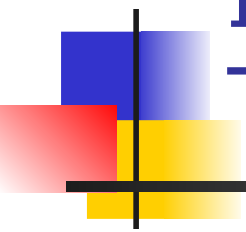
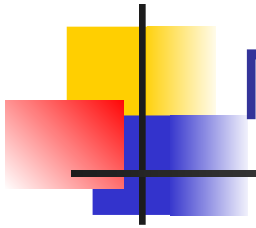


IT4853
Tìm kiếm và trình diễn thông tin

Bài 11. Phân lớp văn bản

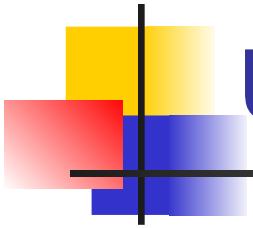
IIR.C13. Text classification and Naive Bayes

Bộ môn Hệ thống thông tin
Viện CNTT & TT



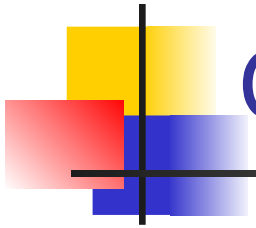
Nội dung chính

- Ứng dụng phân lớp trong tìm kiếm
- Phương pháp Naïve Bayes
- Đánh giá phương pháp phân lớp



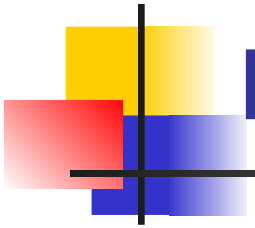
Ứng dụng trong công cụ tìm kiếm

- Xác định ngôn ngữ
  - Các lớp: Tiếng Anh, tiếng Việt, v.v.
- Xác định spam
- Tìm kiếm theo chủ đề
- Truy vấn cố định (standing queries), v.d., Google Alerts
- Phân lớp bình luận: vd., bình luận về phim mang tính khen ngợi hay phê bình, v.v.



Các phương pháp phân lớp

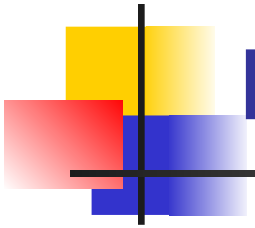
- Theo mức độ tham gia của con người
  - Phân lớp thủ công
    - Người phân lớp, máy hỗ trợ
  - Phân lớp dựa trên luật (bán tự động)
    - Người cung cấp luật, máy phân lớp
  - Xác suất/thống kê (tự động)
    - Người huấn luyện, máy phân lớp



Phương pháp phân lớp thủ công

- Sử dụng ở: Yahoo, ODP, Pubmed;
- Rất chính xác!
- Đơn giản với dữ liệu nhỏ;
- Phức tạp & chi phí cao trên quy mô lớn.

Phân lớp tự động?



Phương pháp phân lớp dựa trên luật

- Ví dụ, Google Alerts;
- Sử dụng môi trường tích hợp hỗ trợ viết luật phân lớp;
  - Thường sử dụng Logic Boolean.
- Có thể đạt độ chính xác rất cao;
- Cần chi phí lớn và khó quản lý.

Ví dụ luật phân lớp

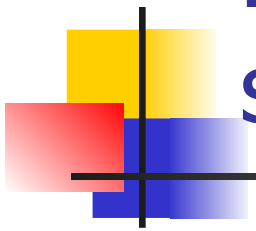
comment line
top-level topic
topic definition modifiers
subtopic topic
evidence topic
topic definition modifier
evidence topic
topic definition modifier
evidence topic
topic definition modifier
evidence topic
topic definition modifier
subtopic

```
# Beginning of art topic definition
art ACCRUE
    /author = "fsmith"
    /date   = "30-Dec-01"
    /annotation = "Topic created
                   by fsmith"
* 0.70 performing-arts ACCRUE
** 0.50 WORD
    /wordtext = ballet
** 0.50 STEM
    /wordtext = dance
** 0.50 WORD
    /wordtext = opera
** 0.30 WORD
    /wordtext = symphony
* 0.70 visual-arts ACCRUE
** 0.50 WORD
    /wordtext = painting
** 0.50 WORD
    /wordtext = sculpture
```

subtopic
subtopic
subtopic

```
* 0.70 film ACCRUE
** 0.50 STEM
    /wordtext = film
** 0.50 motion-picture PHRASE
*** 1.00 WORD
    /wordtext = motion
*** 1.00 WORD
    /wordtext = picture
** 0.50 STEM
    /wordtext = movie
* 0.50 video ACCRUE
** 0.50 STEM
    /wordtext = video
** 0.50 STEM
    /wordtext = vcr
# End of art topic
```

Phương pháp phân lớp dựa trên xác suất/thống kê

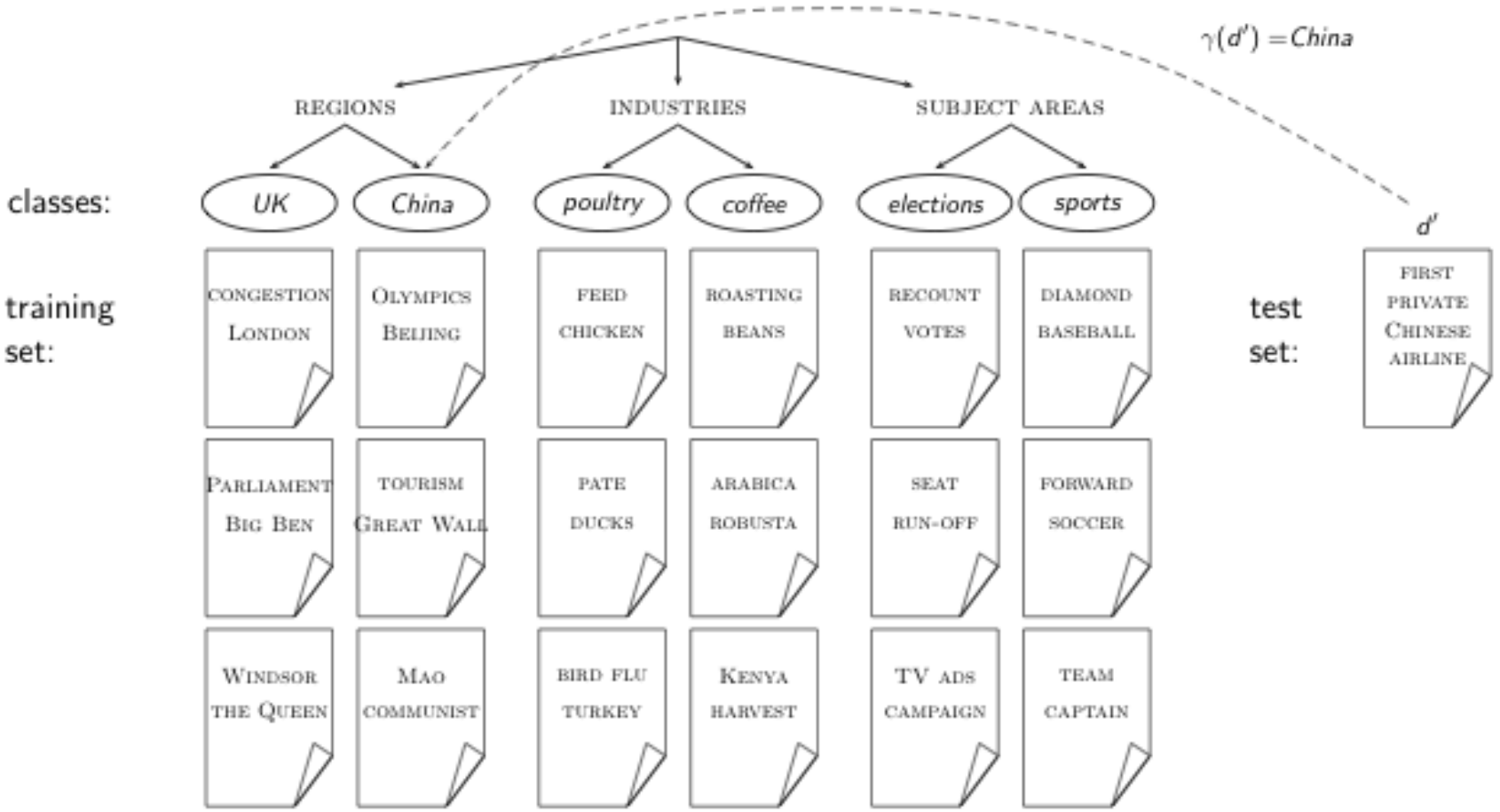


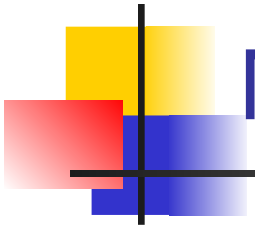
- Bài toán phân lớp văn bản:
  - Huấn luyện: Học có giám sát, nhằm xác định γ;
  - Phân lớp: Sử dụng γ để phân lớp văn bản.
- Tiêu biểu: Naïve Bayes, Rocchio, kNN, SVMs
  - Cần thiết lập bộ dữ liệu huấn luyện;
    - Tuy nhiên không yêu cầu chuyên gia.

Bài toán phân lớp văn bản

- Ký hiệu:
  - D là tập văn bản;
  - C là tập lớp (còn được gọi là tập nhãn).
- Dữ liệu huấn luyện là một phân lớp mẫu
  - TrainingSet= $\{< d, c > | d \in D_{training},\ c \in C\}$, cho biết một số văn bản tiêu biểu thuộc các lớp đã cho.
- **Học:** sử dụng giải thuật huấn luyện để xác định ánh xạ γ gán văn bản với lớp:
  - $\gamma: D_{training} \rightarrow C$
- **Phân lớp:** cho d ∈ X cần xác định γ(d) ∈ C.

Bài toán phân lớp văn bản (2)





Nội dung chính

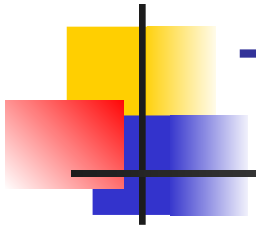
- Ứng dụng phân lớp trong tìm kiếm
- Phương pháp Naïve Bayes
- Đánh giá phương pháp phân lớp

Phân lớp Naïve Bayes

- Phân lớp dựa trên xác suất;
- Xác suất d thuộc c được tính như sau:

$$p(c|d) \propto p(c) \prod_{1 \le k \le n_d} p(t_k|c),$$

- Trong đó:
  - n_d là số lượng từ trong văn bản (độ dài tính theo từ);
  - $p(t_k|c)$ xác suất t_k thuộc c;
  - p(c) là xác suất tiền nghiệm của lớp c.



Tiêu trí xác suất cực đại

- Gán văn bản với lớp có xác suất thuộc cao nhất:

$$\gamma(d) = arg\max_{c \in C} p(c|d)$$

Lấy log

- Tích nhiều đại lượng xác suất nhỏ có thể gây tràn số;
- Kết quả phân lớp không đổi nếu sử dụng logarithm
- Trong thực tế sử dụng công thức sau:

$$\gamma(d) = arg\max_{c \in C}[\log p(c|d)]$$

$$= arg\max_{c \in C}\left[\log p(c) + \sum_{1 \le k \le n_d} \log p(t_k|c)\right]$$

Ước lượng tham số

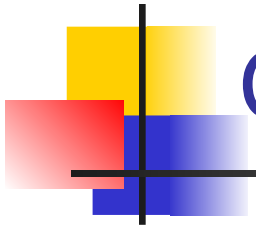
- Xác định p(c) và $p(t_k|c)$ dựa trên dữ liệu luyện:

$$p(c) = \frac{N_c}{N}$$

  - Trong đó N_c là số văn bản của lớp c, N là số văn bản trong bộ dữ liệu luyện
- Xác suất có điều kiện:

$$p(t_k|c) = \frac{cf_{c,t_k}}{\sum_{t \in V} cf_{c,t}}$$

  - Trong đó $cf_{c,t}$ là số lần từ t xuất hiện trong lớp c.



Giá trị 0

- Nếu có một từ t thuộc d nhưng không xuất hiện trong bất kỳ văn bản nào của lớp c thì:
  - $p(t|c) = 0$
  - Kéo theo $p(c|d)=0$.

Giải pháp?

Làm mịn

- Cộng thêm 1:

$$p(t_k|c) = \frac{cf_{c,t_k} + 1}{\sum_{t \in V}(cf_{c,t}+1)} = \frac{cf_{c,t_k} + 1}{\sum_{t \in V} cf_{c,t} + |V|}$$

Giải thuật 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** $\mathsf{cf}_{c,t} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{\mathsf{cf}_{c,t}+1}{\sum_{t'}(\mathsf{cf}_{c,t'}+1)}$
11 **return** $V, prior, condprob$

Giải thuật Naïve Bayes: Phân lớp

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

1 $W \leftarrow \textsc{ExtractTokensFromDoc}(V, d)$
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in W$
5 **do** $score[c] += \log condprob[t][c]$
6 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

Độ phức tạp của Naive Bayes

mode	time complexity
training	$\Theta(\|\mathbb{D}\|L_{\text{ave}} + \|\mathbb{C}\||V|)$
testing	$\Theta(L_{\text{a}} + \|\mathbb{C}\|M_{\text{a}}) = \Theta(\|\mathbb{C}\|M_{\text{a}})$

- L_{ave}: Độ dài trung bình của văn bản luyện, L_a: Độ dài văn bản phân lớp; M_a: Số lượng từ duy nhất trong văn bản phân lớp; D là bộ dữ liệu luyện, V là bộ từ vựng; C là tập lớp.
- Naive Bayes có độ phức tạp tuyến tính so với kích thước dữ liệu luyện và dữ liệu phân lớp. Đây là độ phức tạp tối ưu.

Ví dụ phân lớp Naive Bayes

	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	?

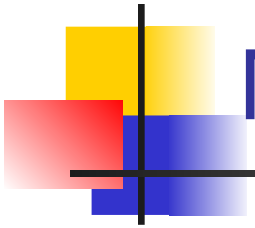
- Ước lượng tham số cho bộ phân lớp Naïve Bayes
- Phân lớp văn bản test (docID = 5)

Ví dụ phân lớp Naive Bayes (2)

Priors: $\hat{P}(c) = 3/4$ and $\hat{P}(\overline{c}) = 1/4$ Conditional probabilities:

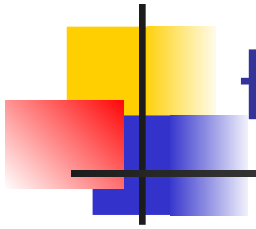
$$\begin{aligned}
\hat{P}(\text{CHINESE}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{TOKYO}|c) = \hat{P}(\text{JAPAN}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{CHINESE}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{TOKYO}|\overline{c}) = \hat{P}(\text{JAPAN}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

$$\begin{aligned}
\hat{P}(c|d_5) &\propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003 \\
\hat{P}(\overline{c}|d_5) &\propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001
\end{aligned}$$



Nội dung chính

- Ứng dụng phân lớp trong tìm kiếm
- Phương pháp Naïve Bayes
- Đánh giá phương pháp phân lớp



Đánh giá kết quả phân lớp

- Phải được thực hiện trên dữ liệu không trùng lặp với dữ liệu huấn luyện;
- Các tiêu chí cơ bản: Độ chính xác (P), Độ đầy đủ (R), F1.

Các độ đo cơ bản

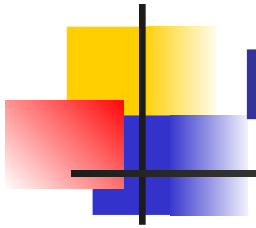
- Thống kê các đại lượng sau đối với một lớp:

	Thuộc lớp	Không thuộc lớp
Dự đoán thuộc lớp	A (TP)	B (FP)
Dự đoán không thuộc lớp	C (FN)	D (TN)

$$\mathrm{P} = \frac{|\mathrm{A}|}{|\mathrm{A} \cup \mathrm{B}|} = \frac{TP}{TP + FP}$$

$$\mathrm{R} = \frac{|\mathrm{A}|}{|\mathrm{A} \cup \mathrm{C}|} = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2\mathrm{PR}}{P + R}$$



Lấy trung bình

- Macro
  - Tính F_1 cho từng lớp;
  - Lấy trung bình các giá trị F_1
- Micro:
  - Thống kê TP, TN, FP, FN cho từng lớp;
  - Lấy tổng các đại lượng thống kê này trên tất cả các lớp;
  - Tính F_1 trên các giá trị tổng hợp này.

Kết quả thực nghiệm: F1 trên Reuters-21578

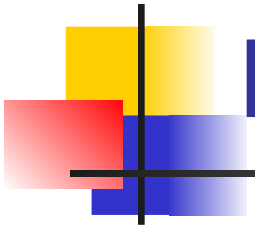
(a)

	NB	Rocchio	kNN	SVM
micro-avg-L (90 classes)	80	85	86	89
macro-avg (90 classes)	47	59	60	60

(b)

	NB	Rocchio	kNN	trees	SVM
earn	96	93	97	98	98
acq	88	65	92	90	94
money-fx	57	47	78	66	75
grain	79	68	82	85	95
crude	80	70	86	85	89
trade	64	65	77	73	76
interest	65	63	74	67	78
ship	85	49	79	74	86
wheat	70	69	77	93	92
corn	65	48	78	92	90
micro-avg (top 10)	82	65	82	88	92
micro-avg-D (118 classes)	75	62	n/a	n/a	87

Bộ phân loại Naïve Bayes tuy đơn giản nhưng hoạt động tương đối tốt so với các bộ phân loại khác.



Bài tập 14.1

Trường hợp mỗi văn bản trong bộ dữ liệu kiểm thử được gán đúng 1 nhãn lớp, đồng thời bộ phân lớp cũng gán đúng một nhãn lớp cho một văn bản, gọi là phân lớp 1 lớp.

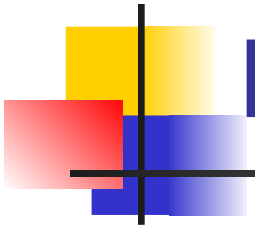
Hãy chứng mình, đối với phân lớp 1 lớp, tổng FP trên tất cả các lớp bằng tổng FN. Nếu lấy trung bình theo micro, thì F_1 tương tự Accuracy.

Bài tập 14.2

Cho bộ văn bản:

(1) He moved from London, Ontario, to London, England.
(2) He moved from London, England, to London, Ontario.
(3) He moved from England to London, Ontario.

Hãy so sánh các biểu diễn túi từ theo mô hình đa thức và mô hình Bernoulli của những văn bản đã cho.



Bài tập 14.3

Ý nghĩa của giả thuyết độc lập với vị trí là: Thông tin từ xuất hiện ở vị trí k cụ thể là không hữu ích. Hãy tìm ngoại lệ.

Thử thiết lập văn bản với với cấu trúc cố định.