Лекція 3. Байєсівські методи класифікації

В цій лекції розглядаються задачі класифікації, які виникають в теорії інформаційного пошуку. Ці задачі відрізняються тим, що в них об'єкт може описуватися або як *мішок*, або як *вектор* ознак.

Розглянемо *мультиноміальний наївний метод Байєса*, який застосовується для розв'язання імовірнісної постановки задачі класифікації об'єктів, які описуються як *мішок* ознак. [1] У цьому методі імовірність того, що об'єкт d належить до класу c, обчислюється в такий спосіб:

$$P(c|d) \approx P(c) \prod_{1 \le k \le n_d} P(t_k|c) \ . \qquad (3.1)$$

Тут $P(t_k|c)$ — умовна імовірність, що ознака t_k буде наявною у об'єкта з класу c, $P(t_k|c)$ — міра правильного розпізнавання класу c за ознакою t_k, $P(c)$ — апріорна імовірність того, що об'єкт належить класу c. Якщо ознаки об'єкта не дозволяють чітко відокремити один клас від іншого, то варто вибрати ті з них, що має більш високу апріорну імовірність. Послідовність $\langle t_1, t_2, ..., t_{n_d} \rangle$ складається з ознак об'єкта d, а n_d — кількість таких ознак у об'єкті d. Наприклад, послідовність $\langle t_1, t_2, ..., t_{n_d} \rangle$ для об'єкта, що є текстом *Beijing and Taipei join the WTO*, який складається з одного речення, може мати вигляд <Beijing, Taipei, join, WTO>, де $n_d = 4$, якщо видалити службові слова.

Мета класифікації — знайти *найкращий* клас для об'єкта. У мультиноміальному наївному методі Байєса найкращим вважається найбільш ймовірний клас, чи клас c_{map}, що має *максимальну апостеріорну імовірність* (MAP).

$$c_{map} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c) \ . \qquad (3.2)$$

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 (3.2) перемножуються кілька умовних ймовірностей, по одній для кожного значення $1 \le k \le n_d$. Це може призвести до переповнення машинної пам'яті. Отже, краще замінити добуток ймовірностей додаванням їх логарифмів. Клас з

[1] Наприклад, сукупність слів у тексті, що можуть повторюватися.

найбільшим значенням логарифма імовірності залишається найбільш ймовірним, тому що $\log(xy) = \log(x) + \log(y)$ і логарифмічна функція монотонна. Отже, у наївному методі Байєса насправді потрібно знайти точку максимуму функції

$$c_{map} = \arg\max_{c \in \mathbb{C}} \left[\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k | c) \right]. \quad (3.3)$$

Рівність (3.3) допускає просту інтерпретацію. Кожен логарифм умовної імовірності $\log \hat{P}(t_k | c)$ — це вага, що вказує, наскільки важлива ознака t_k для класу c. Аналогічно, 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 ймовірностей і ваг ознак характеризує кількість свідчень того, що об'єкт належить класу, а рівність (3.3) ідентифікує клас, якому відповідає найбільша кількість доказів.

Як 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 ймовірностей оцінка має наступний вид.

$$\hat{P}(c) = \frac{N_c}{N'}. \quad (3.4)$$

Тут N_c — кількість об'єктів у класі c, а N — загальна кількість об'єктів.

Оцінимо умовну імовірність $\hat{P}(t | c)$ як відносну частоту ознаки t у об'єкті, що належить класу c.

$$\hat{P}(t | c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}. \quad (3.5)$$

Тут T_{ct} — кількість появ ознаки t у навчальних об'єктах із класу c з урахуванням багаторазових появ терміна в об'єкті. Ця оцінка заснована на *припущенні про позиційну незалежність*: умовні ймовірності появи ознаки однакові незалежно від її позиції в описі об'єкта (у *мішку* ознак), тобто

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

для всіх позицій k_1 і k_2, ознак t і класів c. Таким чином, у моделі виникає єдиний розподіл ознак для всіх позицій k_i; T_{ct} — це

кількість появ ознаки у всіх позиціях k в об'єктах з навчальної множини. Таким чином, ми не обчислюємо різні оцінки для різних позицій і, наприклад, якщо слово двічі зустрічається в тексті на позиціях k_1 і k_2, то

$$\hat{P}\left(t_{k_1}\middle|c\right)=\hat{P}\left(t_{k_2}\middle|c\right).$$

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

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

Тепер умовна імовірність класу *UK* щодо об'єкта *Britain is a member of the WTO*, що складається з одного речення, дорівнює нулю, оскільки в рівності (3.1) ми перемножуємо умовні імовірності для всіх термінів. Очевидно, що модель повинна приписувати класу *UK* високу імовірність, оскільки в реченні 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* ознаки. Навчальні дані ніколи не бувають великими настільки, щоб частота рідких ознак оцінювалася адекватно, як, наприклад, частота терміна WTO у об'єктах класу *UK*.

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

$$\hat{P}\left(t\middle|c\right)=\frac{T_{ct}+1}{\sum_{t'\in V}\left(T_{ct'}+1\right)}=\frac{T_{ct}+1}{\sum_{t'\in V}T_{ct'}+B}\ . \qquad (3.6)$$

Тут $B=|V|$ — кількість ознак. Згладжування Лапласа можна інтерпретувати як апріорний рівномірний розподіл (кожна ознака зустрічається в кожнім класі по одному разі), що потім уточнюється на основі навчальних даних, що надходять. Відзначимо, що це — апріорна імовірність появи *ознаки*, а не *класу*, що оцінюється формулою (3.4) на рівні об'єкта.

Приклад 4.1. Ґрунтуючись на зразках, приведених у таблиці, і зазначених нижче параметрах, потрібно класифікувати тестовий об'єкт.

	docID	*Слова в тексті*	*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	?

$\hat{P}(c)=3/4,\ \hat{P}(\overline{c})=1/4,$

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

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

Знаменники рівні (8 + 6) і (3 + 6), оскільки довжина тексту $text_c$ і $text_{\overline{c}}$ рівні 8 і 3 відповідно, а константа B у рівності (3.6) 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 об'єкт до класу c = *China*. Причина такої класифікації полягає в тім, що три позитивних індикатори входження терміна Chinese у об'єкт d_5 переважують негативні індикатори термінів Japan і Tokio.

Існує два види наївної байєсівської моделі. Одна з них — мультиноміальна — була описана вище. Вона генерує одну ознаку на кожній позиції векторного опису об'єкта.

Альтернативою мультиноміальній моделі є *модель Бернуллі*. Вона еквівалентна бінарній моделі незалежності, що генерує індикатор для кожного терміну словника: 1, якщо термін є присутнім у документі, і 0, якщо відсутнім.

Різні моделі використовують різні стратегії оцінки і різні правила класифікації. У моделі Бернуллі імовірність $\hat{P}(t|c)$

оцінюється як *частка об'єктів із класу c*, що мають ознаку *t*. На противагу йому в мультиноміальній моделі імовірність $\hat{P}(t|c)$ оцінюється як *частка ознаки t* в об'єктах з класу *c*. При класифікації тестового документа на основі моделі Бернуллі використовується бінарна інформація про появу ознаки, що ігнорує кількість входжень цієї ознаки, у той час як мультиноміальна модель відслідковує багаторазові появи терміну в документі. У результаті при класифікації довгих документів модель Бернуллі, як правило, допускає багато помилок. Наприклад, вона може віднести до класу *China* цілу книгу через єдине згадування терміну *China*.

Ці моделі розрізняються також тим, як використовуються слова, що не з'являються в документі. У мультиноміальній моделі ця інформація ніяк не впливає на рішення, а в моделі Бернуллі імовірність відсутності терміну при обчисленні імовірності $P(c|d)$ факторизується. Це пояснюється тим, що тільки моделі Бернуллі враховують інформацію про відсутність терміну явно.

Приклад 3.2. Застосувавши модель Бернуллі до даних, наведених у прикладі 3.1, ми одержали оцінки $\hat{P}(c)$=3/4 і $\hat{P}(\overline{c})$=1/4. Умовні імовірності набувають наступні значення.

$$\hat{P}(\text{Chinese}|c) = (3 + 1)/(3 + 2) = 4/5,$$

$$\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokio}|c) = (0 + 1)/(3 + 2) = 1/5,$$

$$\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) = (1 + 1)/(3 + 2) = 2/5,$$

$$\hat{P}(\text{Chinese}|\overline{c}) = (1 + 1)/(1 + 2) = 2/3,$$

$$\hat{P}(\text{Japan}|\overline{c}) = \hat{P}(\text{Tokyo}|\overline{c}) = (1 + 1)/(1 + 2) = 2/3,$$

$$\hat{P}(\text{Beijing}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) = \hat{P}(\text{Shanghai}|\overline{c}) = (0 + 1)/(1 + 2) = 1/3.$$

Знаменники рівні (3 + 2) і (1 + 2), оскільки в класі *c* існують три документи, а в класі $\overline{c}$ — один документ, а константа *B* у рівності Лагранжа дорівнює двом: для кожного терміну існують два варіанти — вони або входять у документ, або ні.

Ранги тестового документа стосовно цих двох класів такі.

$$\hat{P}(c|d_5) \propto \hat{P}(c)\cdot\hat{P}(\text{Chinese}|c)\cdot\hat{P}(\text{Japan}|c)\cdot\hat{P}(\text{Tokyo}|c)\cdot$$
$$\cdot\left(1-\hat{P}(\text{Beijing}|c)\right)\cdot\left(1-\hat{P}(\text{Shanghai}|c)\right)\cdot\left(1-\hat{P}(\text{Makao}|c)\right)=$$
$$=3/4\cdot 4/5\cdot 1/5\cdot 1/5\cdot(1-2/5)\cdot(1-2/5)\cdot(1-2/5)\approx 0,005.$$

Аналогічно,

$$\hat{P}(\overline{c}|d_5) \propto 1/4\cdot 2/3\cdot 2/3\cdot 2/3\cdot(1-1/3)\cdot(1-1/3)\cdot(1-1/3)\approx 0,022.$$

Отже, класифікатор віднесе тестовий документ до класу $\overline{c}$ = *not-China*. Якщо враховувати тільки бінарний індикатор входження терміну в документа, а не його частоту, то індикатори термінів Japan і Tokyo є індикаторами класу $\overline{c}$ (2/3 > 1/5), а умовні імовірності терміну Chinese для класів c і $\overline{c}$ недостатньо сильно відрізняються друг від друга (4/5 від 2/3) і не впливають на остаточну класифікацію.

Для того щоб краще зрозуміти ці дві моделі, а також припущення, на яких вони засновані, повернемося назад і перевіримо, як ми вивели правила класифікації раніше. Нагадаємо, що рішення про належність до класу набувається шляхом визначення класу з максимальною апостеріорною імовірністю.

$$c_{map} = \arg\max_{c\in\mathbb{C}} \hat{P}(c|d)$$
$$= \arg\max_{c\in\mathbb{C}} \frac{P(d|c)P(c)}{P(d)} = \arg\max_{c\in\mathbb{C}} P(d|c)P(c). \qquad (3.7)$$

Застосовуючи до виразу (3.7) правило Байєса, можна відкинути знаменник, оскільки він є постійним для всіх класів і не впливає на величину argmax.

Вираз (3.7) можна інтерпретувати як опис процесу генерування, характерного для байєсівської класифікації текстів. Для того щоб згенерувати документ, ми спочатку вибираємо клас c з імовірністю $P(c)$. Ці дві моделі відрізняються способом генерування документа по заданому класі відповідно до умовного розподілу $P(d|c)$.

Мультиноміальна модель $P(d|c) = P\left(\langle t_1,...,t_k,...,t_{n_d}\rangle \middle| c\right)$

Модель Бернуллі $P(d|c) = P\left(\langle e_1,...,e_i...,e_M\rangle \middle| c\right)$

Тут $\langle t_1,...,t_k,...,t_{n_d} \rangle$ — послідовність слів, що з'являються в документі d (за винятком службових слів), а $\langle e_1,...,e_i...,e_M \rangle$ — бінарний вектор, що складається з M індикаторів присутності кожного терміну в документі d.

Тепер зрозуміліше, чому критичний крок у розв'язуванні задачі класифікації текстів — вибір представлення документа. Наприклад, такими представленнями документа є послідовності $\langle t_1,...,t_k,...,t_{n_d} \rangle$ і $\langle e_1,...,e_i...,e_M \rangle$. У першому варіанті простір $\mathbb{X}$ є множиною всіх послідовностей термінів (чи, точніше, послідовностей лексем термінів). В другому варіанті простір ознак являє собою множину $[0,1]^M$.

Література

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