2021

Лекция 10. Наивный байесовский классификатор

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

ПАПУЛИН С.Ю. (PAPULIN.STUDY@YANDEX.RU) I

Дополнительные материалы доступны по ссылке: https://github.com/mlmethods

Содержание

1. Наивный байесовский классификатор

Теорема Байеса:

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

Если рассматривать теорему Байеса в контексте задачи классификации для одного наблюдения, то её можно представить следующим образом

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})},$$

где $\boldsymbol{x}$ – вектор признаков размера p; y – метка класс. Для бинарной классификации примем, что $y \in \{0,1\}$, а для многоклассовой – $y \in \{1, \dots, K\}$.

Наша задача найти вероятности $p(y|\boldsymbol{x})$ для каждого класса и выбрать максимальную в качестве предсказания для наблюдения $\boldsymbol{x}$:

$$\hat{y} = \underset{k}{\operatorname{argmax}}\, p(y = k|\boldsymbol{x}) = \underset{k}{\operatorname{argmax}} \left[\frac{p(\boldsymbol{x}|y = k)p(y = k)}{p(\boldsymbol{x})}\right].$$

Так как знаменатель, $p(\boldsymbol{x})$, не зависит от y, то эту вероятность будем рассматривать как константу:

$$p(y|\boldsymbol{x}) \propto p(\boldsymbol{x}|y)p(y)$$

Тогда выражение для предсказания примет вид

$$\hat{y} = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}|y = k)p(y = k).$$

Если предположить, что признаки независимы, то вероятность появления $\boldsymbol{x}$ при заданном классе y можно представить как

$$p(\boldsymbol{x}|y) = \prod_{j=1}^{p} p(x_j|y),$$

где x_j – значение j-ого признака.

В связи с предположением о независимости признаков классификатор называется «наивным». В результате предсказание будет иметь вид

$$\hat{y} = \underset{k}{\operatorname{argmax}}\, p(y = k) \prod_{j=1}^{p} p(x_j|y = k).$$

Теперь осталось найти оценки вероятностей $p(y)$ и $p(x_j|y)$. Для этого можно воспользоваться методом максимального правдоподобия (MLE). Ранее для логистической регрессии функцию правдоподобия записывали следующим образом

$$L(y|X;\theta) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i;\theta),$$

где n – количество независимых одинаково распределённых наблюдений, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$; X – матрица признаков.

Так как апостериорная вероятность, $p(y_i|\boldsymbol{x}_i)$ пропорциональна совместной вероятности, $p(\boldsymbol{x}_i|y_i)p(y_i)$, как было показано ранее, представим функцию правдоподобия в следующем виде

$$L(y|X;\theta) \propto \prod_{i=1}^{n} p(\boldsymbol{x}_i|y_i;\theta)p(y_i;\theta) = \prod_{i=1}^{n} p(y_i;\theta) \prod_{j=1}^{p} p\big(x_{ij}|y_i;\theta\big).$$

В итоге нам необходимо найти такие параметры θ, при которых функция правдоподобия будет иметь максимальное значение

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(y|X;\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i;\theta) \prod_{j=1}^{p} p\big(x_{ij}|y_i;\theta\big)$$

или через логарифм

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log L(y|X;\theta) = \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \log p(y_i;\theta) + \sum_{i=1}^{n}\sum_{j=1}^{p} \log p\big(x_{ij}|y_i;\theta\big)\right]$$

при условии

$$\sum_{k=1}^{K} p(y_i = k;\theta) = 1.$$

После того как найдем $\hat{\theta}$, можно будет рассчитать $p\big(y_i;\hat{\theta}\big)$ и $p\big(\boldsymbol{x}_i|y_i;\hat{\theta}\big)$. Таким образом, предсказания для некоторого наблюдения x_* примет вид

$$\hat{y}_* = \underset{k}{\operatorname{argmax}}\, p\big(y_* = k;\hat{\theta}\big) \prod_{j=1}^{p} p\big(x_{*j}|y = k;\hat{\theta}\big).$$

Так как значения вектора $\boldsymbol{x}$ могут быть как непрерывными, так и дискретными, далее рассмотрим три варианта:

- Для непрерывных значений $\boldsymbol{x} \in \mathbb{R}^p$ – распределение Гаусса (нормальный закон)
- Для бинарных значений $\boldsymbol{x} \in \{0,1\}^p$ – распределение Бернулли
- Для дискетных значений $\boldsymbol{x} \in \mathbb{N}^p$ – полиномиальное (мультиномиальное) распределение

Замечание

Оценки параметров θ также можно найти посредством метода апостериорного максимума (MAP), если сделать предположения о функции распределения этих параметров. Сопряженные априорные вероятности:

- для нормального закона – нормальный закон распределения
- для Бернулли – Бета-распределение
- для полиномиального – распределение Дирихле

2. Наивный байесовский классификатор с распределением Гаусса

Если признаками являются вещественные значения, то их можно представить в виде вектора размера p, то есть $\boldsymbol{x} \in \mathbb{R}^p$. Далее рассмотрим частный случай с бинарным классификатором, в котором $y \in \{0,1\}$.

Для вещественных значений распределение вероятности представляется в виде плотности вероятности. Примем, что признаки независимы и подчиняются нормальному закону распределения (распределению Гаусса). Тогда плотность вероятности для j-ого признака при заданном классе $y = k$ есть

$$p(x_j|y=k) = \mathcal{N}(x_j|\mu_{kj}, \sigma_{kj}^2) = \frac{1}{2\pi\sigma_{kj}^2}\exp\left[-\frac{1}{2\sigma_{kj}^2}(x_j - \mu_{kj})^2\right],$$

где μ_{kj} и σ_{kj}^2 – среднее значение и дисперсия j-ого признака для класса k.

Функция правдоподобия бинарного классификатора записывается следующим образом

$$L(y|X;\theta) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i;\theta) \propto \prod_{i=1}^{n} \underbrace{[p(y=1)p(\boldsymbol{x}_i|y=1)]^{y_i}}_{y_i=1} \underbrace{[p(y=0)p(\boldsymbol{x}_i|y=0)]^{1-y_i}}_{y_i=0}.$$

Обозначим вероятность появления класса 1 при заданном θ как

$$p(y=1;\theta) = \pi.$$

Для наивного байесовского классификатора выражение выше примет вид

$$L(y|X;\theta) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i;\theta) \propto \prod_{i=1}^{n} \underbrace{\prod_{j=1}^{p}[\pi\mathcal{N}(x_{ij}|\mu_{1j},\sigma_{1j}^2)]^{y_i}}_{y_i=1} \underbrace{\prod_{j=1}^{p}[(1-\pi)\mathcal{N}(x_{ij}|\mu_{0j},\sigma_{0j}^2)]^{1-y_i}}_{y_i=0}$$

Запишем его в логарифмическом виде

$$\log L(y|X;\theta) \propto \underbrace{\sum_{i=1}^{n} y_i \log\pi + \sum_{i=1}^{n}(1-y_i)\log(1-\pi)}_{\pi} + \underbrace{\sum_{i=1}^{n} y_i \sum_{j=1}^{p} \log\mathcal{N}(x_{ij}|\mu_{1j},\sigma_{1j}^2)}_{\mu_{1j}\text{ и }\sigma_{1j}^2}$$
$$+ \underbrace{\sum_{i=1}^{n}(1-y_i)\sum_{j=1}^{p}\log\mathcal{N}(x_{ij}|\mu_{0j},\sigma_{0j}^2)}_{\mu_{0j}\text{ и }\sigma_{0j}^2}$$

Таким образом, задача сводится к поиску таких параметров π, μ_{1j}, μ_{0j}, σ_{1j}^2 и σ_{0j}^2, при которых функция правдоподобия примет максимальное значение

$$\underbrace{\hat{\pi}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\sigma}}_1^2, \hat{\boldsymbol{\sigma}}_0^2}_{\hat{\theta}} = \underset{\theta}{\operatorname{argmax}} \log L(y|X;\theta),$$

где $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\sigma}}_1^2, \hat{\boldsymbol{\sigma}}_0^2$ – векторы оценок параметров по всем признакам.

Для этого возьмем частные производные по параметрам, приравняем их к нулю и решим систему уравнений. Если это сделаем, то получим следующие оценки:

$$\hat{\pi} = \frac{n_1}{n_1 + n_0},$$

$$\hat{\mu}_{1j} = \frac{1}{n_1}\sum_{i=1}^{n} 1(y_i = 1)x_{ij},$$

$$\hat{\mu}_{0j} = \frac{1}{n_0}\sum_{i=1}^{n} 1(y_i = 0)x_{ij},$$

$$\hat{\sigma}^2_{1j} = \frac{1}{n_1}\sum_{i=1}^{n} 1(y_i = 1)\left(x_{ij} - \hat{\mu}_{1j}\right)^2,$$

$$\hat{\sigma}^2_{0j} = \frac{1}{n_0}\sum_{i=1}^{n} 1(y_i = 0)\left(x_{ij} - \hat{\mu}_{0j}\right)^2,$$

где n_1 и n_0 – количество наблюдений класса 1 и 0, соответственно.

Вывод

Оценка параметра π

$$\frac{\partial \log L(y, X; \theta)}{\partial \pi} = \frac{\partial}{\partial \pi}\sum_{i=1}^{n} y_i \log \pi + \sum_{i=1}^{n}(1 - y_i)\log(1 - \pi) =$$
$$= \frac{1}{\pi}\sum_{i=1}^{n} y_i - \frac{1}{1-\pi}\sum_{i=1}^{n}(1 - y_i) = \frac{n_1}{\pi} - \frac{n_0}{1-\pi} = 0$$

$$\hat{\pi} = \frac{n_1}{n_1 + n_0}$$

Оценка параметров μ_{1j} и μ_{0j}

$$\frac{\partial \log L(y, X; \theta)}{\partial \mu_{1l}} = \frac{\partial}{\partial \mu_{1l}}\sum_{i=1}^{n} y_i \sum_{j=1}^{p} \log \mathcal{N}\left(x_{ij} | \mu_{1j}, \sigma^2_{1j}\right) = 0$$

$$\log \mathcal{N}\left(x_{ij} | \mu_{1j}, \sigma^2_{1j}\right) = \log \frac{1}{2\pi\sigma^2_{1j}} \exp\left[-\frac{1}{2\sigma^2_{1j}}(x_j - \mu_{1j})^2\right]$$
$$= \underbrace{-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2_{1j}}_{\text{константы по } \mu_{1j}} - \frac{1}{2\sigma^2_{1j}}(x_j - \mu_{1j})^2$$

$$\frac{\partial \log L(y, X; \theta)}{\partial \mu_{1l}} = \frac{\partial}{\partial \mu_{1l}}\sum_{i=1}^{n} y_i \sum_{j=1}^{p}\left[-\frac{1}{2\sigma^2_{j1}}(x_{ij} - \mu_{j1})^2 + \text{const}\right] = \frac{1}{\sigma^2_{1l}}\sum_{i=1}^{n} y_i(x_{il} - \mu_{1l}) = 0$$

$$\sum_{i=1}^{n} y_i(x_{il} - \mu_{1l}) = \sum_{i=1}^{n} y_i x_{il} - \mu_{1l}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} 1(y_i = 1)x_{il} - n_1\mu_{1l} = 0$$

$$\hat{\mu}_{1l} = \frac{1}{n_1}\sum_{i=1}^{n} 1(y_i = 1)x_{ij}$$

Аналогичным образом находим

$$\hat{\mu}_{0l} = \frac{1}{n_0}\sum_{i=1}^{n} 1(y_i = 0)x_{il}$$

Оценка параметров σ_{1j}^2 и σ_{0j}^2

$$\frac{\partial \log L(y, X; \theta)}{\partial \sigma_{1l}^2} = \frac{\partial}{\partial \sigma_{1l}^2}\sum_{i=1}^{n} y_i \sum_{j=1}^{p} \log \mathcal{N}\left(x_{ij}\middle|\mu_{1j}, \sigma_{1j}^2\right) = 0$$

Далее самостоятельно

В результате предсказание для некоторого наблюдения $\boldsymbol{x}_*$ будет иметь вид

$$\hat{y}_* = \underset{k}{\operatorname{argmax}}\, \hat{\pi}_k \prod_{j=1}^{p} \mathcal{N}\left(x_{*j}\middle|\hat{\mu}_{kj}, \hat{\sigma}_{kj}^2\right).$$

Если рассматривать наблюдение как вектор случайных величин с нормальным законом распределения, то функция плотности вероятности для наблюдения $\boldsymbol{x}$ при заданном классе $y = k$ будет иметь вид многомерного нормального распределения

$$p(\boldsymbol{x}|y = k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T {\Sigma_k}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right],$$

где $\boldsymbol{x}$ – вектор признаков; $\boldsymbol{\mu}_k$ – вектор средних значений по каждому признаку для класса k; Σ_k – ковариационная матрица размера $p \times p$ для класса k; $|\Sigma_k|$ – детерминант матрицы Σ_k.

Ковариационная матрица определяет разброс относительно центра класса, $\boldsymbol{\mu}_k$. Элементом матрицы Σ_k является ковариация вида

$$\sigma_{ijk} = \operatorname{cov}\left[\mathrm{x}_{ik}, \mathrm{x}_{jk}\right],$$

где x_i – значения i-го признака по всем наблюдениям класса k, то есть $\mathrm{x}_{ik} = [x_{1ik}\ x_{2ik} \dots\ x_{nik}\ ]^T$.

В данном случае функция правдоподобия представляется как

$$L(y|X; \theta) \propto \prod_{i=1}^{n} \underbrace{[\pi \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_1, \Sigma_1)]^{y_i}}_{y_i=1} \underbrace{[(1-\pi)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_0, \Sigma_0)]^{1-y_i}}_{y_i=0}$$

или через логарифм

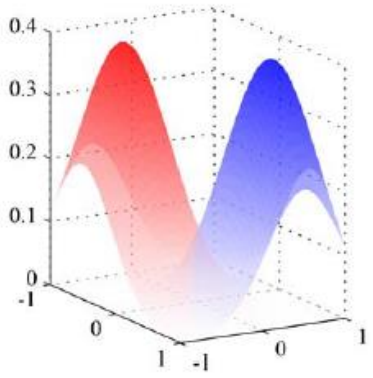
$$\log L(y|X; \theta) \propto \sum_{i=1}^{n} \left[y_i \log \pi \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_1, \Sigma_1) + (1 - y_i)\log\left((1-\pi)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_0, \Sigma_0)\right)\right].$$

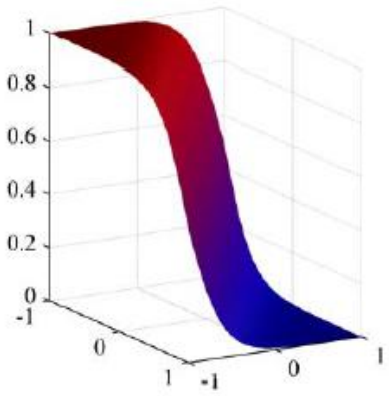
Далее посредством метода максимального правдоподобия аналогичным образом производится оценка параметров.

Такой подход с многомерным нормальным распределением называется *квадратичный дискриминантный анализ* (Quadratic Discriminant Analysis – **QDA**). Если принять, что $\Sigma_1 = \Sigma_0$, то получим *линейный дискриминантный анализ* (Linear Discriminant Analysis – **LDA**).

Пример LDA:

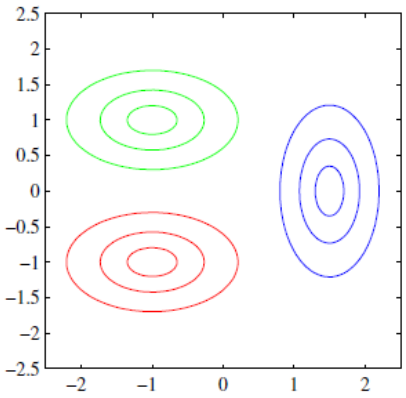
- Слева: условные плотности вероятностей по двум классам $p(x|y=1)$ и $p(x|y=0)$
- Справа: апостериорная вероятность $p(y=1|x)$

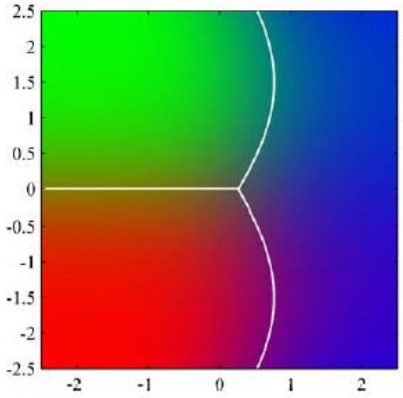




Пример QDA:

- Слева: условные распределения по трем классам
  - Зеленным и красным: одна ковариационная матрица
  - Зеленным и синим: разные ковариационные матрицы
- Справа: апостериорные вероятности и границы принятия решений





Замечание

Если матрицы Σ_k являются диагональными, то этот случай будет соответствовать наивному байесовскому классификатору, рассмотренному ранее. Это следует из того, что при нормальном законе распределения если признаки некоррелированны, то они независимы.

3. Наивный байесовский классификатор с распределением Бернулли

Ранее был приведен случай, когда $\boldsymbol{x}$ является вектором вещественных значений. Если $\boldsymbol{x}$ представить как бинарный вектор

$$\boldsymbol{x} = [x_1\ x_2\ \dots\ x_p]^T,$$

где $x_j \in \{0,1\}$, то вероятность появления $\boldsymbol{x}$ при заданном классе y будет

$$p(\boldsymbol{x}|y) = \prod_{j=1}^{p} \mu_{y,j}^{x_j}\left(1-\mu_{y,j}\right)^{\left(1-x_j\right)},$$

где $\mu_{y,j}$ – вероятность появления 1 для j-ого признака класса $y \in \{1,\dots,K\}$, то есть $p\left(x_j = 1|y\right)$.

В качестве примера бинарного вектора $\boldsymbol{x}$ можно привести бинарный вектор вхождения термов в документ, в котором каждый элемент соответствует определенному терму, а значение 0 или 1 указывает на отсутствие или присутствие терма в документе. Под y в данном случае может подразумеваться тематический класс документа.

Функцию правдоподобия можно записать следующим образом

$$L(y|X;\underbrace{\boldsymbol{\pi},\boldsymbol{\mu}}_{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i;\boldsymbol{\pi},\boldsymbol{\mu}) \propto \prod_{i=1}^{n} p(y_i;\boldsymbol{\pi},\boldsymbol{\mu})p(x_i|y_i;\boldsymbol{\pi},\boldsymbol{\mu}) = \prod_{i=1}^{n} \pi_{y_i} \prod_{j=1}^{p} \mu_{y_i,j}^{x_{i,j}}\left(1-\mu_{y_i,j}\right)^{\left(1-x_{i,j}\right)}$$

где π_{y_i} – вероятность встретить класс $y_i \in \{1,\dots,K\}$.

Неизвестные параметры θ есть $\boldsymbol{\pi}$ и $\boldsymbol{\mu}$.

Замечание

Можно использовать функцию правдоподобия $L(X,y|\theta)$, которая будет соответствовать $L(y|X;\theta)$ следующим образом

$$L(y|X;\theta) \propto L(X,y|\theta)$$

Таким образом, оценка параметров будет иметь вид

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(y|X;\theta) = \underset{\theta}{\operatorname{argmax}} L(X,y|\theta)$$

Выразим функцию правдоподобия через логарифм

$$\log L(y|X;\underbrace{\boldsymbol{\pi},\boldsymbol{\mu}}_{\theta}) \propto \log L(X,y|\underbrace{\boldsymbol{\pi},\boldsymbol{\mu}}_{\theta}) = \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{i,j}\log \mu_{y_i,j} + \left(1-x_{i,j}\right)\log\left(1-\mu_{y_i,j}\right)\right)$$

Соответственно необходимо найти такие $\boldsymbol{\pi}$ и $\boldsymbol{\mu}$, при которых функция правдоподобия примет максимальное значение:

$$\underbrace{\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\mu}}}_{\hat{\theta}} = \underset{\boldsymbol{\pi},\boldsymbol{\mu}}{\operatorname{argmax}} \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{i,j}\log \mu_{y_i,j} + \left(1-x_{i,j}\right)\log\left(1-\mu_{y_i,j}\right)\right)$$

при условии

$$\sum_{k=1}^{K} \pi_k = 1$$

В результате получаем следующие оценки параметров

$$\hat{\mu}_{k,j} = \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = y)\, x_{i,j}}{\sum_{i=1}^{n} \mathbf{1}(y_i = y)}$$

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = k)}{n}$$

Вывод оценок параметров $\hat{\mu}_{k,j}$ и $\hat{\pi}_k$

Функция правдоподобия

$$\log L(X, y| \underbrace{\boldsymbol{\pi}, \boldsymbol{\mu}}_{\theta}) = \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{i,j} \log \mu_{y_i,j} + \left(1 - x_{i,j}\right) \log\left(1 - \mu_{y_i,j}\right)\right)$$

Оценка параметра π_k:

$$\sum_{i=1}^{n} \log \pi_{y_i} = \sum_{i=1}^{n} \sum_{k=1}^{K} \log \pi_k \cdot 1(y_i = k) = \sum_{k=1}^{K} \sum_{i=1}^{n} \log \pi_k \cdot 1(y_i = k) = \sum_{k=1}^{K} \log \pi_k \sum_{i=1}^{n} 1(y_i = k)$$

$$\begin{cases} \dfrac{\partial \log L(\theta)}{\partial \pi_c} + \dfrac{\partial}{\partial \pi_c} \lambda \left[\sum_{k=1}^{K} \pi_k - 1\right] = 0 \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases}$$

Решаем систему (с применением множителя Лагранжа, так как есть ограничение по π_k)

$$\begin{cases} \dfrac{1}{\pi_c} \sum_{i=1}^{n} 1(y_i = c) + \lambda = 0 \\ \sum_{k=1}^{K} \pi_k = 1. \end{cases}$$

Выразим π_c следующим образом

$$\pi_c = -\frac{1}{\lambda} \sum_{i=1}^{n} 1(y_i = c).$$

Подставим полученное выражение во второе уравнение системы

$$-\frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{n} 1(y_i = k) = -\frac{1}{\lambda} n = 1.$$

Получаем

$$\lambda = -n.$$

Подставляем полученное значение в первое уравнение системы

$$\frac{1}{\pi_c} \sum_{i=1}^{n} \mathbf{1}(y_i = c) - n = 0$$

Из этого выводим итоговое выражение

$$\hat{\pi}_c = \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = c)}{n}$$

Оценка параметра $\mu_{y,j}$:

$$\nabla_{\mu_{y,j}} \log L(\theta) = \frac{\partial \log L(\theta)}{\partial \boldsymbol{\mu_{k,j}}} = \mathbf{0}$$

$$\frac{\partial \log L(\theta)}{\partial \mu_{c,j}} = \frac{\partial}{\partial \mu_{c,j}} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{i,j} \log \mu_{y_i,j} + \left(1 - x_{i,j}\right) \log\left(1 - \mu_{y_i,j}\right)\right) = 0$$

$$\frac{\partial}{\partial \mu_{c,j}} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{i,j} \log \mu_{y_i,j} + \left(1 - x_{i,j}\right) \log\left(1 - \mu_{y_i,j}\right)\right) = \sum_{i=1}^{n} 1(y_i = c) \left[\frac{x_{i,j}}{\mu_{c,j}} - \frac{1 - x_{i,j}}{1 - \mu_{c,j}}\right] = 0$$

$$\sum_{i=1}^{n} 1(y_i = c)\left[x_{i,j}\left(1 - \mu_{c,j}\right) - \left(1 - x_{i,j}\right)\mu_{c,j}\right] = \sum_{i=1}^{n} 1(y_i = c)\left[x_{i,j} - \mu_{c,j}\right]$$

$$= \sum_{i=1}^{n} 1(y_i = c) x_{i,j} - \mu_{c,j} \sum_{i=1}^{n} 1(y_i = c) = 0$$

$$\hat{\mu}_{c,j} = \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = c)\, x_{i,j}}{\sum_{i=1}^{n} \mathbf{1}(y_i = c)}$$

Классификация текстовых документов

Представим коллекцию текстовых документов как набор векторов размером $N = |V|$, где $|V|$ – размер словаря. Для одного документа это можно записать следующим образом:

$$d \to b = (b_1, \dots, b_N),$$

где

$$b_j \in \{0,1\}.$$

Значение 1 или 0 определяет присутствие или отсутствие терма в документе, соответственно.

Модель Бернулли

$$x_{i,j} \sim b_{i,j}$$

$$\mu_{y,j} \sim p(t_i|c)$$

$$\pi_y \sim p(c)$$

Для одного документа:

$$p(d|c) \sim p(\boldsymbol{b}|c) = \prod_{j=1}^{N}\left[p(t_j|c)^{b_j} \cdot \left(1 - p(t_j|c)\right)^{(1-b_j)}\right].$$

Обучение

Для коллекции документов функция правдоподобия будет иметь вид:

$$L(y|D;\theta) = p(c|D;\theta) \propto \prod_{i=1}^{N_D} p(c_i)p(d_i|c_i) = \prod_{i=1}^{N_D} p(c_i) \prod_{j=1}^{N}\left[p(t_i|c_i)^{b_{ij}} \cdot \left(1 - p(t_j|c_i)\right)^{(1-b_{ij})}\right],$$

где N_D – количество документов в коллекции; c_i – класс документа d_i.

Неизвестными параметрами θ в данном случае являются $p(c)$ и $p(t_j|c)$. Для их оценки используется метод максимального правдоподобия (MLE), как было показано ранее.

Оценка параметров:

- Оценка вероятности встретить документ класса c:

$$\hat{p}(\mathrm{c}) = \frac{N_\mathrm{c}}{N_D},$$

где N_c – количество документов класса c.

- Оценка вероятности встретить терм t_j в документах класса c:

$$\hat{p}(t_j|\mathrm{c}) = \frac{\mathrm{df}_\mathrm{c}(t_j)}{N_\mathrm{c}}$$

где $\mathrm{df}_\mathrm{c}(t_j)$ – количество документов класса c, содержащих терм t_j.

Если в документе встречается терм, для которого $\mathrm{df}_\mathrm{c}(t_j) = 0$, то $p(d_j|c) = 0$ и $p(c|d_j) = 0$. Чтобы избежать нулевых вероятностей используется сглаженная версия вычисления оценки вероятности:

$$\hat{p}_{smooth}(t_j|c) = \frac{\mathrm{df}_c(t_j) + 1}{N_c + 2}$$

Замечание

В общем случае после обучения может получиться так, что в коллекции документов для некоторых классов будут отсутствовать некоторые термы из словаря. Так как вероятность появления документа при заданном классе рассчитывается как произведение вероятностей появления каждого отдельного терма при том же условии, то отсутствие терма в обучающем множестве для этого класса сводит всю вероятность к нулю.

Предсказание

Для некоторого нового документа d_*:

$$\hat{y} = \underset{c}{\operatorname{argmax}}\, p(\mathrm{c}|d_*) = \underset{c}{\operatorname{argmax}}\, p(c)p(d_*|c) = \underset{c}{\operatorname{argmax}} \left[\hat{p}(c) \cdot \prod_{j=1}^{N} \left[\hat{p}(t_j|c)^{b_{*j}} \cdot \left(1 - \hat{p}(t_j|c)\right)^{(1-b_{*j})} \right] \right]$$

4. Наивный байесовский классификатор с полиномиальным распределением

Если вектор $\boldsymbol{x}$ является вектором натуральных чисел, то есть

$$\boldsymbol{x} = \left[x_1\ x_2\ \ldots\ x_p\right]^T,$$

где $x_j \in \mathbb{N}$, то вероятность появления $\boldsymbol{x}$ при заданном классе y будет

$$p(\boldsymbol{x}|y) = \prod_{j=1}^{p} \mu_{y,j}^{x_j}$$

В этом случае апостериорную вероятность можно выразить следующим образом

$$L(y|X; \underbrace{\boldsymbol{\pi}, \boldsymbol{\mu}}_{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i; \boldsymbol{\pi}, \boldsymbol{\mu}) \propto \prod_{i=1}^{n} p(y_i;) p(\boldsymbol{x}_i|y_i) = \prod_{i=1}^{n} \pi_{y_i} \prod_{j=1}^{p} \mu_{y_i,j}^{x_{i,j}}$$

Неизвестные параметры: $\boldsymbol{\pi}$ и $\boldsymbol{\mu}$

Нам необходимо найти такие параметры θ, при которых функция правдоподобия примет максимальное значение

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log L(\theta).$$

Функция правдоподобия в логарифмической форме будет иметь вид

$$\log L(y|X; \underbrace{\boldsymbol{\pi}, \boldsymbol{\mu}}_{\theta}) \propto \log L(X, y|\theta) = \log p(X, y|\theta) = \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n} \sum_{j=1}^{p} x_{i,j} \log \mu_{y_i,j}.$$

Оценка параметра π_k:

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = k)}{N}$$

Оценка параметра $\mu_{k,j}$:

$$\hat{\mu}_{k,j} = \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = k)\, x_{i,j}}{\sum_{j=1}^{p} \sum_{i=1}^{n} \mathbf{1}(y_i = k) x_{i,j}}$$

Классификация текстовых документов

$$d \to x = (x_1, \ldots, x_N),$$

где

$$x_i \in \mathbb{N}_{0+} \text{ или } \mathbb{R}_{0+}$$

Для одного документа:

$$p(d|c) = \prod_{j=1}^{N} p(t_j|c)^{x_j}$$

Замечание

Полиномиальное (мультиномиальное) распределение позволяет оценить вероятность наступления события с определенным исходом некоторое количество раз. Для текстового документа мы оцениваем вероятность появления терма некоторое количество раз. Количество термов представляется в виде вектора x:

$$x = (x_1, \dots, x_N), \text{где } x_i \in \mathbb{N}_{0+}$$

Как правило, для текстовых документов вместо вектора количества вхождения термов используется вектор TF-IDF, то есть вектор вещественных значений:

$$x = (x_1, \dots, x_N), \text{где } x_i \in \mathbb{R}_{0+}$$

Это, казалось бы, нарушает условие применения полиномиального распределения. Тем не менее на практике часто именно использование вектора TF-IDF дает более приемлемый результат при классификации текстовых документов.

Обучение

Для коллекции документов:

$$L(y|D;\theta) = p(c|D;\theta) \propto \prod_{i=1}^{N_D} p(c_i)p(d_i|c_i) = \prod_{i=1}^{N_D} p(c_i) \prod_{j=1}^{N} p(t_{ij}|c_i)^{x_{ij}}$$

Для оценки вероятностей $p(c)$ и $p(t_j|c)$ используется метод максимального правдоподобия (MLE).

Оценка параметров:

- Оценка вероятности встретить документ класса c:

$$\hat{p}(\mathrm{c}) = \frac{N_\mathrm{c}}{N_D},$$

где N_c – количество документов класса c.

- Оценка вероятности встретить терм t_j в документах класса c:

$$\hat{p}(t_j|\mathrm{c}) = \frac{\mathrm{tf}_c(t_j)}{\sum_{k=1}^{N} \mathrm{tf}_c(t_k)} = \frac{\mathrm{tf}_c(t_j)}{n_c}$$

где $\mathrm{tf}_c(t_j)$ – количество терма t_j в документах класса c; n_c – количество термов в документах класса c.

Чтобы избежать нулевых вероятностей:

$$\hat{p}_{smooth}^{\alpha=1}(t_i|c) = \frac{1 + \mathrm{tf}_c(t_i)}{N + \sum_{k=1}^{N} \mathrm{tf}_c(t_k)}$$

или в общем виде

$$\hat{p}_{smooth}^{\alpha}(t_i|c) = \frac{\alpha + \mathrm{tf}_c(t_i)}{\alpha N + \sum_{k=1}^{N} \mathrm{tf}_c(t_k)}$$

Предсказание

Для некоторого нового документа d_*:

$$\hat{y} = \underset{c}{\operatorname{argmax}}\, p(\mathrm{c}|d_*) = \underset{c}{\operatorname{argmax}}\, p(c)p(d_*|c) = \underset{c}{\operatorname{argmax}} \left[\hat{p}(c) \cdot \prod_{j=1}^{N} \hat{p}(t_j|c)^{x_{*j}} \right]$$

Список литературы

1. Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg. URL: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
2. Text Classification using Naive Bayes by Hiroshi Shimodaira. URL: https://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn07-notes-nup.pdf
3. Chapter 13 Text classification and Naive Bayes // Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, Cambridge University Press. 2008. URL: http://www-nlp.stanford.edu/IR-book/