Машинное обучение, ФКН ВШЭ
Семинар №21

1 Определение Гауссовского процесса

Напомним, что случайным процессом называется семество $\{\xi_t\}_{t \in T}$ случайных велечин ξ_t, индексированных некоторым параметром t.

Случайный процесс называется слабостационарным, если $\mathbb{E}\xi_t = \text{const}$, $\text{cov}(\xi_{t'}, \xi_{t''}) = K(t', t'') = K(t' - t'')$.

Гауссовский процесс $\{\xi_t\}_{t \in T}$ — это такой случайный процесс, у которого вектор, составленный из произвольного конечного подмножества случайных величин $(\xi_{t_1}, \dots, \xi_{t_k})$, имеет многомерное нормальное распределение:

$$(\xi_{t_1}, \dots, \xi_{t_k}) \sim \mathcal{N}(\mu, \Sigma), \quad \mu_j = \mathbb{E}\xi_{t_j}, \quad \Sigma_{ij} = \text{cov}(\xi_{t_i}, \xi_{t_j}) = K(t_i, t_j).$$

2 Построение регрессии с помощью гауссовского процесса

Изучим, как понятие гауссовских процессов может быть применено в задаче восстановления регрессии.

Будем считать, что индексирующим множеством T является пространство объектов $\mathbb{X}$, а значения целевой переменной для объектов некоторой выборки $X = \{(x_i, t_i)\}_{i=1}^{\ell}$ состоят из двух слагаемых:

$$t_i = y_i + \varepsilon_i,$$

где y_i — неизвестное истинное значение целевой зависимости на объекте x_i, а $\varepsilon_i \sim \mathcal{N}(0, \beta^{-1})$ — шум, полученный в результате измерения. Будем предполагать, что распределение вектора истинных значений целевых зависимостей на объектах выборки y нормально и

$$p(y) = \mathcal{N}(y \,|\, 0, K),$$

где $K = (K(x_i, x_j))_{i,j=1}^{\ell}$ — ковариационная матрица выборки для некоторой заранее заданной ковариационной функции. Тогда для распределения вектора значений целевой переменной для объектов выборки t верно следующее:

$$p(t \,|\, y) = \mathcal{N}(t \,|\, y, \beta^{-1} I)$$

Для решения задачи нам необходимо научиться предсказывать значение целевой переменной для любого нового объекта — для этого приведём решения нескольких сопутствующих задач в более общих постановках.

Задача 2.1. *Докажите следующее тождество для матрицы, заданной в блочном виде:*

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix},$$

где $M = (A - BD^{-1}C)^{-1}$.

Решение.

Проверяется непосредственно переменожением матриц.

■

Задача 2.2. *Для заданных плотностей*

$$\begin{aligned} p(x) &= \mathcal{N}(x \mid \mu, \Lambda^{-1}), \\ p(y \mid x) &= \mathcal{N}(y \mid Ax + b, L^{-1}), \end{aligned}$$

необходимо найти плотность $p(y)$.

Решение. Рассмотрим вектор

$$z = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Из свойств многомерного нормального распределения известно, что вектор z распределён нормально — найдём параметры данного распределения. Распишем $\log p(z)$:

$$\log p(z) = -\frac{1}{2}(z - \mu_z)^T \Sigma_z^{-1}(z - \mu_z) + \mathrm{const} = -\frac{1}{2} z^T \Sigma_z^{-1} z + z^T \Sigma_z^{-1} \mu_z + \mathrm{const}$$

Теперь запишем то же выражение с использованием формулы полной вероятности:

$$\begin{aligned} \log p(z) &= \log p(x) + \log p(y \mid x) \\ &= -\frac{1}{2}(x - \mu)^T \Lambda (x - \mu) \\ &\quad -\frac{1}{2}(y - Ax - b)^T L (y - Ax - b) + \mathrm{const} \end{aligned}$$

Рассмотрим все члены второго порядка относительно x, y в обоих представлениях $\log p(z)$. Заметим, что их можно записать в виде

$$\begin{aligned} &-\frac{1}{2} x^T (\Lambda + A^T L A) x - \frac{1}{2} y^T L y + \frac{1}{2} y^T L A x + \frac{1}{2} x^T A^T L y \\ &= -\frac{1}{2} \begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -\frac{1}{2} z^T \Sigma_z^{-1} z \end{aligned}$$

Отсюда получили, что

$$\Sigma_z^{-1} = \begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix}$$

Используя тождество из предыдущей задачи, получаем:

$$\Sigma_z = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1} A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}$$

Аналогично для членов первого порядка относительно x, y имеем:

$$x^T \Lambda \mu - x^T A^T L b + y^T L b = \begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} \Lambda\mu - A^T L b \\ Lb \end{pmatrix}$$

Получаем

$$z^T \begin{pmatrix} \Lambda\mu - A^T L b \\ Lb \end{pmatrix} = z^T \Sigma_z^{-1} \mu_z$$

Отсюда

$$\mu_z = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1} A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix} \begin{pmatrix} \Lambda\mu - A^T L b \\ Lb \end{pmatrix} = \begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}$$

Итого, получаем плотность $p(y)$:

$$p(y) = \mathcal{N}(y \,|\, A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

■

Таким образом, теперь может быть выражена плотность $p(t)$, позволяющая описывать произвольное количество объектов из заданного пространства объектов:

$$p(t) = \int p(t \,|\, y) p(y) dy = \int \mathcal{N}(t \,|\, y, \beta^{-1} I) \mathcal{N}(y \,|\, 0, K) dy = \mathcal{N}(t \,|\, 0, K + \beta^{-1} I)$$

Задача 2.3. *Известно, что состоящий из 2 частей вектор x*

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix},$$

имеет многомерное нормальное распределение с параметрами

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}.$$

Найти плотность $p(x_a \,|\, x_b)$.

Решение.

Аналогично решению предыдущей задачи запишем $\log p(x_a \,|\, x_b)$:

$$\log p(x_a \,|\, x_b) = -\frac{1}{2}(x_a - \mu_{a\,|\,b})^T \Sigma_{a\,|\,b}^{-1} (x_a - \mu_{a\,|\,b}) + \text{const} = -\frac{1}{2} x_a^T \Sigma_{a\,|\,b}^{-1} x_a + x_a^T \Sigma_{a\,|\,b}^{-1} \mu_{a\,|\,b} + \text{const}$$

$$\log p(x_a \,|\, x_b) = \log p(x) - \log p(x_b)$$

Пусть $\Lambda = \Sigma^{-1}$ и может быть представлена в блочном виде аналогично Σ. Тогда для $\log p(x)$ верно:

$$\begin{aligned}\log p(x) = \log p(x_a, x_b) = -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + \text{const} = \\ -\frac{1}{2}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b) \\ -\frac{1}{2}(x_b-\mu_b)^T\Lambda_{ba}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T\Lambda_{bb}(x_b-\mu_b) + \text{const}.\end{aligned}$$

Отсюда получаем:

$$\begin{aligned}\mu_{a\,|\,b} =& \Sigma_{a\,|\,b}(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)) \\ \Sigma_{a\,|\,b} =& \Lambda_{aa}^{-1}\end{aligned}$$

Используя тождество из задачи 2.1, получим, что

$$\begin{pmatrix}\Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb}\end{pmatrix}^{-1} = \begin{pmatrix}\Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb}\end{pmatrix}$$

$$\begin{aligned}\Lambda_{aa} &= (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}, \\ \Lambda_{ab} &= -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}.\end{aligned}$$

Отсюда получаем ответ:

$$\begin{aligned}p(x_a \,|\, x_b) &= \mathcal{N}(x_a \,|\, \mu_{a\,|\,b}, \Sigma_{a\,|\,b}), \\ \mu_{a\,|\,b} &= \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \\ \Sigma_{a\,|\,b} &= \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}.\end{aligned}$$

■

Положим в условии предыдущей задачи $x = (t_{\ell+1}, t_1, \ldots, t_\ell)^T, x_a = (t_{\ell+1})^T$, $x_b = (t_1, \ldots, t_\ell)^T$. Отметим также, что ранее нами было получено распределение $p(t_1, \ldots, t_\ell, t_{\ell+1})$. откуда можем получить

$$p(t_{\ell+1} \,|\, t_1, \ldots, t_\ell) = \mathcal{N}(k^T C_\ell^{-1} t, K(x_{\ell+1}, x_{\ell+1}) + \beta^{-1} - k^T C_\ell^{-1} k),$$

где $C_\ell = K + \beta^{-1} I$, $k = (K(x_1, x_{\ell+1}), \ldots, K(x_\ell, x_{\ell+1}))$.

3 Обучение ковариационной функции

Отметим, что в приведенном методе результат регрессии сильно зависит от выбора ковариационной функции $K(x, y)$, описывающей сходство объектов x, y. В качестве ковариационных функций можно, в частности, рассматривать одну из следующих:

- $K(x, y) = C$ – константная,
- $K(x, y) = \langle x, y\rangle$ – линейная,
- $K(x, y) = \exp(-\frac{(x-y)^2}{2l^2})$ – экспоненциальная,
- $K(x, y) = \exp(|x - y|/l)$ – процесс Орнштейна-Уленбека.

Распространненой практикой является обучение параметров ковариационной функции после выбора её общего вида. Рассмотрим функцию

$$K(x, y) = \exp\left(-\sum_{j=1}^{d} \theta_j (x_j - y_j)^2\right)$$

Здесь θ_j, $j = \overline{1, d}$ являются обучаемыми параметрами и могут быть найдены путем максимизации логарифма правдоподобия:

$$\log p(t \mid \Theta) \propto -\frac{1}{2}\log|C_\ell| - \frac{1}{2}t^T C_\ell^{-1} t \to \max_{\Theta}$$

$$\frac{\partial}{\partial\theta_j}\log p(t \mid \Theta) = -\frac{1}{2}\mathrm{Tr}\left(C_\ell^{-1}\frac{\partial C_\ell}{\partial\theta_j}\right) + \frac{1}{2}t^T C_\ell^{-1}\frac{\partial C_\ell}{\partial\theta_j}C_\ell^{-1} t$$