Математические основы искусственного интеллекта
Уравнение регрессии

Солодушкин Святослав Игоревич

Кафедра вычислительной математики и компьютерных наук,
УрФУ имени первого Президента России Б.Н. Ельцина

Ноябрь 2021

Определение

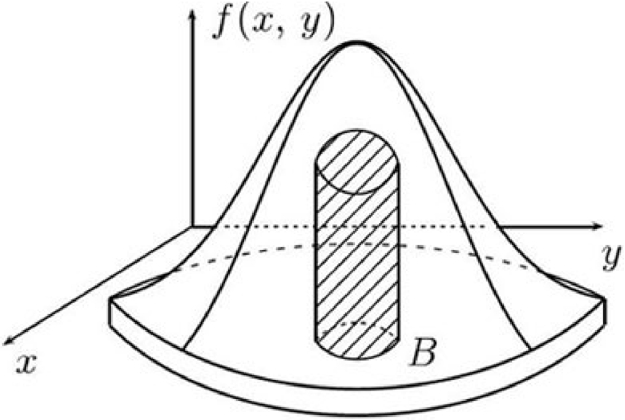
Говорят, что двумерная случайная величина ξ, η имеет двумерное нормальное (гауссовское) распределение с параметрами, если ξ, η имеет следующую плотность распределения:

$$f_{\xi\eta}(x, y) = \frac{1}{2\pi\sigma_\xi\sigma_\eta\sqrt{1 - r_{\xi\eta}^2}} \times$$

$$\times e^{-\frac{1}{2(1-r_{\xi\eta}^2)}\left[\frac{(x-m_\xi)^2}{\sigma_\xi^2} - 2r_{\xi\eta}\frac{x-m_\xi}{\sigma_\xi}\frac{y-m_\eta}{\sigma_\eta} + \frac{(y-m_\eta)^2}{\sigma_\eta^2}\right]}, \quad (x, y) \in \mathbb{R}^2.$$

Можно доказать, что m_ξ, m_η — математические ожидания, $\sigma_\xi^2, \sigma_\eta^2$ — дисперсии, $r_{\xi\eta}$ — коэффициент корреляции случайных величин ξ, η.

Плотность распределения



Некоррелированность влечет независимость

Если составляющие двумерной случайной величины некоррелированы, то они и независимы.

$$f_{\xi\eta}(x,y) = \frac{1}{2\pi\sigma_\xi\sigma_\eta} \times e^{-\frac{1}{2}\left[\frac{(x-m_\xi)^2}{\sigma_\xi^2}+\frac{(y-m_\eta)^2}{\sigma_\eta^2}\right]} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_\xi} e^{-\frac{(x-m_\xi)^2}{2\sigma_\xi^2}} \times \frac{1}{\sqrt{2\pi}\sigma_\eta} e^{-\frac{(y-m_\eta)^2}{2\sigma_\eta^2}} = f_\xi(x)\, f_\eta(y).$$

Для нормально распределенных составляющих двумерной случайной величины понятие некоррелированности и независимости равносильны.

Теорема

Если двумерная случайная величина ξ, η распределена нормально, то ξ, η связаны линейной связью.

Доказательство. Сделаем замену $u = \dfrac{x - m_\xi}{\sigma_\xi}, \quad v = \dfrac{y - m_\eta}{\sigma_\eta}$. Получим

$$f_{\xi\eta}(u, v) = \frac{1}{2\pi\sigma_\xi\sigma_\eta\sqrt{1 - r^2}} \times e^{-\frac{1}{2(1-r^2)}\left[u^2 - 2ruv + v^2\right]}.$$

Найдем маргинальную функцию распределения

$$f_\xi(u) = \int\limits_{-\infty}^{+\infty} f_{\xi\eta}(u, v)\, dv = \frac{1}{\sqrt{2\pi}\sigma_\xi} e^{-\frac{u^2}{2}}.$$

Найдем условную плотность распределения

$$f_\eta(y|\xi = x) = \frac{1}{\sqrt{2\pi}\sigma_\eta\sqrt{1-r^2}} e^{-\frac{(v-ru)^2}{2\sqrt{1-r^2}}} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_\eta\sqrt{1-r^2}} e^{-\frac{1}{2\sigma_\eta^2(1-r^2)}\left(y-\left[m_\eta - r\frac{\sigma_\eta}{\sigma_\xi}(x-m_\xi)\right]\right)^2}.$$

Условное математическое ожидание

$$\mathrm{E}(\eta|\xi = x) = m_\eta - r\frac{\sigma_\eta}{\sigma_\xi}(x - m_\xi).$$

Критерий линейнойсти регрессии

так что

$$\mu_{2v} = \mu'_{2v} - (\mu'_{1v})^2 = (1-\rho^2)\left\{2\rho^2 v + \frac{1}{2}n(1-\rho^2)^2\right\}. \qquad (28.23)$$

Соотношения (28.22) и (28.23) показывают, что регрессии среднего и дисперсии величины u по v линейны.

Критерии линейности регрессии

28.5 Пусть $\psi(t_1, t_2) = \log\varphi(t_1, t_2)$ — совместная п. ф. с. величин x и y. Сейчас мы докажем следующий факт: если регрессия величины y по x линейна, так что

$$\mu'_{1x} = \mathsf{M}(y \mid x) = \beta_0 + \beta_1 x, \qquad (28.24)$$

то

$$\left[\frac{\partial\psi(t_1, t_2)}{\partial t_2}\right]_{t_2=0} = i\beta_0 + \beta_1\frac{\partial\psi(t_1, 0)}{\partial t_1}; \qquad (28.25)$$

и наоборот, если выполнено некоторое условие полноты, то (28.25) не только необходимо, но и достаточно для (28.24).

Используя (28.24), из (28.9) при $r = 1$ находим

$$\left[\frac{\partial\varphi(t_1, t_2)}{\partial t_2}\right]_{t_2=0} = i\int_{-\infty}^{\infty}\exp(it_1x)\, g(x)(\beta_0 + \beta_1 x)\, dx = \qquad (28.26)$$

$$= i\beta_0\varphi(t_1, 0) + \beta_1\frac{\partial}{\partial t_1}\varphi(t_1, 0). \qquad (28.27)$$

Полагая в (28.27) $\psi = \log\varphi$ и деля обе части на $\varphi(t_1, 0)$, получаем (28.25).

Обратно, если имеет место соотношение (28.25), то, используя (28.9), перепишем его в виде

$$i\int_{-\infty}^{\infty}\exp(it_1x)(\beta_0 + \beta_1 x - \mu'_{1x})\, g(x)\, dx = 0. \qquad (28.28)$$

Теперь видим, что соотношение (28.28) влечет тождественно по x

$$\beta_0 + \beta_1 x - \mu'_{1x} = 0, \qquad (28.29)$$

если только семейство $\exp(it_1x)g(x)$ полно. Следовательно, мы получили (28.24).

Прямая среднеквадратичной регрессии

Для упрощения математической стороны изложения, мы наложим ограничения[1] на случайные величины ξ и η. Пусть условное распределение η относительно своего среднего (которое, как и раньше, является функцией от x) одно и то же для любого x, т. е. только среднее значение η изменяется при изменении x. Говорят, что η имеет «однородные ошибки».

Таким образом, существует случайная величина ε такая, что

$$\eta = M[\eta|\xi = x] + \varepsilon.$$

В частности, когда регрессия линейная, имеем

$$\eta = \beta_1\xi + \beta_0 + \varepsilon.$$

[1]Эти ограничения не сильно обременительны и в прикладных задачах обычно выполняются хотя бы приближенно.

Прямая среднеквадратичной регрессии

Априори известен только вид уравнения, а конкрентные значения коэффициентов β_0 и β_1 неизвестны, естественно выбрать их так, чтобы, зная значение, которое в эксперименте приняла величина ξ, наиболее точно спрогнозировать знечение, которое примет величина η.
Таким образом мы приходим к постановке задачи построения линейной регресси:

$$M[\eta - \beta_1 \xi - \beta_0]^2 \xrightarrow[\beta_0,\, \beta_1]{} \min .$$

Прямая среднеквадратичной регрессии

Теорема

Линейная среднеквадратическая регрессия η на ξ имеет вид

$$f(x) = m_\eta + r\frac{\sigma_\eta}{\sigma_\xi}(x - m_\xi),$$

где m_η, m_ξ, σ_η, σ_ξ — математические ожидания и средние квадратические отклонения случайных величин η и ξ, соответственно, r — коэффициент корреляции случайных величин η и ξ.

Прямая среднеквадратичной регрессии

Рассмотрим функцию $F(\beta_0, \beta_1) = M[\eta - \beta_1 \xi - \beta_0]^2$.
Пользуясь формулами

$$M[\xi^2] = M^2[\xi] + D[\xi]$$

$$M[\xi\eta] = M[\xi]M[\eta] + cov(\xi, \eta) = M[\xi]M[\eta] + r\sigma_\xi\sigma_\eta,$$

получим

$$F(\beta_0, \beta_1) = \sigma_\eta^2 + \beta_1^2\sigma_\xi^2 - 2r\sigma_\xi\sigma_\eta\beta_1 + (m_\eta - \beta_1 m_\xi - \beta_0)^2.$$

$$F(\beta_0, \beta_1) = \sigma_\eta^2 + \beta_1^2\sigma_\xi^2 - 2r\sigma_\xi\sigma_\eta\beta_1 + (m_\eta - \beta_1 m_\xi - \beta_0)^2.$$

Исследуем функцию F на минимум, приравняв частные производные к нулю

$$\frac{\partial F}{\partial \beta_0} = -2(m_\eta - \beta_1 m_\xi - \beta_0) = 0,$$

$$\frac{\partial F}{\partial \beta_1} = 2\beta_1\sigma_\xi^2 - 2r\sigma_\xi\sigma_\eta = 0.$$

Откуда получаем

$$\beta_0 = m_\eta - r\frac{\sigma_\eta}{\sigma_\xi}m_\xi, \qquad \beta_1 = r\frac{\sigma_\eta}{\sigma_\xi}.$$

Интеграл Пуассона

$$I = \int\limits_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}.$$

$$I = \int\limits_{-\infty}^{\infty} e^{-ax^2}\, dx = \frac{1}{\sqrt{a}}\,\Gamma\left(\frac{1}{2}\right).$$

Для сведения $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Задачи

Двумерная случайная величина задана плотностью распределения

$$f_{\xi\eta}(x, y) = \frac{3\sqrt{3}}{\pi} e^{-4x^2-6xy-9y^2}.$$

Найти условные законы распределения составляющих. Найти уравнение регрессии η на ξ.