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:

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

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 машинного обучения

Пусть $X = \{x_1, \dots, x_\ell\}$ — выборка, $\mathbb{X}$ — множество всех возможных объектов, Y — множество ответов. В байесовском подходе предполагается, что обучающие объекты и ответы на них $(x_1, y_1), \dots, (x_\ell, y_\ell)$ независимо выбираются из некоторого распределения $p(x, y)$, заданного на множестве $\mathbb{X} \times Y$. Данное распределение можно переписать как

$$p(x, y) = p(y)p(x \,|\, y),$$

где $p(y)$ определяет вероятности появления каждого из возможных ответов и называется *априорным распределением*, а $p(x \,|\, y)$ задает распределение объектов при фиксированном ответе y и называется *функцией правдоподобия*.

Если известны априорное распределение и функция правдоподобия, то по формуле Байеса можно записать *апостериорное распределение* на множестве ответов:

$$p(y \,|\, x) = \frac{p(x \,|\, y)p(y)}{\int_s p(x \,|\, s)p(s)ds} = \frac{p(x \,|\, y)p(y)}{p(x)},$$

где знаменатель не зависит от y и является нормировочной константой.

§2.1 Оптимальные байесовские правила

Пусть на множестве всех пар ответов $Y \times Y$ задана функция потерь $L(y, s)$. Наиболее распространенным примером для задач классификации является ошибка классификации $L(y, s) = [y \neq s]$, для задач регрессии — квадратичная функция потерь $L(y, x) = (y - s)^2$. *Функционалом среднего риска* называется матожидание функции потерь по всем парам (x, y) при использовании алгоритма $a(x)$:

$$R(a) = \mathbb{E}L(y, a(x)) = \int_Y \int_{\mathbb{X}} L(y, a(x))p(x, y)dxdy.$$

Если распределение $p(x, y)$ известно, то можно найти алгоритм $a_*(x)$, оптимальный с точки зрения функционала среднего риска.

2.1.1 Классификация

Начнем с задачи классификации с множеством ответом $Y = \{1, \dots, K\}$ и функции потерь $L(y, s) = [y \neq s]$. Покажем, что минимум функционала среднего риска достигается на алгоритме

$$a_*(x) = \arg\max_{y \in Y} p(y \,|\, x).$$

Для произвольного классификатора $a(x)$ выполнена следующая цепочка неравенств [?]:

$$\begin{aligned}
R(a) = &\int_Y \int_{\mathbb{X}} L(y, a(x)) p(x,y) dx dy = \\
&= \sum_{y=1}^{K} \int_{\mathbb{X}} [y \neq a(x)] p(x,y) dx = \\
&= \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x,y) dx = \left\{ \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x,y) dx + \int_{\mathbb{X}} p(x, a(x)) dx = 1 \right\} = \\
&= 1 - \int_{\mathbb{X}} p(x, a(x)) dx \geqslant \\
&\geqslant 1 - \int_{\mathbb{X}} \max_{s \in Y} p(x, s) dx = \\
&= 1 - \int_{\mathbb{X}} p(x, a_*(x)) dx = \\
&= R(a_*)
\end{aligned}$$

Таким образом, средний риск любого классификатора $a(x)$ не превосходит средний риск нашего классификатора $a_*(x)$.

Мы получили, что оптимальный байесовский классификатор выбирает тот класс, который имеет наибольшую апостериорную вероятность. Такой классификатор называется *MAP-классификатором* (maximum a posteriori).

2.1.2 Регрессия

Перейдем к задаче регрессии и функции потерь $L(y, x) = (y - s)^2$. Нам пригодится понятие условного матожидания:

$$\mathbb{E}(y \,|\, x) = \int_Y y p(y \,|\, x) dy.$$

Преобразуем функцию потерь [?]:

$$\begin{aligned}
L(y, a(x)) &= (y - a(x))^2 = (y - \mathbb{E}(y \,|\, x) + \mathbb{E}(y \,|\, x) - a(x))^2 = \\
&= (y - \mathbb{E}(y \,|\, x))^2 + 2\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) + (\mathbb{E}(y \,|\, x) - a(x))^2.
\end{aligned}$$

Подставляя ее в функционал среднего риска, получаем:

$$\begin{aligned}
R(a) = &\int_Y \int_{\mathbb{X}} L(y, a(x)) p(x,y) dx dy = \\
&= \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t \,|\, x))^2 p(x,y) dx dy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t \,|\, x) - a(x))^2 p(x,y) dx dy + \\
&+ 2 \int_Y \int_{\mathbb{X}} \big(y - \mathbb{E}(t \,|\, x)\big)\big(\mathbb{E}(t \,|\, x) - a(x)\big) p(x,y) dx dy.
\end{aligned}$$

Разберемся сначала с последним слагаемым. Заметим, что величина $\big(\mathbb{E}(t\,|\,x) - a(x)\big)$ не зависит от y, и поэтому ее можно вынести за интеграл по y:

$$\begin{aligned}
&\int_Y \int_{\mathbb{X}} \big(y - \mathbb{E}(t\,|\,x)\big)\big(\mathbb{E}(t\,|\,x) - a(x)\big)p(x,y)dxdy = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t\,|\,x) - a(x)\big) \int_Y \big\{\big(y - \mathbb{E}(t\,|\,x)\big)p(x,y)\big\}\, dydx = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t\,|\,x) - a(x)\big) \left\{\int_Y yp(x,y)dy - \int_Y \mathbb{E}(t\,|\,x)p(x,y)dy\right\} dx = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t\,|\,x) - a(x)\big) \left\{p(x)\int_Y yp(y\,|\,x)dy - \mathbb{E}(t\,|\,x)\int_Y p(x,y)dy\right\} dx = \\
&= \int_{\mathbb{X}} \big(\mathbb{E}(t\,|\,x) - a(x)\big) \underbrace{\{p(x)\mathbb{E}(t\,|\,x) - p(x)\mathbb{E}(t\,|\,x)\}}_{=0} dx = \\
&= 0
\end{aligned}$$

Получаем, что функционал среднего риска имеет вид

$$R(a) = \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t\,|\,x))^2 p(x,y)dxdy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t\,|\,x) - a(x))^2 p(x,y)dxdy.$$

От алгоритма $a(x)$ зависит только второе слагаемое, и оно достигает своего минимума, если $a(x) = \mathbb{E}(t\,|\,x)$. Таким образом, оптимальная байесовская функция регрессии для квадратичной функции потерь имеет вид

$$a_*(x) = \mathbb{E}(y\,|\,x) = \int_Y yp(y\,|\,x)dy.$$

Иными словами, мы должны провести «взвешенное голосование» по всем возможным ответам, причем вес ответа равен его апостериорной вероятности.

§2.2 Особенности байесовских алгоритмов

Основной проблемой оптимальных байесовских алгоритмов, о которых шла речь в предыдущем разделе, является невозможность их построения на практике, поскольку нам никогда неизвестно распределение $p(x, y)$. Данное распределение можно попробовать восстановить по обучающей выборке, при этом существует два подхода — параметрический и непараметрический. Сейчас мы сосредоточимся на параметрическом подходе.

Допустим, распределение на парах «объект-ответ» зависит от некоторого параметра θ: $p(x, y\,|\,\theta)$. Тогда получаем следующую формулу для апостериорной вероятности:

$$p(y\,|\,x, \theta) \propto p(x\,|\,y, \theta)p(y),$$

где выражение «$a \propto b$» означает «a пропорционально b». Для оценивания параметров применяется *метод максимального правдоподобия*:

$$\theta_* = \arg\max_\theta L(\theta) = \arg\max_\theta \prod_{i=1}^{\ell} p(x_i\,|\,y_i, \theta),$$

где $L(\theta)$ — функция правдоподобия. Примером такого подхода может служить *нормальный дискриминантный анализ*, где предполагается, что функции правдоподобия являются нормальными распределениями:

$$a(x) = \arg\max_{y \in Y} p(y)p(x \mid y),$$
$$p(x \mid y) = \mathcal{N}(x \mid \mu_y, \Sigma_y).$$

Параметрами алгоритма являются средние μ_y и ковариационные матрицы классов Σ_y, которые оцениваются по выборке методом максимального правдоподобия.

Если предположить, что ковариационные матрицы классов равны, и оценивать их по всей выборке, то мы получим алгоритм, называемый *линейным дискриминантом Фишера*. Можно показать, что он является линейным:

$$a(x) = \arg\max_{y \in Y}(\langle w_y, x\rangle + w_{0y}),$$

причем $w_y = \Sigma^{-1}\mu_y$. В случае двух классов ($Y = \{-1, +1\}$) классификатор принимает вид

$$a(x) = \text{sign}\left(\langle w, x\rangle + b\right) \quad w = \Sigma^{-1}(\mu_2 - \mu_1). \tag{2.1}$$

Заметим, что мы ранее уже вводили данный алгоритм, но с другой точки зрения — ранее пытались минимизировать внутриклассовую и максимизировать межклассовую дисперсию при помощи линейного классификатора, и полученное нами тогда решение полностью совпадает с приведенным выше.

§2.3 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 $x = (x_1, \dots, x_d)$ может быть представлена в следующем виде:

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

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

$$a(x) = \arg\max_{y \in Y} p(y \mid x) = \arg\max_{y \in \mathbb{Y}} \left(\ln p(y) + \sum_{j=1}^{d} \ln p(x_j \mid y)\right).$$

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.