Прикладная статистика и анализ данных

Съезд II

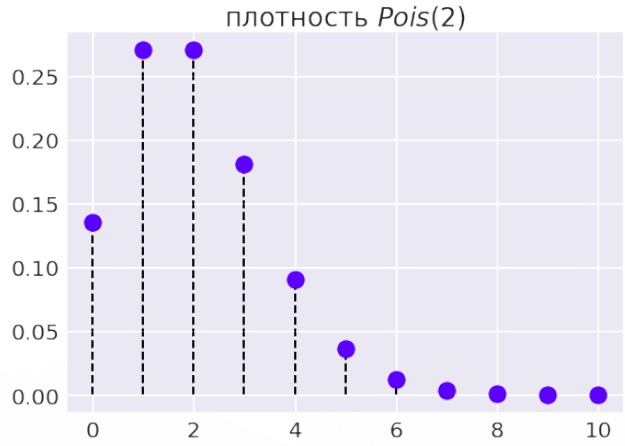
Обобщенная модель линейной регрессии

Пуассоновское распределение

$Pois(\lambda): \quad p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \; x \in \mathbb{Z}_+$

Смысл: число событий,

произошедших за единицу времени



Условия:

1. события происходят с фиксированной интенсивностью λ.
2. независимо друг от друга.

Утверждение: время между двумя событиями имеет распр. $Exp(\lambda)$ (*см. пуассоновские случайные процессы*)

Примеры:

1. число клиентов в час
2. число запросов на сервер за минуту

Интенсивность не постоянна, может зависеть от каких-то факторов.

Обобщенная модель линейной регрессии

Гауссовская линейная модель

Ожидаемый отклик:

$y = \mu_\theta(x) = x^T\theta.$

Наблюдаемый отклик:

$Y_i = x_i^T\theta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$

или $Y_i \sim \mathcal{N}(\mu_\theta(x_i), \sigma^2)$

Оценка отклика:

$\widehat{y} = x^T\widehat{\theta}.$

Generalized Linear Models (GLM)

Ожидаемый отклик:

$y = \mu_\theta(x)$, причем $g(\mu_\theta(x)) = x^T\theta$,

т.е. g — линеаризация ожид. отклика

Наблюдаемый отклик:

$Y_i \sim \mathsf{P}_{\mu_\theta(x_i)}$,

где $\{\mathsf{P}_\psi \mid \psi \in \Psi\}$ — семейство распр.

Оценка отклика:

$\widehat{y} = g^{-1}\left(x^T\widehat{\theta}\right).$

Натуральный отклик

$y \in \mathbb{Z}_+$ — значения наблюдаемого отклика

$\mu_\theta(x) = \mathsf{E}_x Y$ — ожидаемый отклик

$Y_i \sim Pois(\mu_\theta(x_i)) = \mathsf{P}_{\mu_\theta(x_i)}$ — наблюдаемый отклик

Линеаризация ожид. отклика: $g(z) = \ln z$

т.е. $g(\mu_\theta(x)) = \ln \mathsf{E}_x Y = x^T\theta$

Тогда $\mu_\theta(x) = g^{-1}(x^T\theta) = exp(x^T\theta)$

$\Longrightarrow$ это **пуассоновская регрессия**

Смысл: пытаемся приблизить интенсивность с помощью регрессии.

Примечание. Интенсивность имеет неравномерный масштаб. Пусть $\lambda = 1$.
В два раза чаще это $\lambda = 2$, в два раза реже это $\lambda = 1/2$.
Логарифмирование это исправляет.

Свойства GLM

В качестве $\widehat{\theta}$ берется ОМП (ищется численно)

$$L_X(\theta) = \prod_{i=1}^{n} p_{\mu_\theta(x_i)}(Y_i) = \prod_{i=1}^{n} p_{g^{-1}(x_i^T\theta)}(Y_i) \longrightarrow \max_\theta$$

Если $\{\mathrm{P}_\psi \mid \psi \in \Psi\}$ лежит в экспоненциальном классе, то $\widehat{\theta}$:

1. существует и единственна;
2. состоятельна;
3. асимптотически нормальна: $\sqrt{I(\theta)}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}(0, I_d)$,
   где $I(\theta) = \left(-\mathrm{E}\frac{\partial^2 \log L_X(\theta)}{\partial\theta_j \partial\theta_k}\right)_{jk}$ — информационная матрица Фишера.

Для пуассоновской регрессии $I(\theta) = \sum_{i=1}^{n} e^{x_i^T\theta} x_i x_i^T = X^T V(\theta) X$,
где $V(\theta) = \mathrm{diag}\left[e^{x_i^T\theta}\right]$.

Асимпт. доверительные интервалы в GLM

Для параметров ($\Longrightarrow$ критерий для гипотезы $\mathsf{H}_0 \colon \theta_j = 0$)

$$\theta_j \quad \in \quad \left(\widehat{\theta}_j \pm z_{1-\alpha/2} \sqrt{\left(I^{-1}(\widehat{\theta}) \right)_{jj}} \right)$$

Для преобразованного ожидаемого отклика

$$x_0^T \theta \quad \in \quad \left(x_0^T \widehat{\theta} - \delta, x_0^T \widehat{\theta} + \delta \right)$$

Для ожидаемого отклика

$$\mu(x_0) = g^{-1}(x_0^T \theta) \quad \in \quad \left[g^{-1} \left(x_0^T \widehat{\theta} - \delta \right), g^{-1} \left(x_0^T \widehat{\theta} + \delta \right) \right],$$

$$\delta = z_{1-\alpha/2} \sqrt{x_0^T I^{-1}(\widehat{\theta}) x_0}$$