Curs 6: Clasificarea surselor de informații

Clasificarea Bayes Naivă. Modelul Bernoulli

1 noiembrie 2018

Se dau (1) o mulțime $\mathbb{C} = \{c_1, c_2, \ldots\}$ de clase de obiecte și (2) un obiect o dintr-o mulțime de obiecte test T.

Se cere să se decidă la ce clasă (sau clase) de obiecte aparține obiectul o.

Noi studiem doar metode de clasificare a documentelor de către sistemele de extragere a informațiilor:

- Obiectele sunt documente indexate de către sistemul de IR
- O clasă poate fi definită în mai multe feluri:
  1. ca mulțime de răspunsuri relevante pentru o cerere de informare.
  2. ca mulțime de documente care se referă la un anumit subiect sau categorie

Clasificări bazate pe cereri de informare

Cererile de informare pot fi:

- tranzitorii: se solicită o singură dată. Sunt cereri caracteristice pentru sistemele ad-hoc de extragere a informațiilor.
- permanente: se solicită periodic, de exemplu cererea `multicore computer chips` solicitată zilnic pentru a urmări dezvoltarea procesoarelor multicore.

Caracteristici ale cererilor permanente de informare:

- Se solicită periodic dintr-o colecție de documente care crește în timp
- Pentru a obține cât mai multe răspunsuri relevante, sistemul de IR **rafinează** cererea. De exemplu

  `multicore computer chips`

  poate fi rafinată la

  `(multicore OR multi-core) AND (chip OR processor OR microprocessor)`

  Rafinarea poate deveni tot mai complexă, în timp.

Probleme de clasificare în sistemele de IR
Aplicații

1. În pașii de preprocesare care produc indecșii inversați: detecția codificării documentelor (ASCII, UNICODE UTF-8, etc.); segmentarea cuvintelor, detecția limbajului unui document.
2. Detecția automată a paginilor spam (care nu se indexează)
3. Detecția automată a conținutului sexual explicit: documente de acest tip sunt incluse doar dacă se activează o opține de căutare, precum `SafeSearch`.
4. Detecția sentimentelor (engl. *sentiment detection*): clasificarea automată a unei recenzii de film sau produs ca fiind pozitivă sau negativă. Această capabilitate este utilă, de ex., pt. a găsi recenzii negative înainte de a cumăra un aparat foto.
5. Sortarea email-urilor și plasarea lor în directoare precum: *anunturi, email-uri de la familie si prieteni*, etc.
6. Motoare de căutare verticală (sau pentru un anumit subiect): limitează căutarea la documente referitoare la un anumit subiect.
7. Unele sisteme de extragere ad-hoc a informațiilor au funcții de calcul al scorurilor de importanță bazate pe un clasificator de document.

Metode de clasificare a documentelor

- **Manuală**, de către un operator uman sau librar care definește/aplică reguli de clasificare
  - elaborarea regulilor necesită îdemânare (de ex., cum să scrie expresii regulate); găsirea unui specialist poate fi dificilă
  - clasificarea manuală este costisitoare pentru colecții mari de documente.
- **Bazată pe machine learning**. Criteriul de decizie al acestei metode este învățat automat din date de antrenament/test.

Exemplu: clasificarea statistică a documentelor text:

- se bazează pe un număr bun de documente de antrenare pentru fiecare clasă.
- marcarea documentelor (adică indicarea claselor la care aparține fiecare document din colecția de date de antrenament) se face de către un operator uman.

- Presupunem că fiecare document d este descris ca element al unei mulțimi $\mathbb{X}$ numită spațiu de documente, și că mulțimea de clase este $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$.
  - De obicei, $\mathbb{X}$ este un spațiu multidimensional.
- Mulțimea de antrenare $\mathbb{D}$ este o submulțime finită a lui $\mathbb{X} \times \mathbb{C}$. De exemplu $\langle d, c \rangle \in \mathbb{D}$ poate fi

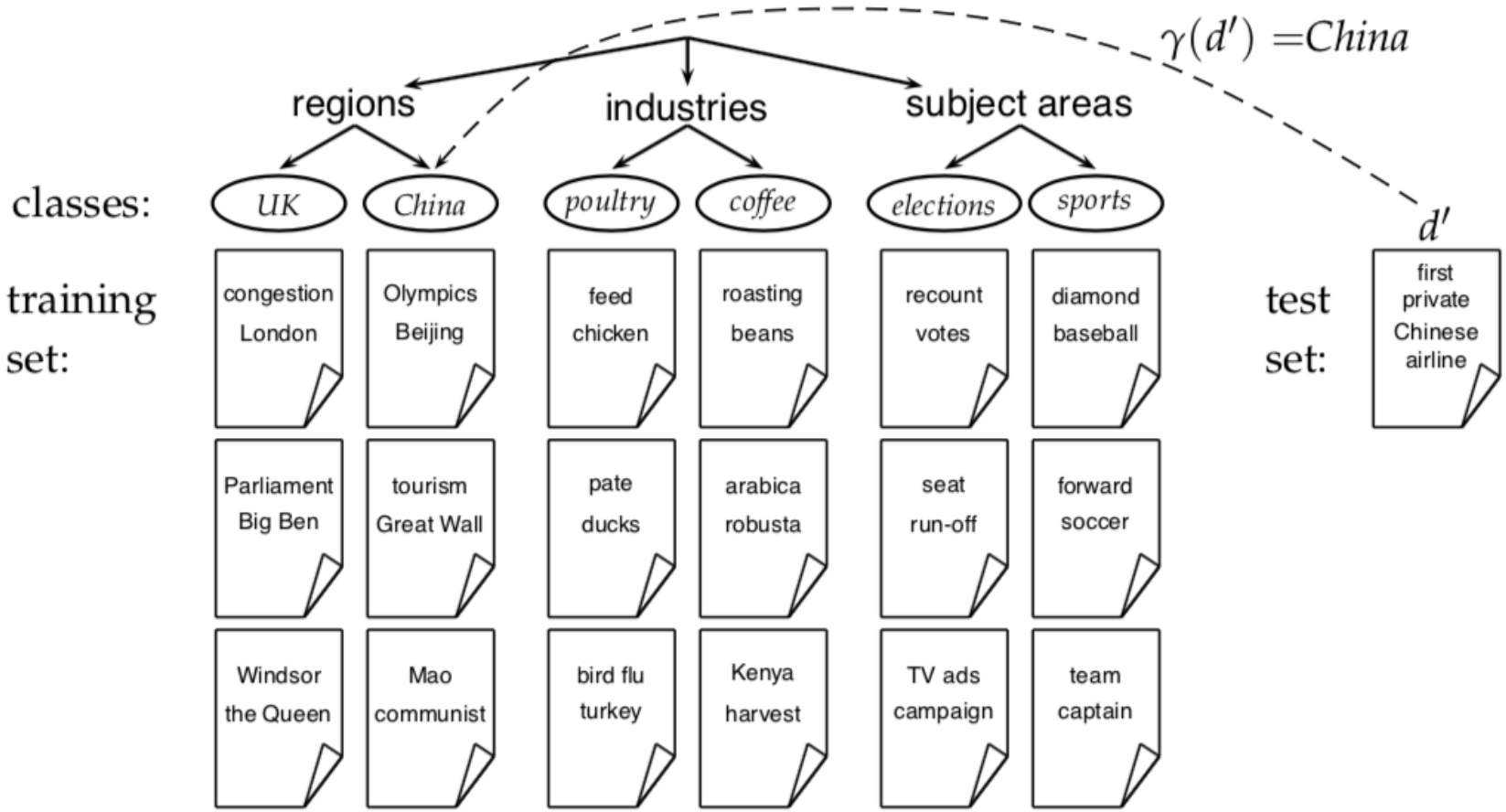
$$\langle d, c \rangle = \langle \text{Beijing joins the World Trade Organization}, \textit{China} \rangle$$

- Un clasificator este o funcție $\gamma : \mathbb{X} \to \mathbb{C}$.
- O metodă de învățare a unui clasificator γ din o mulțime de antrenare $\mathbb{D}$ este o funcție Γ care ia ca argument o mulțime de antrenare $\mathbb{D}$ și calculează un clasificator $\gamma = \Gamma(\mathbb{D})$.

Acest tip de învățare este supervizată pt. că există un supervizor care controlează procesul de învățare, prin intermediul mulțimii $\mathbb{D}$.

- Metoda de învățare Γ este robustă pentru o mulțime de antrenare $\mathbb{D}$ și o mulțime de documente test T dacă procentul de documente din T clasificate greșit de către clasificatorul $\gamma = \Gamma(\mathbb{D})$ este mic.
- **Deziderat** al metodelor de învățare: să fie robuste, aducă să minimizeze eroarea de clasificare a documentelor din T.

$$\mathbb{C} = \{UK, China, poultry, coffee, elections, sports\}$$



PREMISĂ: Clasificarea Bayes naivă se aplică atunci când mț. de exemple de antrenare $\mathbb{D}$ și mț. de teste T sunt similare, adică au aceeași distribuție (vezi cursul următor.) Rezultă că

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

adică, probab. ca $d \in D$ să fie în clasa $c \in \mathbb{C}$ este proporțională cu $P(c) \prod_{1 \leq k \leq n_d} P(t_k \mid c)$ unde

- $P(t_k \mid c)$ este probab. de apariție a termenului t_k în un document din clasa c.
- $P(c)$ este probab. ca un document oarecare să fie în clasa c.
- n_d este nr. de token-uri în d. De ex., dacă conținutul lui d este "Beijing and Taipei join the WTO" atunci secvența de token-uri din d este $\langle$Beijing,Taipei,join,WTO$\rangle$, iar $n_d = 4$.

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

unde $\hat{P}(c)$ și $\hat{P}(t_k \mid c)$ sunt estimările valorilor lui $P(c)$ și $P(t_k \mid c)$ obținute din mulțimea de antrenare $\mathbb{D}$, iar $\arg\max_{c \in \mathbb{C}} expr$ este o valoare a lui $c \in \mathbb{C}$ pentru care *expr* are valoarea maximă.
Pentru a evita erorile de calcul numeric al lui $\prod_{1 \le k \le n_d} \hat{P}(t_k \mid c)$, observăm că c_{map} coincide cu

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \log \left(\hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k \mid c) \right)$$

$$= \arg\max_{c \in \mathbb{C}} \left(\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k \mid c) \right)$$

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

unde $N = |\mathbb{D}|$ și $N_c = |\{d \mid (d, c) \in \mathbb{D}\}|$.

$$\hat{P}(t_k \mid d) := \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

unde T_{ct} este numărul de apariții ale termenului t în documentele din $\{d \mid (d, c) \in \mathbb{D}\}$, incluzând aparițiile multiple.

- De obicei, se preferă eliminarea probab. condiționale cu valoare 0 $\Rightarrow$ se aplică metoda de netezire Laplace:

$$\hat{P}(t_k \mid d) := \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'}) + B}$$

unde V este vocabularul de termeni din mulțimea de antrenare și $B = |V|$.

Clasificarea Bayes naivă

Exemplu de estimare a parametrilor $\hat{P}(c)$ și $\hat{P}(t_k \mid d)$, și de clasificare

$\mathbb{C} = \{China, \overline{China}\}$, $\mathbb{D} = \{(d_1, China), (d_2, China), (d_3, China), (d_4, \overline{China})\}$

	docID	termeni în document	în $c = China$?
mț. de antrenare	1	Chinese Beijing Chinese	da
	2	Chinese Chinese Shanghai	da
	3	Chinese Macao	da
	4	Tokyo Japan Chinese	nu
docum. test	5	Chinese Chinese Chinese Tokyo Japan	?

$N_c = 3$, $N_{\overline{c}} = 1, N = 4$, deci $\hat{P}(c) = \frac{3}{4}$ și $\hat{P}(\overline{c}) = \frac{1}{4}$

$$\begin{aligned} \hat{P}(\text{Chinese} \mid c) &= (5+1)/(8+6) = 6/14 = 3/7 \\ \hat{P}(\text{Tokyo} \mid c) = \hat{P}(\text{Japan} \mid c) &= (0+1)/(8+6) = 1/14 \\ \hat{P}(\text{Chinese} \mid \overline{c}) &= (1+1)/(3+6) = 2/9 \\ \hat{P}(\text{Tokyo} \mid \overline{c}) = \hat{P}(\text{Japan} \mid \overline{c}) &= (1+1)/(3+6) = 2/9 \end{aligned}$$

$\Rightarrow \hat{P}(c \mid d_5) \propto 3/4 \cdot (3/7)^3 \cdot (1/14)^2 \approx 0.0003$

$\hat{P}(\overline{c} \mid d_5) \propto 1/4 \cdot (2/9)^3 \cdot (2/9)^2 \approx 0.0001$, deci $d_5 \in c = China$

AntrenareBayesNaiv($\mathbb{C}, \mathbb{D}$)
1 V :=ExtrageVocabular($\mathbb{D}$)
2 N :=NumaraDocumente($\mathbb{D}$)
3 **for** fiecare $c \in \mathbb{C}$ **do**
4 N_c :=NumaraDocumenteDinClasa($\mathbb{D}, c$)
5 $prior[c] := N_c/N$
6 $text_c$:=ConcateneazaToateTexteleDocumentelorDinClasa($\mathbb{D}, c$)
7 **for** fiecare termen $t \in V$ **do**
8 T_{ct} :=NumaraAparitiileTermenului($text_c, t$)
9 **for** fiecare termen $t \in V$ **do**
10 $condprob[t][c] := \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

AplicaBayesNaiv($\mathbb{C}$, V, *prior*, *condprob*, d)
1 W :=ExtrageTermeniiDinDoc(V, d)
2 **for** fiecare $c \in \mathbb{C}$ **do**
3 $\quad$ *score*[c] := log *prior*[c]
4 $\quad$ **for** fiecare $t \in W$ **do**
5 $\quad\quad$ *score*[c]+ = log *condprob*[t][c]
6 **return** arg $\max_{c \in \mathbb{C}}$ *score*[c]

Diferă de modelul Bayes naiv prin felul cum estimează $\hat{P}(t \mid c)$:

- în modelul Bayes naiv $\hat{P}(t \mid c)$:=fracțiunea conținutului documentelor din c formată din apariții ale lui t
- în modelul Bernoulli $\hat{P}(t \mid c)$:=fracțiunea documentelor din clasa c care conțin termenul t

AntrenareBernoulli($\mathbb{C}, \mathbb{D}$)
1 V :=ExtrageVocabular($\mathbb{D}$)
2 N :=NumaraDocumente($\mathbb{D}$)
3 **for** fiecare $c \in \mathbb{C}$ **do**
4 $\quad N_c$:=NumaraDocumenteDinClasa($\mathbb{D}, c$)
5 $\quad prior[c] := N_c/N$
6 $\quad$ **for** fiecare termen $t \in V$ **do**
7 $\quad\quad N_{ct}$:=NumaraDocumenteledinClasaCareContinTermenul($\mathbb{D}, c, t$)
8 $\quad condprob[t][c] := \frac{N_{ct}+1}{N_c+2}$
9 **return** $V, prior, condprob$

AplicaBernoulli($\mathbb{C}$, V, *prior*, *condprob*, d)
1 V_d :=ExtrageTermeniiDinDoc(V, d)
2 **for** fiecare $c \in \mathbb{C}$ **do**
3 $\quad$ *score*[c] := log *prior*[c]
4 $\quad$ **for** fiecare $t \in V$ **do**
5 $\quad\quad$ **if** $t \in V_d$ **then**
6 $\quad\quad\quad$ *score*[c]+ = log *condprob*[t][c]
7 $\quad\quad$ **else** *score*[c]+ = log(1 − *condprob*[t][c])
8 **return** arg max$_{c \in \mathbb{C}}$*score*[c]

Modelul Bernoulli

Exemplu de estimare a parametrilor $\hat{P}(c)$ și $\hat{P}(t_k \mid d)$, și de clasificare

	docID	termeni în document	în $c = China$?
mț. de antrenare	1	Chinese Beijing Chinese	da
	2	Chinese Chinese Shanghai	da
	3	Chinese Macao	da
	4	Tokyo Japan Chinese	nu
docum. test	5	Chinese Chinese Chinese Tokyo Japan	?

$N_c = 3$, $N_{\overline{c}} = 1, N = 4$, deci $\hat{P}(c) = \frac{3}{4}$ și $\hat{P}(\overline{c}) = \frac{1}{4}$

$$\begin{aligned}
\hat{P}(\text{Chinese} \mid c) &= (3+1)/(3+2) = 4/5 \\
\hat{P}(\text{Tokyo} \mid c) = \hat{P}(\text{Japan} \mid c) &= (0+1)/(3+2) = 1/5 \\
\hat{P}(\text{Beijing} \mid c) = \hat{P}(\text{Macao} \mid c) = P(\text{Shanghai} \mid c) &= (1+1)/(3+2) = 2/5 \\
\hat{P}(\text{Chinese} \mid \overline{c}) = \hat{P}(\text{Japan} \mid \overline{c}) = \hat{P}(\text{Tokyo} \mid \overline{c}) &= (1+1)/(1+2) = 2/3 \\
\hat{P}(\text{Beijing} \mid \overline{c}) = \hat{P}(\text{Macao} \mid \overline{c}) = \hat{P}(\text{Shanhai} \mid \overline{c}) &= (0+1)/(1+2) = 1/3
\end{aligned}$$

$\hat{P}(c \mid d_5) \propto \hat{P}(c) \cdot \hat{P}(\text{Chinese} \mid c) \cdot \hat{P}(\text{Japan} \mid c) \cdot \hat{P}(\text{Tokyo} \mid c) \cdot (1 - \hat{P}(\text{Beijing} \mid c)) \cdot (1 - \hat{P}(\text{Shanghai} \mid c)) \cdot (1 - \hat{P}(\text{Macao} \mid c)) \approx 0.005$

și $\hat{P}(\overline{c} \mid d_5) \propto 1/4 \cdot 2/3 \cdot 2/3 \cdot 2/3 \cdot (1 - 1/3) \cdot (1 - 1/3) \cdot (1 - 1/3) \approx 0.022$, deci

$d_5 \in \overline{China}$

Bibliografie

1. *Text classification and Naive Bayes* (Cap. 13) din Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze: *An Introduction to Information Retrieval*. Ediție online (c) 2009 Cambridge UP. `http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf`