Машинное обучение, ФКН ВШЭ
Семинар №18

1 Оценка параметров многомерного нормального распределения

На лекции обсуждались различные методы восстановления плотности по выборке, в том числе параметрические методы. В этом случае предполагается, что распределение выбирается из некоторого параметрического семейства (например, нормальное распределение или смесь нормальных), после чего по выборке оцениваются значения параметров.

Пусть имеется выборка $X = \{x_i\}_{i=1}^{\ell}$, $x_i \in \mathbb{R}^d$, полученная из многомерного нормального распределения:

$$p(x) = \mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},\ \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d}.$$

Выведем оценки на параметры многомерного нормального распределения по выборке, но сначала напомним некоторые факты из матричного дифференцирования, которые потребуются нам для вывода оценок:

$$\nabla_x \left(a^T x\right) = a,$$
$$\nabla_x \left(x^T A x\right) = (A + A^T)x,$$
$$\nabla_A \left(\det A\right) = (\det A)\, A^{-T},$$
$$\nabla_A \left(x^T A y\right) = xy^T,$$

где $a, x, y \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$.

Задача 1.1. *Выведите оценку максимального правдоподобия на вектор матожиданий μ по выборке X.*

Решение.

Будем максимизировать правдоподобие выборки X:

$$p(X \mid \mu, \Sigma) = \prod_{i=1}^{\ell} \mathcal{N}(x_i \mid \mu, \Sigma) \to \max_{\mu}.$$

Перейдем к логарифму:

$$\log p(X \mid \mu, \Sigma) = -\frac{\ell}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{\ell}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) + \text{const}.$$

Найдем производную по μ и приравняем ее к нулю:

$$\begin{aligned}\nabla_\mu \log p(X \mid \mu, \Sigma) &= -\frac{1}{2}\nabla_\mu \left(\sum_{i=1}^{\ell} x_i^T \Sigma^{-1} x_i - 2\sum_{i=1}^{\ell} x_i^T \Sigma^{-1}\mu + \sum_{i=1}^{\ell} \mu^T \Sigma^{-1}\mu \right) = \\ &= -\frac{1}{2}\left(-2\sum_{i=1}^{\ell} \underbrace{\Sigma^{-T}}_{=\Sigma^{-1}} x_i + \sum_{i=1}^{\ell} 2\Sigma^{-1}\mu \right) = \\ &= \Sigma^{-1}\left(\ell\mu - \sum_{i=1}^{\ell} x_i \right) = \\ &= 0.\end{aligned}$$

Домножая слева на матрицу Σ, получаем

$$\mu = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i.$$

■

Задача 1.2. *Выведите оценку максимального правдоподобия на ковариационную матрицу Σ по выборке X.*

Решение.

Для удобства перейдем в правдоподобии к матрице точности $\Lambda = \Sigma^{-1}$:

$$\log p(X \mid \mu, \Lambda) = -\frac{\ell}{2}\log\det\Lambda^{-1} - \frac{1}{2}\sum_{i=1}^{\ell}(x_i - \mu)^T\Lambda(x_i - \mu) + \text{const}.$$

Найдем производную по Λ и приравняем ее к нулю:

$$\begin{aligned}\nabla_\Lambda \log p(X \mid \mu, \Lambda) &= -\frac{\ell}{2}\nabla_\Lambda \log\det\Lambda^{-1} - \frac{1}{2}\sum_{i=1}^{\ell}\nabla_\Lambda (x_i - \mu)^T\Lambda(x_i - \mu) = \\ &= \frac{\ell}{2}\underbrace{\Lambda^{-T}}_{=\Lambda^{-1}} - \frac{1}{2}\sum_{i=1}^{\ell}(x_i - \mu)(x_i - \mu)^T = 0.\end{aligned}$$

Отсюда

$$\Sigma = \Lambda^{-1} = \frac{1}{\ell}\sum_{i=1}^{\ell}(x_i - \mu)(x_i - \mu)^T.$$

■

2 Байесовские методы классификации

§2.1 Вероятностная постановка задачи классификации

Пусть $\mathbb{X}$ — множество объектов, $\mathbb{Y}$, $|\mathbb{Y}| < \infty$, — множество имён классов, множество $\mathbb{X} \times \mathbb{Y}$ является веротностным пространством с плотностью распределения $p(x, y) = P(y)p(x \,|\, y)$. Вероятности появления объектов каждого из классов $P_y = P(y)$ называются *априорными вероятностями классов*. Плотности распределения $p_y(x) = p(x \,|\, y)$ называются *функциями правдоподобия классов*. Вероятностная постановка задачи классификации разбивается на 2 независимые подзадачи:

1. Имеется простая выборка $X = \{(x_i, y_i)\}_{i=1}^{\ell}$ из неизвестного распределения $p(x, y) = P_y p_y(x)$. Требуется построить эмпирические оценки априорных вероятностей $\hat{P}_y$ и функций правдоподобия $\hat{p}_y(x)$ для каждого из классов $y \in \mathbb{Y}$.

2. По известным плотностям распределения $p_y(x)$ и априорным вероятностям P_y всех классов $y \in \mathbb{Y}$ построить алгоритм $a(x)$, минимизирующий вероятность ошибочной классификации.

Методы восстановления плотности, позволяющие найти $\hat{p}_y(x)$ в первой задаче, обсуждались на лекции, априорная же вероятность может быть оценена как $\hat{P}_y = \frac{\sum_{i=1}^{\ell}[y_i=y]}{\ell}$ согласно закону больших чисел. Заметим, что данная оценка является несмещенной лишь в том случае, если все наблюдаемые объекты заносились в обучающую выборку. На практике применяются и другие способы формирования выборок — в частности, в задачах с несбалансированными классами. В таких случаях оценка $\hat{P}_y$ должна делаться из других содержательных соображений.

Таким образом, мы умеем решать первую из задач. Выведем теперь решение второй задачи.

§2.2 Функционал среднего риска

Для начала запишем, как должен выглядеть оптимизируемый функционал. Напомним, что целью является минимизация вероятности ошибки классификатора. Зафиксируем некоторый классификатор $a(x)$, и пусть он разбивает множество $\mathbb{X}$ на непересекающиеся области $A_y = \{x \in \mathbb{X} \,|\, a(x) = y\}$, $y \in \mathbb{Y}$. Тогда вероятность того, что появится объект класса y и алгоритм $a(\cdot)$ отнесет его к классу s, будет равна $P_y \mathbb{P}(A_s \,|\, y) = P_y \int_{A_s} p_y(x)dx$. Пусть также каждой паре классов $(y, s) \in \mathbb{Y} \times \mathbb{Y}$ поставлена в соответствие некоторая величина потери λ_{ys} при отнесении классификатором объекта класса y к классу s. Обычно полагают $\lambda_{yy} = 0$, $\lambda_{ys} > 0$ при $y \neq s$. Отсюда можем записать ожидаемую величину потери при классификации объектов алгоритмом $a(x)$:

$$R(a) = \sum_{y \in \mathbb{Y}} \sum_{s \in \mathbb{Y}} \lambda_{ys} P_y \mathbb{P}(A_s \,|\, y).$$

Данная величина называется *функционалом среднего риска*.

Значения λ_{ys}, как правило, известны из экспертных оценок и, вообще говоря, могут различаться для различных классов. Однако часто можно полагать $\lambda_{ys} \equiv \lambda_y$, $s \neq y$, т.е. что величина потери зависит лишь от истинной классификации объекта, а не от класса, к которому объект был ошибочно отнесен.

Везде далее будем считать $\lambda_{ys} = [y \neq s]$, то есть что все ошибки классификатора равноценны. В этом случае средний риск $R(a)$ совпадает с вероятностью ошибки алгоритма a:

$$R(a) = \sum_{y\in\mathbb{Y}} \sum_{s\in\mathbb{Y}\setminus\{y\}} P_y \mathbb{P}(A_s \,|\, y).$$

§2.3 Оптимальное байесовское решающее правило

Задача 2.1. *Пусть известны априорные вероятности P_y и функции правдоподобия $p_y(x)$. Покажите, что минимум среднего риска $R(a)$ достигается на алгоритме*

$$a(x) = \arg\max_{y\in\mathbb{Y}} P_y p_y(x).$$

Решение. Для произвольного $t \in \mathbb{Y}$ имеем:

$$R(a) = \sum_{y\in\mathbb{Y}} \sum_{s\in\mathbb{Y}\setminus\{y\}} P_y \mathbb{P}(A_s \,|\, y) = \sum_{y\in\mathbb{Y}} \sum_{s\in\mathbb{Y}} [y \neq s] P_y \mathbb{P}(A_s \,|\, y) = \sum_{y\in\mathbb{Y}} [y \neq t] P_y \mathbb{P}(A_t \,|\, y) + \sum_{y\in\mathbb{Y}} \sum_{s\in\mathbb{Y}\setminus\{t\}} [y \neq s] P_y \mathbb{P}(A_s \,|\, y).$$

По формуле полной вероятности имеем $\mathbb{P}(A_t \,|\, y) = 1 - \sum_{s\in\mathbb{Y}\setminus\{t\}} \mathbb{P}(A_s \,|\, y)$, поэтому:

$$R(a) = \sum_{y\in\mathbb{Y}} [y \neq t] P_y - \sum_{y\in\mathbb{Y}} \sum_{s\in\mathbb{Y}\setminus\{t\}} [y \neq t] P_y \mathbb{P}(A_s \,|\, y) + \sum_{y\in\mathbb{Y}} \sum_{s\in\mathbb{Y}\setminus\{t\}} [y \neq s] P_y \mathbb{P}(A_s \,|\, y).$$

Заметим, что первое слагаемое не зависит от классификатора $a(x)$ и что $[y \neq = t] = 1 - [y = t]$, $[y \neq s] = 1 - [y = s]$. Отсюда имеем:

$$\begin{aligned} R(a) &= \text{const}(a) + \sum_{y\in\mathbb{Y}} \sum_{s\in\mathbb{Y}\setminus\{t\}} ([y \neq s] - [y \neq t]) P_y \mathbb{P}(A_s \,|\, y) = \\ &= \text{const}(a) + \sum_{s\in\mathbb{Y}\setminus\{t\}} \sum_{y\in\mathbb{Y}} ([y = t] - [y = s]) P_y \mathbb{P}(A_s \,|\, y) = \\ &= \text{const}(a) + \sum_{s\in\mathbb{Y}\setminus\{t\}} (P_t \mathbb{P}(A_s \,|\, t) - P_s \mathbb{P}(A_s \,|\, s)) = \\ &= \text{const}(a) + \sum_{s\in\mathbb{Y}\setminus\{t\}} \int_{A_s} (P_t p_t(x) - P_s p_s(x))\, dx. \end{aligned}$$

Обозначим $g_y(x) = P_y p_y(x)$, тогда

$$R(a) = \text{const}(a) + \sum_{s\in\mathbb{Y}\setminus\{t\}} \int_{A_s} (g_t(x) - g_s(x)) dx.$$

Пусть $a^* = \arg\min_a R(a)$ — искомый оптимальный алгоритм, $\{A_s^*\}_s$ — соответствующее ему разбиение пространства объектов $\mathbb{X}$.

Заметим, что $R(a)$ распадается (за исключением аддитивной константы) на $|\mathbb{Y}| - 1$ слагаемых $I(A_s) = \int_{A_s} (g_t(x) - g_s(x)) dx$, каждое из которых зависит только от

одной области A_s. Минимум $I(A_s)$ достигается в том случае, когда A_s совпадает с областью неположительности подынтегрального выражения. Отсюда в силу произвольности t имеем:

$$A_s^* = \{x \in \mathbb{X} \,|\, g_s(x) \geq g_t(x)\, \forall t \in \mathbb{Y} \setminus \{s\}\}\,.$$

С другой стороны, $A_s^* = \{x \in \mathbb{X} \,|\, a^*(x) = s\}$, поэтому $a^*(x) = s$ тогда и только тогда, когда $s = \arg\max_{y\in\mathbb{Y}} g_y(x) = \arg\max_{y\in\mathbb{Y}} P_y p_y(x)$.

■

Полученное выражение называют *байесовским решающим правилом*.

§2.4 Наивный байесовский классификатор

Как было сказано ранее, при применении байесовского классификатора необходимо решить задачу восстановления плотности $p_y(x)$ для каждого класса $y \in \mathbb{Y}$. Данная задача является довольно трудоёмкой и не всегда может быть решена, особенно в случае большого количества признаков, — в частности, если объектами являются тексты, приходится работать с крайне большим числом признаков, и восстановление плотности многомерного распределения не представляется возможным.

Для разрешения этой проблемы сделаем предположение о независимости признаков. В этом случае функция правдоподобия класса y для объекта $x = (x_1, \ldots, x_d)$ может быть представлена в следующем виде:

$$p_y(x) = \prod_{j=1}^{d} p_{yj}(x_j),$$

где $p_{yj}(x_j)$ — одномерная плотность распределения j-ого признака объектов класса $y \in \mathbb{Y}$. В этом случае формула байесовского решающего правила примет следующий вид:

$$a(x) = \arg\max_{y\in\mathbb{Y}} P_y p_y(x) = \arg\max_{y\in\mathbb{Y}} \left(P_y \prod_{j=1}^{d} p_{yj}(x_j) \right) =$$
$$= \arg\max_{y\in\mathbb{Y}} \left(\ln P_y + \sum_{j=1}^{d} \ln p_{yj}(x_j) \right).$$

Предположение о независимости признаков существенно облегчает задачу, поскольку вместо решения задачи восстановления d-мерной плотности необходимо решить d задач восстановления одномерных плотностей. Полученный классификатор называется *наивным байесовским классификатором*.

Плотности отдельных признаков могут быть восстановлены различными способами (параметрическими и непараметрическими). Среди параметрических способов чаще всего используются нормальное распределение (для вещественных признаков), распределение Бернулли и мультиномиальное распределение (для дискретных признаков), благодаря которым получаются различные применяющиеся на практике модели.

2.4.1 Гауссовский наивный байесовский классификатор

Данный классификатор используется для задач с непрерывными признаками. В этом случае в формуле наивного байесовского классификатора в качестве плотности распределения каждого признака используется плотность нормального распределения с оцененными по выборке значениями параметров (матожидание и стандартное отклонение признака) — например, при помощи метода максимума правдоподобия.

2.4.2 Наивный байесовский классификатор Бернулли

Распределение Бернулли используется в случае, если признаки в выборке являются бинарными — например, для задач классификации текстов, где признаками является наличие или отсутствие определенного слова в тексте. В качестве плотности распределения каждого признака используется распределение Бернулли с оцененным по выборке значением параметра. На практике данная модель широко используется для классификации коротких текстов.

2.4.3 Мультиномиальный наивный байесовский классификатор

Мультиномиальное распределение используется, если признаки по смыслу являются счетчиками или частотами. По этой причине данная модель так же, как и в случае распределения Бернулли, используется для классификации текстов. Так же, как и раньше, для конкретного признака используется мультиномиальное распределение с оцененными по выборке значениями параметров.