoulli model
- 13.4 Properties of Naive Bayes
- 13.5 Feature selection
- 13.6 Evaluation of text classification
- 13.7 References and further reading

Outline

- 13.1 The text classification problem
- 13.2 Naive Bayes text classification
- 13.3 The Bernoulli model
- 13.4 Properties of Naive Bayes
- 13.5 Feature selection
- 13.6 Evaluation of text classification
- 13.7 References and further reading

13.4 Properties of Naive Bayes

- Conditional independence assumption

$$P(d|c) = P(\langle t_1, \ldots, t_{n_d} \rangle | c) = \prod_{1 \le k \le n_d} P(X_k = t_k | c)$$

- Positional independence assumption

- → bag of words model (词袋模型)

Outline

- 13.1 The text classification problem
- 13.2 Naive Bayes text classification
- 13.3 The Bernoulli model
- 13.4 Properties of Naive Bayes
- 13.5 Feature selection
- 13.6 Evaluation of text classification
- 13.7 References and further reading

13.5 Feature selection

```
SelectFeatures(𝔻, c, k)
  V ← ExtractVocabulary(𝔻)
  L ← []
  for each t ∈ V        t: term
  do A(t, c) ← ComputeFeatureUtility(𝔻, t, c)
       Append(L, ⟨A(t, c), t⟩)
  return FeaturesWithLargestValues(L, k)
```

- How do we compute *A(t, c)*, the **feature utility**?
  - E.g., Frequency, mutual information, Chi-square

13.5 Feature selection

- Mutual information (MI)

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U=e_t, C=e_c) \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t)P(C=e_c)}$$

- **"how much information"** the term contains about the class and vice versa
- Notes: when *U* and *C* are **independent**, ***I(U; C) = 0***

13.5 Feature selection

- Calculation (MI)
  - Based on maximum likelihood estimates, the formula we actually use is:

$$
\begin{aligned}
I(U;C) \quad = \quad & \frac{N_{11}}{N}\log_2\frac{N\times N_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{N\times N_{01}}{N_{0.}N_{.1}} \\
& + \frac{N_{10}}{N}\log_2\frac{N\times N_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{N\times N_{00}}{N_{0.}N_{.0}}
\end{aligned}
$$

  - Notes:
    - N_{10}: number of documents that contain t ($e_t = 1$) and are not in c ($e_c = 0$);
    - N_{11}: number of documents that contain t ($e_t = 1$) and are in c ($e_c = 1$);
    - N_{01}: number of documents that do not contain t ($e_t = 0$) and are in c ($e_c = 1$);
    - N_{00}: number of documents that do not contain t ($e_t = 0$) and are not in c ($e_c = 0$);
    - $N = N_{00} + N_{01} + N_{10} + N_{11}$.

13.5 Feature selection

- Example 1 (MI)

c: class

t: term

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{EXPORT}} = 1$	$N_{11} = 49$	$N_{10} = 27652$
$e_t = e_{\text{EXPORT}} = 0$	$N_{01} = 141$	$N_{00} = 774106$

Plug these values into formula:

$$
\begin{aligned}
I(U;C) &= \frac{49}{801948}\log_2\frac{801948\times 49}{(49+27652)(49+141)} \\
&+ \frac{141}{801948}\log_2\frac{801948\times 141}{(141+774106)(49+141)} \\
&+ \frac{27652}{801948}\log_2\frac{801948\times 27652}{(49+27652)(27652+774106)} \\
&+ \frac{774106}{801948}\log_2\frac{801948\times 774106}{(141+774106)(27652+774106)} \\
&\approx 0.000105
\end{aligned}
$$

13.5 Feature selection

- Example 2 (MI)

Class: *coffee*

term	MI
COFFEE	0.0111
BAGS	0.0042
GROWERS	0.0025
KG	0.0019
COLOMBIA	0.0018
BRAZIL	0.0016
EXPORT	0.0014
EXPORTERS	0.0013
EXPORTS	0.0013
CROP	0.0012

Class: *sports*

term	MI
SOCCER	0.0681
CUP	0.0515
MATCH	0.0441
MATCHES	0.0408
PLAYED	0.0388
LEAGUE	0.0386
BEAT	0.0301
GAME	0.0299
GAMES	0.0284
TEAM	0.0264

13.5 Feature selection

observed frequency

expected frequency

- Chi-square χ^2

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

- Calculation

For example, E_{11} is the expected frequency of t and c occurring together in a document assuming that term and class are **independent**.

$$\begin{aligned} E_{11} &= N \times P(t) \times P(c) = N \times \frac{N_{11}+N_{10}}{N} \times \frac{N_{11}+N_{01}}{N} \\ &= N \times \frac{49+141}{N} \times \frac{49+27652}{N} \approx 6.6 \end{aligned}$$

class

term

	$e_{poultry} = 1$		$e_{poultry} = 0$	
$e_{\text{export}} = 1$	$N_{11} = 49$	$E_{11} \approx 6.6$	$N_{10} = 27{,}652$	$E_{10} \approx 27{,}694.4$
$e_{\text{export}} = 0$	$N_{01} = 141$	$E_{01} \approx 183.4$	$N_{00} = 774{,}106$	$E_{00} \approx 774{,}063.6$

13.5 Feature selection

- χ^2 is a measure of how much expected counts E and observed counts N **deviate** from each other.

- A high value of χ^2 indicates that the hypothesis of **independence**, which implies that expected and observed counts are similar, is **incorrect**.

→ 值越大，**t**和**c**的相关性就越大

Summary

- 13.1 The text classification problem
- 13.2 Naive Bayes text classification
- 13.3 The Bernoulli model
- 13.4 Properties of Naive Bayes
- 13.5 Feature selection
- 13.6 Evaluation of text classification
- 13.7 References and further reading