Лекція 4. Наївний метод Байєса [1]

Розглянемо ще один метод навчання з учителем — *мультиномиальний наївний метод Байеса* (Naive Bayes — NB). У цьому методу імовірність того, що документ *d* належить класу *c*, обчислюється в такий спосіб.

$$P(c|d) \approx P(c) \prod_{1 \le k \le n_d} P(t_k|c) \ . \tag{4.1}$$

Тут $P(t_k|c)$ — умовна імовірність, що термін t_k з'явиться в документі з класу *c*, [1] $P(t_k|c)$ — міра правильного розпізнавання класу *c* по терміну t_k, $P(c)$ — апріорна імовірність того, що документ належить класу *c*. Якщо терміни документа не дозволяють чітко відокремити один клас від іншого, то варто вибрати той з них, що має більш високу апріорну імовірність. Послідовність $\langle t_1, t_2, ..., t_{n_d} \rangle$ складається з лексем документа *d*, що є частиною словника, використовуваного для класифікації, а n_d — кількість таких лексем у документі *d*. Наприклад, послідовність $\langle t_1, t_2, ..., t_{n_d} \rangle$ для документа *Beijing and Taipei join the WTO*, що складається з одного речення, може мати вигляд <Beijing, Taipei, join, WTO>, де $n_d = 4$, якщо видалити стоп-слова and і the.

Мета класифікації текстів — знайти *найкращий* клас для документа. У методі *NB* найкращим вважається найбільш ймовірний клас, чи клас c_{map}, що має *максимальну апостеріорну імовірність* (MAP).

$$c_{map} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c) \ . \tag{4.2}$$

Ми пишемо $\hat{P}$, а не *P*, тому що не знаємо справжніх параметрів $P(c)$ і $P(t_k|c)$, а можемо лише оцінити їх за допомогою навчальних множин.

У рівності (4.2) перемножуються кілька умовних імовірностей, по одній для кожного значення $1 \le k \le n_d$. Це може призвести до переповнення машинної пам'яті. Отже, краще замінити добуток імовірностей додаванням их логарифмів. Клас з найбільшим значенням логарифма імовірності залишається найбільш ймовірним, тому що $\log(xy) = \log(x) + \log(y)$ і логарифмічна функція монотонна. Отже, у наївному методі Байеса насправді потрібно знайти точку максимуму наступної функції.

$$c_{map} = \arg\max_{c \in \mathbb{C}} \left[\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k|c) \right] \ . \tag{4.3}$$

Рівність (4.3) допускає просту інтерпретацію. Кожен логарифм умовної імовірності $\log \hat{P}(t_k|c)$ — це вага, що вказує, наскільки важливий термін t_k для класу *c*. Аналогично, апріорна імовірність $\log \hat{P}(c)$ — це вага, що характеризує відносну частоту класу *c*. Ті класи, що зустрічаються більш часто, частіше є правильними, ніж рідкісні. Таким чином, ця сума логарифмів імовірностей і ваг термінів характеризує кількість свідчень того, що документ належить класу, а рівність (4.3) ідентифікує клас, якому відповідає найбільша кількість доказів.

Пока ми обмежимося інтуїтивною інтерпретацією мультиномиальної наївної байєсівської моделі і відкладемо його формальний вивід.

Как оцінити імовірності $\hat{P}(c)$ і $\hat{P}(t_k|c)$? Спочатку спробуємо одержати оцінку максимальної правдоподібності, що являє собою відносну частоту і відповідає найбільш ймовірній величині кожного параметра при заданих навчальних даних. Для апріорних імовірностей оцінка має наступний вид.

$$\hat{P}(c) = \frac{N_c}{N'} \ . \tag{4.4}$$

Тут N_c — кількість документів у класі *c*, а *N* — загальна кількість документів.

Оцінимо умовну імовірність $\hat{P}(t|c)$ как відносну частоту терміна *t* у документі, що належить класу *c*.

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \ . \tag{4.5}$$

Тут T_{ct} — кількість появ терміна *t* у навчальних документах із класу *c* з урахуванням багаторазових появ терміна в документі. Ця оцінка заснована на *припущенні про позиційну незалежність*: умовні імовірності появи терміну однакові незалежно від його позиції в документі, тобто

$$P(X_{k_1} = t|c) = P(X_{k_2} = t|c) \ .$$

для всіх позицій k_1 і k_2, термінів *t* і класів *c*. Таким чином, у моделі виникає єдиний розподіл термов для всіх позицій k_i; T_{ct} — це кількість появ терміна у всіх позиціях *k* у документах з навчальної множини. Таким чином, ми не обчислюємо різні оцінки для різних позицій і, наприклад, якщо слово двічі зустрічається в документі на позиціях k_1 і k_2, то $\hat{P}(t_{k_1}|c) = \hat{P}(t_{k_2}|c)$.

---

[1] Чому тут використаний знак пропорційності, а не рівності?

З оцінкою максимальної правдоподібності зв'язана одна проблема: якщо пари термін-клас не зустрічаються в навчальних даних, то оцінка MLE дорівнює нулю. Наприклад, якщо термін WTO у навчальних даних зустрічається тільки в документах класу *China*, то оцінки MLE для інших класів, наприклад, класу *UK*, дорівнюють нулю.

$$\hat{P}(WTO|UK) = 0 \ .$$

Тепер умовна імовірність класу *UK* щодо документа *Britain is a member of the WTO*, що складається з одного речення, дорівнює нулю, оскільки в рівності (4.1) ми перемножуємо умовні імовірності для всіх термінів. Очевидно, что модель повинна привласнювати класу *UK* високу імовірність, оскільки в пропозиції зустрічається термін *Britain*. Втім, не можна просто відкинути нульову імовірність для терміна WTO, незалежно від того наскільки багато є свідчень на користь класу *UK*, забезпечених іншими ознаками. Ця оцінка дорівнює нулю через *рідкість* терміна. Навчальні дані ніколи не бувають великими настільки, щоб частота рідких термінів оцінювалася адекватно, як, наприклад, частота терміна WTO у документах класу *UK*.

Для того щоб позбутися від нуля, ми використовуємо *згладжування Лапласа* (Laplace smoothing), просто додаючи одиницю до кожної частоти.

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{\sum_{t' \in V} T_{ct'} + B} \ . \qquad (4.6)$$

Тут $B = |V|$ — кількість термінів у словнику. Згладжування Лапласа можна інтерпретувати як апріорний рівномірний розподіл (кожен термін зустрічається в кожнім класі по одному разі), що потім уточнюється на основі навчальних даних, що надходять. Відзначимо, что це — апріорна імовірність появи *терміну*, а не *класу*, що оцінюється формулою (4.4) на рівні документа.

Уведемо тепер всі елементи, необхідні для навчання і застосування в наївному байєсівському класифікаторі. Повний алгоритм приведений на мал. 4.1.

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, $prior$, $condprob$

ApplyMultinomialNB($\mathbb{C}$, V, $prior$, $condprob$, d)
1 $W \leftarrow$ 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** $\arg\max_{c \in \mathbb{C}} score[c]$

Рис. 4.1. Наївний байєсівський алгоритм (мультиномиальна модель): навчання і тестування

Приклад 4.1. Ґрунтуючись на зразках, приведених у табл. 4.1 і зазначених нижче параметрах, потрібно класифікувати тестовий документ.

$\hat{P}(c)$ =3/4, $\hat{P}(\bar{c})$ =1/4, $\hat{P}$(Chinese|c) = (5 + 1)/(8 + 6) = 6/14 = 3/7,

$\hat{P}$(Tokio|c) = $\hat{P}$(Japan|c) = (0 + 1)/(8 + 6) = 1/14,

$\hat{P}$(Chinese|$\bar{c}$) = (1 + 1)/(3 + 6) = 2/9,

$\hat{P}$(Tokio|$\bar{c}$) = $\hat{P}$(Japan|$\bar{c}$) = (1 + 1)/(3 + 6) = 2/9.

Знаменники рівні (8 + 6) і (3 + 6), оскільки довжина тексту $text_c$ і $text_{\bar{c}}$ рівні 8 і 3 відповідно, а константа B у рівності (4.6) дорівнює 6, тому що словник складається із шести термінів. Таким чином,

$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0{,}0003,$

$$\hat{P}(c|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0{,}0001.$$

Отже, класифікатор віднесе тестовий документ до класу *c* = *China*. Причина такої класифікації полягає в тім, что три позитивних індикатори входження терміна Chinese у документ d_5 переважують негативні індикатори термінів Japan і Tokio.

Таблиця 4.1. Дані для оцінки параметрів

	docID	*Слова в документі*	*c = China?*
навчальна множина	1	Chinese Beijing Chinese	так
	2	Chinese Chinese Shanghai	так
	3	Chinese Makao	так
	4	Tokio Japan Chinese	ні
тестова множина	5	Chinese Chinese Chinese Tokio Japan	?

Чому дорівнює часова складність алгоритму NB? Складність обчислення параметрів дорівнює $\Theta(|\mathbb{C}||V|)$, оскільки множина параметрів складається з $|\mathbb{C}||V|$ умовних імовірностей і $|\mathbb{C}|$ апріорних. Для обчислення цих параметрів необхідна певна попередня робота (витяг терміна зі словника, підрахунок і т.д.), яку можна виконати за один прохід по навчальним даним. Отже, часова складність цього компонента дорівнює $\Theta(|\mathbb{D}|L_{ave})$, де $|\mathbb{D}|$ — кількість документів, а L_{ave} — середня довжина документа, $|\mathbb{C}|$ — кількість класів, $|V|$ — кількість термінів у словнику.

Тут ми використовували позначення $\Theta(|\mathbb{D}|L_{ave})$ как варіант позначення $\Theta(T)$, де T — довжина навчальної колекції. Це нестандартне рішення, оскільки величини $\Theta(\cdot)$ не визначаються для середніх величин. Ми виражаємо часову складність через величини $\mathbb{D}$ і L_{ave}, оскільки вони є основними статистичними показниками, що характеризують навчальні колекції.

Часова складність алгоритму ApplyMultinomialNB, представленого на мал. 4.1, дорівнює $\Theta(|\mathbb{C}|L_a)$. Нехай L_a — довжина тексту в лексемах, а M_a — розмір його лексікону. Алгоритм ApplyMultinomialNB можна модифікувати так, щоб его часова складність стала рівною $\Theta(L_a + |\mathbb{C}|M_a)$. На закінчення відзначимо, что $\Theta(L_a + |\mathbb{C}|M_a) = \Theta(|\mathbb{C}|M_a)$, оскільки при фіксованій константі b виконується нерівність $L_a < b|C|M_a$. [2]

Оцінки часової складності приведені в табл. 4.2. У целом, $|\mathbb{C}|M_a < |\mathbb{D}|L_{ave}$, тому складність навчання і тестування лінійної залежить від часу, необхідного для сканування даних. Оскільки дані проглядаються як мінімум один раз, метод NB можна вважати оптимальним за часовою складністю. Ця ефективність є однією з основних причин, по яких метод NB одержав таке широке поширення.

Таблиця 4.2. Дані для оцінки параметрів

Алгоритм	*Тимчасова складність*
Навчання	$\Theta(\|\mathbb{D}\|L_{ave} + \|\mathbb{C}\|\|V\|)$
Тестування	$\Theta(L_a + \|\mathbb{C}\|M_a) = \Theta(\|\mathbb{C}\|M_a)$

Література

1. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск.— М.: Изд-во «Вильямс», 2011.

[2] Припускаємо, що довжина тестових документів обмежена. Для надзвичайно довгих тестових документів величина L_a перевищує $b|C|M_a$.