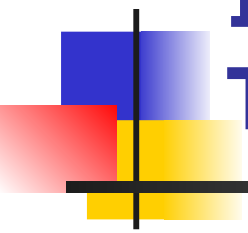
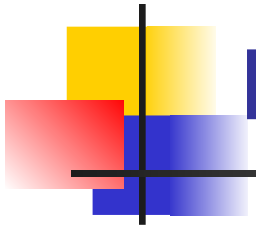


IT4853
Tìm kiếm và trình diễn thông tin

Bài 12. Phân lớp văn bản (2)

IIR.C13. Text classification and Naive Bayes

Bộ môn Hệ thống thông tin
Viện CNTT & TT



Nội dung chính

- **Các mô hình Naïve Bayes**
- **Trích chọn đặc trưng.**

Multinomial Naïve Bayes

- **Huấn luyện:**

TrainMultinomialNB($\mathbb{C}, \mathbb{D}$)

1 $V \leftarrow$ ExtractVocabulary($\mathbb{D}$)
2 $N \leftarrow$ CountDocs($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ CountDocsInClass($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c/N$
6 $text_c \leftarrow$ ConcatenateTextOfAllDocsInClass($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ CountTokensOfTerm($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

Multinomial Naïve Bayes (2)

- **Phân lớp:**

```
ApplyMultinomialNB(ℂ, V, prior, condprob, d)
1  W ← ExtractTokensFromDoc(V, d)
2  for each c ∈ ℂ
3  do score[c] ← log prior[c]
4       for each t ∈ W
5       do score[c] += log condprob[t][c]
6  return arg max_{c∈ℂ} score[c]
```

Bernoulli Naïve Bayes

- **Huấn luyện:**

TRAINBERNOULLINB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c / N$
6 **for each** $t \in V$
7 **do** $N_{ct} \leftarrow$ COUNTDOCSINCLASSCONTAININGTERM($\mathbb{D}, c, t$)
8 $condprob[t][c] \leftarrow (N_{ct} + 1) / (N_c + 2)$
9 **return** $V, prior, condprob$

Bernoulli Naïve Bayes (2)

- **Phân lớp:**

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

1 $V_d \leftarrow$ EXTRACTTERMSFROMDOC(V, d)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in V$
5 **do if** $t \in V_d$
6 **then** $score[c] \mathrel{+}= \log condprob[t][c]$
7 **else** $score[c] \mathrel{+}= \log(1 - condprob[t][c])$
8 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

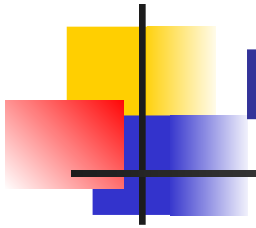
Bernouli NB

▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

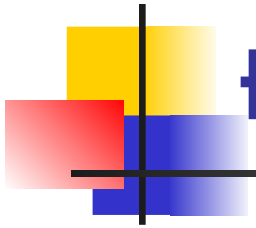
$$\hat{P}(\text{Chinese}|c) = (3+1)/(3+2) = 4/5$$
$$\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokyo}|c) = (0+1)/(3+2) = 1/5$$
$$\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) = (1+1)/(3+2) = 2/5$$
$$\hat{P}(\text{Chinese}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Japan}|\overline{c}) = \hat{P}(\text{Tokyo}|\overline{c}) = (1+1)/(1+2) = 2/3$$
$$\hat{P}(\text{Beijing}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) = \hat{P}(\text{Shanghai}|\overline{c}) = (0+1)/(1+2) = 1/3$$

$$\begin{aligned}\hat{P}(c|d_5) &\propto \hat{P}(c)\cdot\hat{P}(\text{Chinese}|c)\cdot\hat{P}(\text{Japan}|c)\cdot\hat{P}(\text{Tokyo}|c)\\ &\quad\cdot(1-\hat{P}(\text{Beijing}|c))\cdot(1-\hat{P}(\text{Shanghai}|c))\cdot(1-\hat{P}(\text{Macao}|c))\\ &= 3/4\cdot 4/5\cdot 1/5\cdot 1/5\cdot(1-2/5)\cdot(1-2/5)\cdot(1-2/5)\\ &\approx 0.005\end{aligned}$$



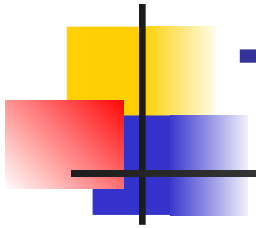
Nội dung chính

- Các mô hình Naïve Bayes;
- **Trích chọn đặc trưng.**



Đặc trưng nhiễu

- **Đặc trưng nhiễu là những đặc trưng mà khi thêm vào văn bản sẽ làm tăng lỗi phân lớp;**
- **Giả sử một từ hiếm t không chứa thông tin liên quan đến lớp c nhưng lại xuất hiện trong các văn bản của lớp c.**
- **Vì *t* là từ hiếm nên bộ phân lớp sau huấn luyện có thể coi *t* như một tín hiệu mạnh để xếp các văn bản chứa t vào lớp c.**
  - **Hiện tượng này được gọi là *overfitting***



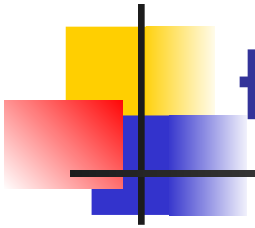
Trích chọn đặc trưng

- **Quá trình loại bỏ các đặc trưng nhiễu gọi là trích chọn đặc trưng:**
  - **Giúp phân lớp chính xác hơn;**
  - **Tăng tốc độ (nhờ giảm khối lượng dữ liệu cần xử lý).**

Giải thuật trích chọn đặc trưng

SELECTFEATURES($\mathbb{D}, c, k$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $L \leftarrow []$
3 **for each** $t \in V$
4 **do** $A(t, c) \leftarrow$ COMPUTEFEATUREUTILITY($\mathbb{D}, t, c$)
5 APPEND($L, \langle A(t, c), t \rangle$)
6 **return** FEATURESWITHLARGESTVALUES(L, k)

How do we compute A, the feature utility?



Độ hữu ích của đặc trưng

- **Độ hữu ích của đặc trưng:**
  - **Tần suất – lựa chọn những từ xuất hiện thường xuyên nhất.**
  - **Hàm lượng thông tin – lựa chọn từ với Hàm lượng thông tin cao nhất;**
  - **χ^2: Chi bình phương**

Hàm lượng thông tin: Mutual Information; Information Gain.

Hàm lượng thông tin

- **Cách tính I:**

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

$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}}$$

N_{11} số văn bản thuộc lớp c chứa t; N_{10} số văn bản chứa t không thuộc lớp c; N_{01} #không chứa t, thuộc lớp c; N_{00} #không thuộc lớp c không chứa t.
$N = N_{11} + N_{10} + N_{01} + N_{00}$ là tổng số văn bản.

Ví dụ tính hàm lượng thông tin, poultry/EXPORT

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{EXPORT}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{EXPORT}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

Plug these values into formula:

$$
\begin{aligned}
I(U;C) &= \frac{49}{801{,}948}\log_2\frac{801{,}948\cdot 49}{(49+27{,}652)(49+141)} \\
&\quad + \frac{141}{801{,}948}\log_2\frac{801{,}948\cdot 141}{(141+774{,}106)(49+141)} \\
&\quad + \frac{27{,}652}{801{,}948}\log_2\frac{801{,}948\cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)} \\
&\quad + \frac{774{,}106}{801{,}948}\log_2\frac{801{,}948\cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)} \\
&\approx 0.000105
\end{aligned}
$$

Kết quả trích chọn đặc trưng trên Reuters

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

Chi bình phương

- **Dùng để đánh giá tính độc lập của hai sự kiện:**
  - **Phân lớp văn bản: sự kiện xuất hiện lớp và sự kiện xuất hiện từ.**
- **Xếp hạng từ theo đại lượng sau:**
  - **Chọn** $$X^2(\mathbb{D},t,c)=\sum_{e_t\in\{0,1\}}\sum_{e_c\in\{0,1\}}\frac{(N_{e_te_c}-E_{e_te_c})^2}{E_{e_te_c}}$$

- **Chi bình phương nhỏ thể hiện mối liên hệ chặt chẽ giữa sự xuất hiện của từ và sự xuất hiện của lớp, thể hiện khả năng từ là một**

Chi bình phương (2)

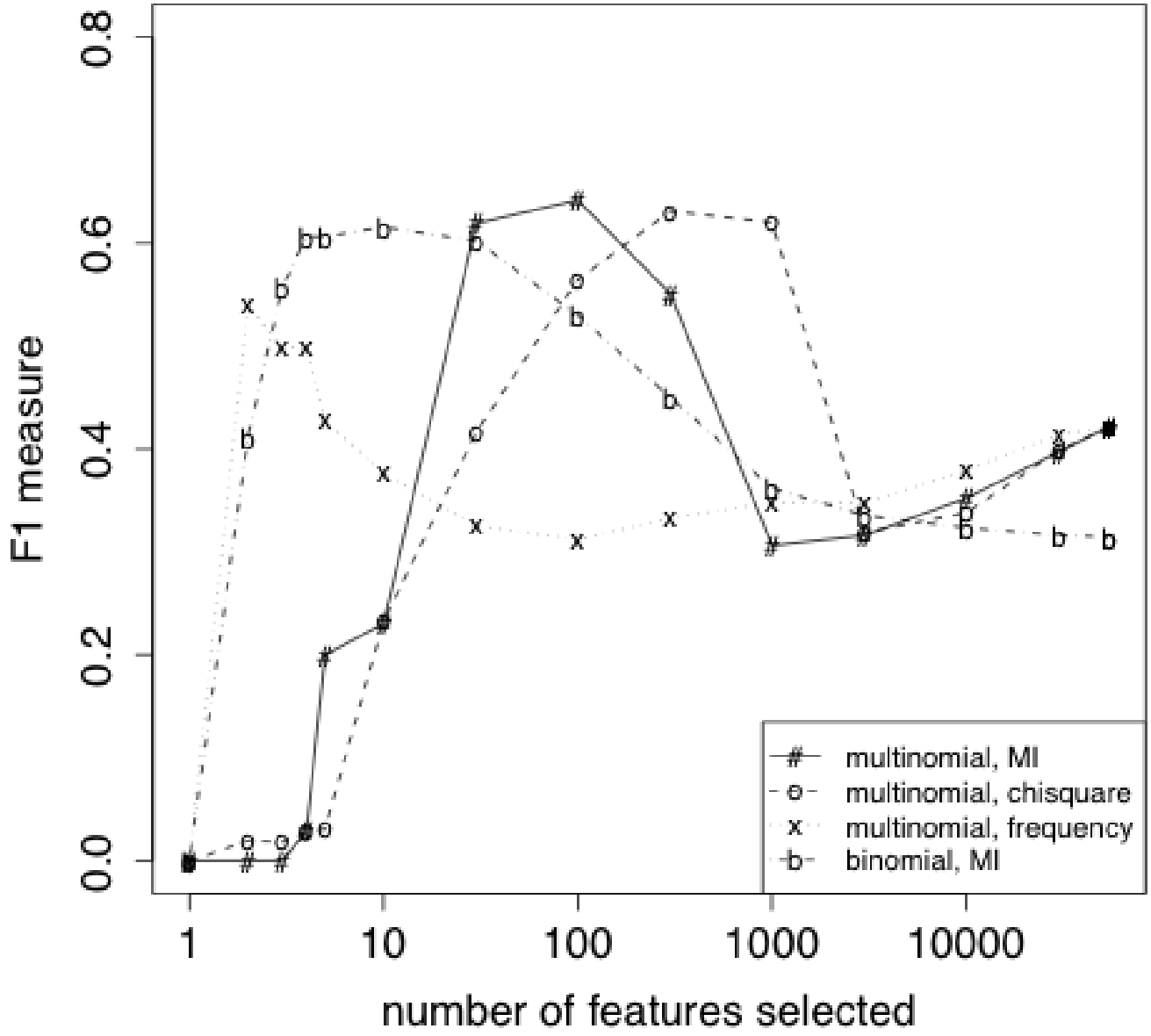
	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$	
$e_t = e_{\text{EXPORT}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$	Plug
$e_t = e_{\text{EXPORT}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$	

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

$$
X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \approx 284
$$

$$
X^2(\mathbb{D}, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11} N_{00} - N_{10} N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}
$$

Hai công thức là tương đương.



(multinomial = multinomial Naive Bayes, binomial = Bernoulli Naive Bayes)

Bài tập 15.1

- **Hãy lập ma trận nhầm lẫn cho cặp "Kyoto/JAPAN", tương tự cặp EXPORT/poultry;**
- **Hãy tính MI cho cặp Kyoto/JAPAN;**
- **Hãy thử thiết lập ma trận nhầm lẫn bất kỳ sao cho MI = 0**

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{EXPORT}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{EXPORT}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

Bộ dữ liệu

	docID	words in document	in c = Japan?
training set	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
	5	London	no

Bài tập 15.2

- Hãy tính $I(U_t, C_c)$ và $X^2(D, t, c)$ trong hai trường hợp:
  - Từ t và lớp c hoàn toàn độc lập;
  - Từ t và lớp c hoàn toàn phụ thuộc.

Bài tập 15.3

Cho dữ liệu thống kê đối với bốn từ của lớp coffee như sau:

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

Hãy lựa chọn hai từ theo Chi-bình phương? và theo MI?

Bài tập 15.4

Đối với các đại lượng trong công thức tính Chi-bình phương, hãy chứng minh:

$$|N_{11} - E_{11}| = |N_{10} - E_{10}| = |N_{01} - E_{01}| = |N_{00} - E_{00}|$$

Bài tập 15.5

	docID	words in document	in $c = China$?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

Trên cơ sở bộ dữ liệu đã cho hãy (i) thiết lập bộ phân lớp Naïve Bayes đã thức (ii) Áp dụng phân lớp văn bản kiểm thử (iii) thiết lập bộ phân lớp Bernoulli (iv) Áp dụng phân lớp văn bản kiểm thử.

Không cần xác định những tham số không dùng đến trong phân lớp.