Лекція 14
Баєсева класифікація та регресія

§73 Спам-фільтр на основі класифікатора Баєса

Задача фільтрації спама

- Задача класифікації листів на 2 класи — спам (spam) і не спам (ham):

$$Y = \{\text{spam}, \text{ham}\}$$

- Перші спам-фільтри використовували наївний баєсів класифікатор
- Прийдемо до нього, відштовхнувшись від задачі фільтрації спама

Приклади спамних листів:

- Hi! :) Purchase Exclusive Tabs Online http://...
- We Offer Loan At A Very Low Rate Із 3%. Із Interested, Kindly Contact Із, Reply by email ...@hotmail.com
- Купіть за супер-знижкою 0.99%! Станьте…

Фільтруємо спам: навчання

- Обчислимо для кожного слова w з колекції текстів кількість листів з ним n_{ws} у спамі (spam) і кількість листів з ним n_{wh} в «не спамі» (ham)

- Оцінити ймовірність появи кожного слова w у спамному і в неспамному тексті:

$$P(\mathrm{w} \mid \mathrm{spam}) = n_{\mathrm{ws}}/n_{\mathrm{s}}$$

$$P(\mathrm{w} \mid \mathrm{ham}) = n_{\mathrm{wh}}/n_{\mathrm{h}}$$

Фільтруємо спам: застосування

Одержавши текст листа, для якого потрібно визначити, відноситься він до спаму чи ні, ми можемо:

1. Оцінити **ймовірність появи всього тексту** в класі «спам» і в класі «не спам» просто добутком імовірностей слів

$$P(\text{text}\,|\,\text{spam}) \approx P(\text{w}_1\,|\,\text{spam})P(\text{w}_2\,|\,\text{spam})...P(\text{w}_N\,|\,\text{spam})$$

$$P(\text{text}\,|\,\text{ham}) \approx P(\text{w}_1\,|\,\text{ham})P(\text{w}_2\,|\,\text{ham})...P(\text{w}_N\,|\,\text{ham})$$

«Наївна» гіпотеза: входження слів у текст – незалежні події

2. Вибрати той клас, у якому імовірність появи цього тексту більше:

$$a(\text{text}) = \text{argmax}\, P(\text{text}\,|y)$$

Це майже правильний алгоритм

Уточнюємо алгоритм

- Насправді

$$a(\text{text}) = \underset{y}{\operatorname{argmax}} P(\text{text}\,|y)$$

- Нам потрібна ймовірність, що **цей текст належить якомусь класу** , $P(y|\,\text{text})$ а на $P(\text{text}\,|y)$

- Правильний алгоритм повинен вибирати той клас, для якого більше ймовірність : $P(y|\,\text{text})$

$$a(\text{text}) = \underset{y}{\operatorname{argmax}}\, P(y|\,\text{text})$$

Але як же її оцінити?

Теорема Баєса

$$P(y|\,\text{text}) = P(\text{text}\,|y)P(y)/P(\text{text})$$

Імовірність появи тексту $P(\text{text})$ однакова для обох класів *y*:

Тому

$$\underset{y}{\text{argmax}}\, P(y|\,\text{text}) = \underset{y}{\text{argmax}}\, P(\text{text}\,|y)P(y)$$

Спам-фільтр на наївному баєсів класифікаторі

- Навчання:

$$P(\mathrm{w}\,|\,\mathrm{spam}) = n_{\mathrm{ws}}/n_{\mathrm{s}}$$

$$P(\mathrm{w}\,|\,\mathrm{ham}) = n_{\mathrm{wh}}/n_{\mathrm{h}}$$

- Застосування:

$$P(\mathrm{new\ \ text}\,|\,\mathrm{spam}) =$$
$$= P(\mathrm{w}_1\,|\,\mathrm{spam})P(\mathrm{w}_2\,|\,\mathrm{spam})...P(\mathrm{w}_N\,|\,\mathrm{spam})$$

$$P(\mathrm{new\ \ text}\,|\,\mathrm{ham}) =$$
$$= P(\mathrm{w}_1\,|\,\mathrm{ham})P(\mathrm{w}_2\,|\,\mathrm{ham})...P(\mathrm{w}_N\,|\,\mathrm{ham})$$

- Віднести текст до такого класу в (ham або spam), для якого буде більше величина

$$P(y)P(\mathrm{new\ \ text}\,|y)$$

Фільтрація спама: що ще не врахували

- Ніяк не використана інформація, що є в заголовку листа й адресі відправника.
- Якщо слово w не зустрічається в жодному з навчальних текстів для якогось класу, його ймовірності $P(w|y)$ відразу оцінюється нулем. А значить імовірність того, що текст належить класу y відразу оцінюється нулем, що може бути досить невдалим рішенням.
- Припустимо є слово w_1, що не входило в навчальні тексти першого класу (зі спамом), і слово w_2, що не входило в навчальні тексти другого класу (без спама). Тоді, якщо обидва ці слова є в деякому тексті, обидві ймовірності $P(y_1|spam)$ і $P(y_2|ham)$ будуть дорівнюють нулю, а значить не можна буде віднести текст до будь-якого з класів.

§74 Наївний баєсів класифікатор

Баєсів класифікатор

Нехай деякий **об'єкт має вектор ознак x**. Необхідно визначити, **до якого класу y слід віднести цей об'єкт**.

Баєсів класифікатор **$a(x)$ відносить об'єкт до такого класу, імовірність якого максимальна за умови, що реалізувався цей об'єкт**:

$$a(x) = \operatorname{argmax}_y P(y|x) = \operatorname{argmax}_y P(x|y)P(y).$$

$$a(x) = \operatorname*{argmax}_y P(y|x)$$

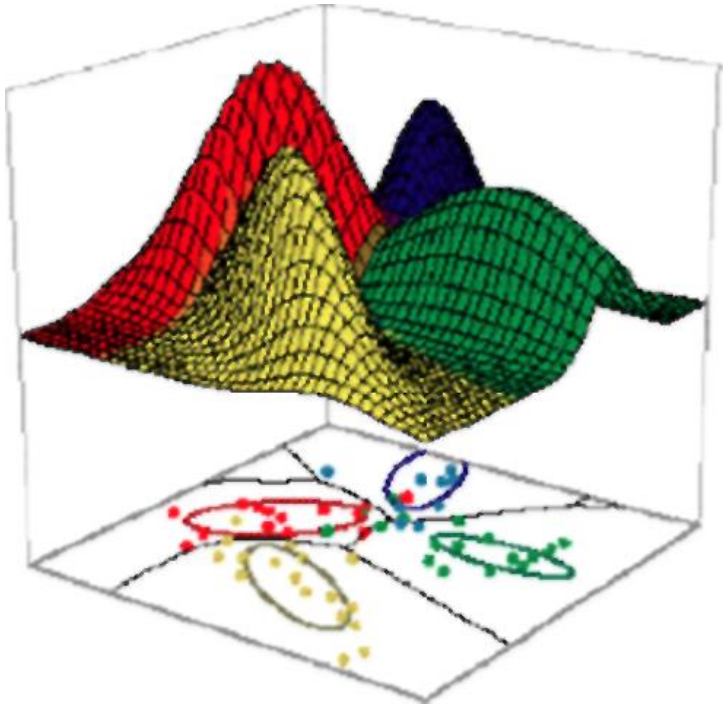
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$a(x) = \operatorname*{argmax}_y P(x|y)P(y)$$

Баєсів класифікатор

$$a(x) = \underset{y}{\operatorname{argmax}} P(x|y)P(y)$$

Якщо $P(y)$ однакові для всіх класів - ми просто вибираємо клас, густина якого більше в точці x



Необхідність використання теореми Баєса

Безпосереднє **обчислення $P(y|x)$** полягає у тому, що необхідно розглянути множину об'єктів, які **мають ознаковий опис x**, і знайти частку класу y серед цієї множини. Але можливих **ознакових описів величезна кількість, а значить навряд чи в навчальній вибірці буде достатня кількість об'єктів для всякого можливого x.** Таким чином, **не маємо змоги обчислювати $P(y|x)$ безпосередньо й при ходиться застосовувати теорему Баєса**.

Теорема Баєса дозволяє переходити до $P(x|y)$, тобто фактично до густини розподілу x *за умови* класу y (у випадку дійсних ознак). Останню величину вже можна оцінювати за навчальною вибіркою.

Застосування класифікатора відбувається в такий спосіб:

$$a(x) = \text{argmax}_y P(x|y)P(y).$$

Що оцінюється за навчальною вибіркою

- $P(x|y)$— імовірність побачити набір ознак ***x*** у класі ***y***, якщо ***x*** дискретний
- Якщо координати вектора x — дійсні, $P(x|y)$ — густина розподілу ***x***
- Саме цю величину й можна оцінювати за навчальною вибіркою
- А потім підставляти в класифікатор:

$$a(x) = \operatorname*{argmax}_y P(x|y)P(y)$$

Проблема нестачі даних

- Приклад: у навчальній вибірці **100 000** об'єктів з **10 000** ознак
- **100 000** точок у просторі розмірності **10 000** -дуже мало
- Наприклад, якщо x — бінарний, то в нього може бути $\mathbf{2^{10000}}$ значень, що набагато більше **100 000**
- Тому відновити $P(x|y)$ як функцію від ознак x досить важко

§75 Відновлення розподілів (частина 1)

Безпосередньо відновити розподіл $P(x|y)$ **не має можливості через проблему нестачі даних**.

Наївний баєсів класифікатор

Рішення – **звести задачу відновлення $P(x|y)$ від оцінки функції багатьох змінних до оцінки функцій однієї змінної**

- Використання «наївного» баєсів класифікатора, тобто баєсів класифікатора:

$$a(x) = \underset{y}{\operatorname{argmax}} P(y|x) = \underset{y}{\operatorname{argmax}} P(x|y)P(y)$$

- і «наївної» гіпотези, що густина розподілу розписується як добуток густин за кожною ознакою:

$$P(x|y) = P(x_{(1)}|y)P(x_{(2)}|y)...P(x_{(N)}|y)$$

де $x_{(k)}$ — k-а ознака об'єкта x.

Відновлення розподілів $P(x_{(k)}|y)$

- При навчанні необхідно **визначити за навчальною вибіркою розподіл $P(x_{(k)}|y)$ і апріорні ймовірності класів $P(y)$.**

- Оцінити апріорні ймовірності класів на основі вибірки можна так:

$$P(y) \approx \frac{\ell_y}{\ell},$$

де l_y — кількість об'єктів класу y у навчальній вибірці, а l — розмір навчальної вибірки. Якщо відношення часток класів у навчальній вибірці не відбиває їхнє реальне співвідношення, апріорні ймовірності класів повинні бути взяті із зовнішніх даних.

- Розподіл $P\left(x_{(k)}\middle|y\right)$ можна оцінити **як частку об'єктів з даним значенням ознаки $x_{(k)}$ серед об'єктів класу y**:

$$P(x_{(k)}|y) = \frac{1}{\ell_y}\#(x_{(k)}, y).$$

- Таким чином, для бінарних ознак:

$$P(x_{(k)} = 0|y) = \frac{1}{\ell_y}\#(x_{(k)} = 0, y), \qquad P(x_{(k)} = 1|y) = \frac{1}{\ell_y}\#(x_{(k)} = 1, y).$$

Приклад: класифікація текстів

- Класифікатор текстів можна побудувати в такий спосіб. За навчальною вибіркою будується словник всіх вхідних у тексти навчальної вибірки слів. Кожний текст буде характеризуватися вектором з бінарних ознак: $x_{(k)} = 1$, якщо слово w_k є присутнім у тексті, а якщо не є присутнім — $x_{(k)} = 0$.
- Після цього можна відновити розподіл як це описано вище:

$$P(x_{(k)} = 0|y) = \frac{1}{\ell_y}\#(x_{(k)} = 0, y), \qquad P(x_{(k)} = 1|y) = \frac{1}{\ell_y}\#(x_{(k)} = 1, y).$$

- Після цього можна застосувати наївний баєсів класифікатор і в такий спосіб вирішити задачу класифікації.

$$a(x) = \operatorname*{argmax}_y P(y|x) = \operatorname*{argmax}_y P(x|y)P(y)$$

$$P(x|y) = P(x_{(1)}|y)P(x_{(2)}|y)...P(x_{(N)}|y)$$

Згладжування ймовірностей

- Якщо в навчальній вибірці серед множини об'єктів певного класу y ніколи **не зустрічалося якесь значення t** деякої ознаки $x_{(k)}$, то ймовірність $\boldsymbol{P(x_{(k)} = t|y) = 0}$.
- Оскільки у виразі, який потрібно максимізувати, знаходиться добуток таких імовірностей, увесь цей вираз буде дорівнює нулю.

$$P(x_{(1)}|y)P(x_{(2)}|y)...P(x_{(N)}|y) = 0$$

- Таким чином, **будь-який об'єкт**, тільки на підставі того, що значення ознаки $x_{(k)} = t$, не було віднесеним до класу y, що, загалом кажучи, неправильно.
- Уникнути такої ситуації можна за допомогою згладжування ймовірності, наприклад у такий спосіб:

$$P(x_{(k)} = 1|y) = \frac{\#(x_{(k)} = 1, y) + a}{\ell_y + a + b}, \qquad P(x_{(k)} = 0|y) = \frac{\#(x_{(k)} = 0, y) + b}{\ell_y + a + b}.$$

Константи a і b вибираються таким чином, що якість алгоритму була найбільшою.

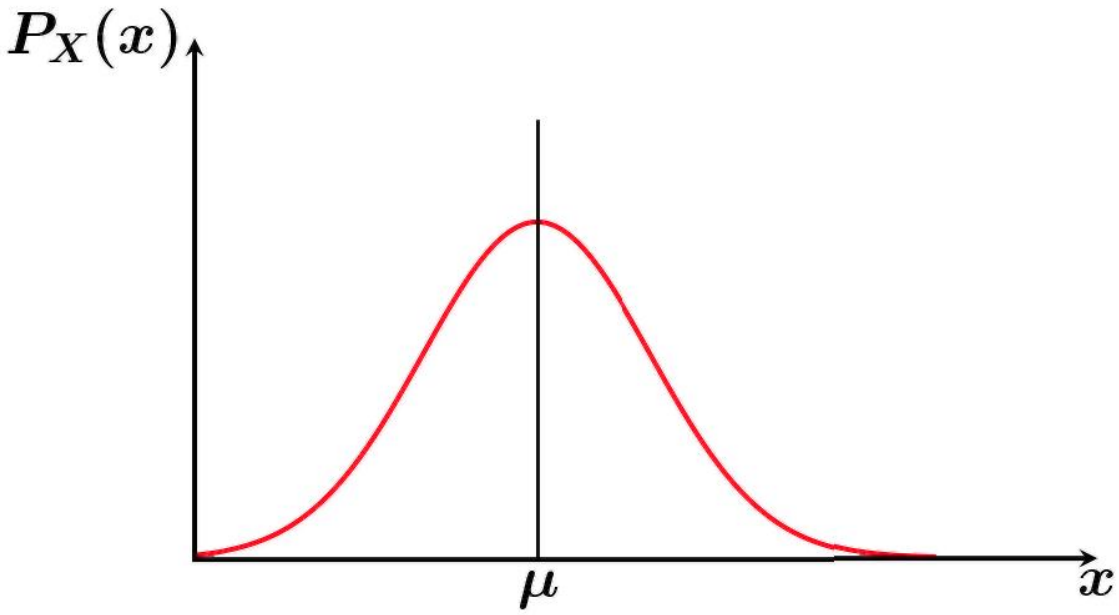
§76 Відновлення розподілів (частина 2)

Параметричне відновлення розподілу

- Розглянутий раніше спосіб відновлення розподілів для бінарних ознак **не годиться у випадку дійсних ознак**.
- Можна припустити, що розподіл має якийсь певний вигляд: пуасонівський, експоненціальний або нормальний, і спробувати відновити його. Це метод називається **методом параметричного відновлення розподілів**.

Нормальний розподіл:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Нормальний розподіл:

Повністю визначається значеннями двох параметрів: математичного очікування і дисперсії, які можна визначити:

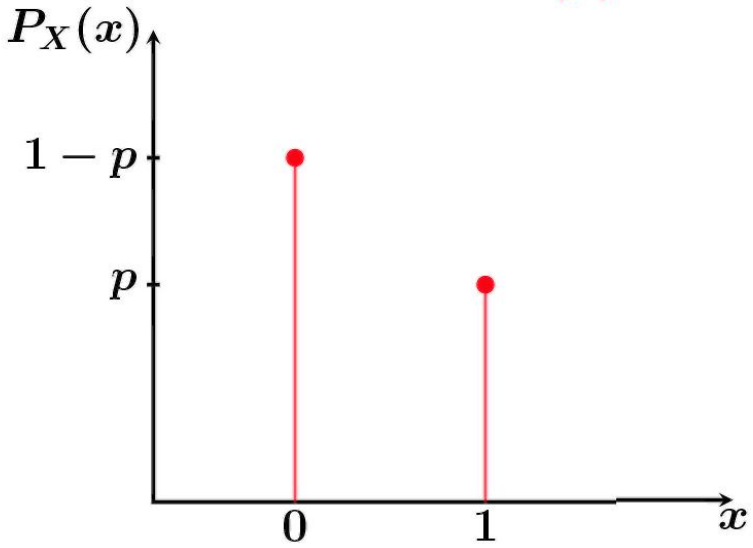
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

Незміщений варіант оцінки для дисперсії:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

Розподіл Бернуллі

$$X \sim Bernoulli(p)$$



$$P(x=1) = p$$
$$P(x=0) = 1-p$$

Характеризується одним параметром — імовірність того, що випадкова величина приймає значення 1.

Цей параметр можна оцінити:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} [x_i = 1]$$

Рекомендації з вибору розподілів

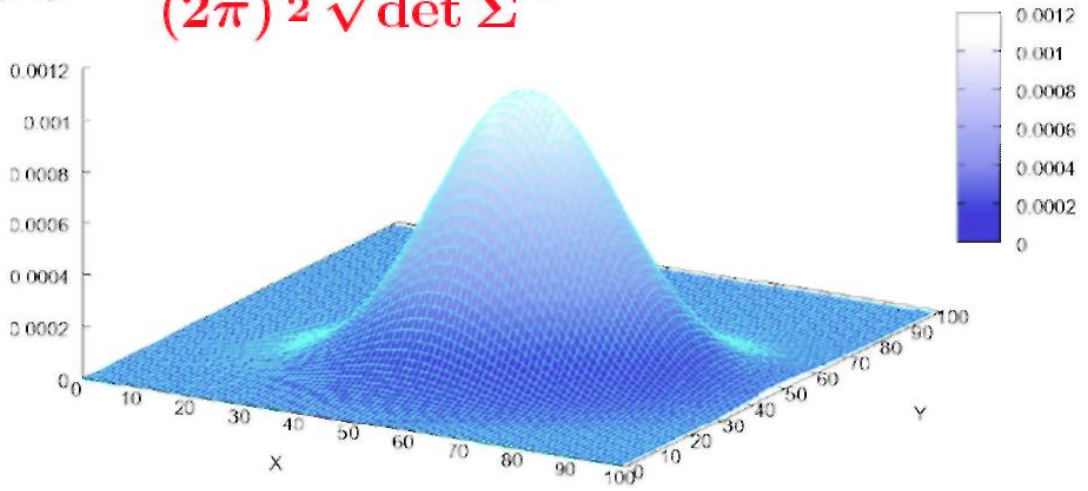
Загальні рекомендації з вибору розподілу при використанні методу параметричного відновлення:

- Якщо вирішується задача, пов'язана з текстами або якимись іншими **розрядженими дискретними ознаками**, то добре підходить мультиноміальний розподіл.
- Якщо в задачі є **неперервні ознаки** з невеликим розкидом, то можна спробувати використовувати **нормальний розподіл**.
- Для неперервних ознак з великим розкидом потрібні більш «розмазані», ніж нормальний, розподіли.

При цьому не обов'язково обмежуватися наївним байєсевським класифікатором. Проблему нестачі даних можна вирішувати за допомогою параметричної оцінки багатомірних розподілів, але рішення шукати в якімсь вузькому класі так, щоб воно визначалося невеликим числом параметрів

Приклад: багатомірний нормальний розподіл

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$



Параметри: вектор середніх μ і матриця коваріацій $\sum$

Недоліки підходу

- Виникає більше параметрів, чим в «наївному» підході
- В «наївному» підході параметри — це n середніх і n дисперсій, а у випадку багатомірного нормального розподілу — вектор розміру n і матриця розміру $n \times n$.
- Оцінка параметрів може вийти неправильної через нестачу даних
- Часто потрібно виконувати «нестійкі» операції - наприклад, обернення матриць, які майже вироджені

Існує також непараметричне відновлення густини.

§77 Мінімізація ризику

У цьому підрозділі розглянемо інший погляд на баєсову класифікацію й виконаємо узагальнення на випадок задачі регресії.

Баєсіва регресія

- Баєсів класифікатор визначається виразом:

$$a(x) = \operatorname*{argmax}_y P(y)P(x|y)$$

- Застосувати таку ж формулу у випадку регресії не можна, тому що навряд чи вийде відновити розподіл $P(x|y)$, оскільки y — **дійсне** число.
- Якщо скористатися при вирішенні задачі регресії виразом:

$$a(x) = \operatorname*{argmax}_y P(y|x)$$

то це буде відповідати пошуку максимуму функції густин за y при обраному x. Не очевидно, що це буде гарним рішенням задачі регресії.

Штрафи за помилки

Часто буває необхідно по-різному штрафувати алгоритм за різні типи помилок.

- **Задачі класифікації** нафтових родовищ із двома класами **«є нафта»** і **«немає нафти»** помилковий позитивний результат – більше критична помилка, тому що буріння свердловини вимагає величезних грошових і часових витрат.
- **У задачах регресії** штрафи за помилки ще **більш природні**: тому що шукану залежність ідеально відновити неможливо, потрібно саме мінімально відхилитися від її. У задачах регресії як міра відхилення часто використовуються квадратичні втрати (MSE) і сума модулів відхилення (MAE):

$$\mathbf{MSE} = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - a(x_i))^2$$

$$\mathbf{MAE} = \frac{1}{\ell}\sum_{i=1}^{\ell}[y_i - a(x_i)]$$

Більше загальний підхід

- Нехай для деякого об'єкта x необхідно зробити прогноз $a(x)$. Яка саме задача, задача регресії або класифікації, розглядається, не має значення. Нехай також y — правильна відповідь, а функція $L(y, a(x))$ визначає величину помилки алгоритму й задається залежно від розглянутої задачі й бажаних властивостей алгоритму.
- У задачі класифікації можна використовувати як функцію $L(y, a(x))$:

$$L(y, a(x)) = [y \neq a(x)].$$

Такий вибір функції приведе до того, що отриманий класифікатор буде вже розглянутим раніше баєсів класифікатором, і про це буде розказано пізніше.

- У задачі регресії використовується квадратична функція:

$$L(y, a(x)) = (y - a(x))^2.$$

Оптимальний баєсів класифікатор

Введем **функціонал ризику** $R(a, x)$, що визначається як умовне математичне очікування втрат при відомому $\boldsymbol{x}$ і відповіді алгоритму a:

$$R(a(x), x) = \mathbb{E}(L(y, a(x))|x)$$

Можна будувати відповіді алгоритму таким чином, щоб мінімізувати очікувані втрати:

$$a(x) = \underset{s}{\operatorname{argmin}} R(s, x)$$

- У випадку задачі класифікації:

$$R(a(x), x) = \mathbb{E}(L(y, a(x))|x) = \sum_{y \in Y} L(y, a(x)) P(y|x)$$

$$a(x) = \operatorname{argmin}_s R(s, x) = \operatorname{argmin}_s \sum_{y \in Y} L(y, s) P(y|x) = \operatorname{argmin}_s \sum_{y \in Y} L(y, s) P(y) P(x|y)$$

- Такий класифікатор називається **оптимальним баєсів класифікатором**, тому що він мінімізує очікувані втрати.
- Реальний класифікатор, звичайно, не буде оптимальним через використання ймовірнісних оцінок, а не істинних імовірностей.

Оптимальний баєсів регресор

Для задачі регресії вирази виглядають аналогічним чином:

$$R(a(x), x) = \mathbb{E}(L(y, a(x))|x) = \\ = \int_{y \in Y} L(y, a(x)) p(y|x) dy$$

$$\boxed{a(x)} = \operatorname{argmin}_s R(s, x) = \\ = \boxed{\operatorname{argmin}_s \int_{y \in Y} L(y, s) p(y|x) dy}$$

На практиці даний результат використовується не для вирішення задачі регресії, а щоб проаналізувати різні функції втрат.

Функціонал середнього ризику

Функціонал середнього ризику

$$R(a) = \mathbb{E}_x R(a(x), x)$$

дозволяє оцінити, наскільки добре працює алгоритм у середньому, а не для конкретного x.

Для визначеності далі розглядається випадок задачі класифікації об'єктів з дискретними ознаками.

$$R(a) = \sum_{x \in X} R(a(x), x) P(x)$$

Оскільки $R(s, x) \geq \min_s R(s, x)$, вірна наступна оцінка знизу для $R(a)$:

$$R(a) = \sum_{x \in X} R(a(x), x) P(x) \geq \sum_{x \in X} \min_s R(s, x) P(x).$$

Таким чином, оптимальний баєсів класифікатор мінімізує не тільки функціонал ризику, але й функціонал середнього ризику.

§78 Мінімізація ризику й аналіз функції втрат

Оптимальний баєсів класифікатор

Можна показати, що оптимальний баєсів класифікатор:

$$a(x) = \operatorname{argmin}_s R(s,x) = \operatorname{argmin}_s \sum_{y \in Y} L(y,s) P(y|x),$$

у випадку, якщо функція втрат дорівнює індикатору того, що прогноз алгоритму $a(x)$ не збігся із правильною відповіддю y:

$$L(y, a(x)) = [y \neq a(x)],$$

переходить у знайомий баєсів класифікатор.

Дійсно, якщо підставити функцію втрат у вираз, який мінімізується:

$$\sum_{y\in Y} L(s,y)P(y|x) = \sum_{y\in Y\setminus\{s\}} P(y|x) = \sum_{y\in Y} P(y|x) - P(s|x) \to \min_s \quad \Longrightarrow \quad P(s|x) \to \max_s,$$

Таким чином, класифікатор має вигляд:

$$a(x) = \operatorname{argmax}_y P(y|x) = \operatorname{argmax}_y P(y)P(x|y),$$

тобто є баєсів класифікатором.

Квадратична функція втрат у регресії

Виявляється, такий підхід годиться для аналізу різних функцій втрат у задачі регресії. Наприклад, у випадку квадратичної функції втрат:

$$\int_Y (t-y)^2 p(y|x)dy \to \min_t .$$

Мінімум можна знайти, якщо прирівняти похідну по t до нуля:

$$\frac{\partial}{\partial t}\int_Y (t-y)^2 p(y|x)dy = 2\int_Y (t-y)p(y|x)dy = 2\left(t\int_Y p(y|x)dy - \int_Y yp(y|x)dy\right) = 0.$$

Тому що $p(y|x)$ — густина імовірності, $\int_Y p(y|x)dy = 1$:

$$a(x) = t = \int_Y yp(y|x)dy = \mathbb{E}(y|x).$$

Таким чином, прогноз алгоритму повинен дорівнювати умовному математичному очікуванню E($y|x$).

Абсолютне відхилення

У випадку, коли функція втрат - абсолютне відхилення:

$$\int_Y |t-y|p(y|x)dy \to \min_t,$$

розрахунки виконуються точно також. Єдиний нюанс полягає в тому, що модуль не має похідної в нулі, тому точку $y = t$ варто завчасно виключити з області інтегрування (важливо, що це не змінить значення інтеграла):

$$\frac{\partial}{\partial t}\int_Y |t-y|p(y|x)dy = \frac{\partial}{\partial t}\int_{Y\setminus\{t\}} |t-y|p(y|x)dy = \int_{Y\setminus\{t\}} \text{sign}(t-y)p(y|x)dy =$$
$$= \int_{\{t>y\}} p(y|x)dy - \int_{\{t<y\}} p(y|x)dy = P(\{t>y\}|x) - P(\{t<y\}|x) = 0.$$

Таким чином, з огляду на, що $P(\{t=y\}|x) = 0$, можна одержати:

$$P(\{t>y\}|x) = P(\{t<y\}|x) = \frac{1}{2}.$$

Інакше кажучи, відповідь алгоритму оцінює 1/2 квантиль (медіану).

Оцінка ймовірності

Розглядається задача бінарної класифікації $Y = \{0, 1\}$. Необхідно, щоб алгоритм класифікації оцінював імовірність того, що об'єкт належить до першого класу $p = P(1|x)$. Виявляється, що одержати необхідний результат можна, вибравши як функцію втрат так звану функцію Log loss:

$$L(y, a(x)) = -y \ln a(x) - (1 - y) \ln(1 - a(x))$$

Умова мінімальності втрат тоді приймає вигляд:

$$\sum_{y \in Y} \Big(- y \ln t - (1 - y) \ln(1 - t) \Big) P(y|x) = -(1 - p) \ln(1 - t) - p \ln t \to \min_t,$$

де використане позначення $p = P(1|x)$. Мінімум можна знайти обчисленням похідної за t:

$$a(x) = t = p,$$

Отримуємо необхідний результат:

$$\frac{\partial}{\partial t}\big(-(1-p)\ln(1-t)-p\ln t\big)=\frac{1-p}{1-t}-\frac{p}{t}=\frac{t-p}{(1-t)t}=0.$$

тобто відповідь алгоритму t повинен дорівнювати ймовірності p того, що об'єкт належить до першого класу.

Обґрунтування методу аналізу функції втрат

Варто нагадати, що в баєсів класифікації мінімізується саме функціонал середнього ризику:

$$R(a) = \mathbb{E}_{x,y} L(y, a(x)).$$

Оскільки помилка Q на навчальній вибірці є емпіричною оцінкою функціонала середнього ризику:

$$Q = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i)) \sim \mathbb{E}_{x,y} L(y, a(x)),$$

результати наведеного вище методу аналізу функції втрат залишаються вірні не тільки у випадку використання баєсів класифікатора або баєсів регресії, але й для довільного методу рішення, у ході якого мінімізується помилка на навчальній вибірці.

§79 Комп'ютерний проект 6: Вибір сімейства розподілів в наївному Байєсі