Спецкурс "Регрессия"

Шкляев А.В.

15 января 2019 г.

1 Общая линейная модель

1.1 Экспоненциальное семейство

Распределение, имеющее плотность или дискретное распределение вида

$$f(x;\theta) = d(\theta)b(x)\exp(x\theta)$$

называют распределением, принадлежащим натуральному однопараметрическому экспоненциальному семейству. К таким распределениям относятся, например, бернуллиевское, геометрическое, пуассоновское, гамма (с известным параметром формы), нормальное (с известной дисперсией). Параметр θ при этом называется натуральным параметром, $\sum_{i=1}^{n} x_i$ является достаточной статистикой.

Пример 1. Для нормального распределения $\mathcal{N}(\mu, 1)$

$$f(x;\mu) = \exp\left(-\frac{\mu^2}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\cdot\exp\left(x\mu\right).$$

Тогда натуральным параметром будет $\theta = \mu$.

Пример 2. Для $Bernoulli(p)$

$$f(x;p) = p^x(1-p)^{1-x} = \exp(x\ln p + (1-x)\ln(1-p)) = (1-p)\exp\left(x\ln\frac{p}{1-p}\right).$$

Таким образом, натуральным параметром будет

$$\theta = \ln\frac{p}{1-p}.$$

Пример 3. Для $Poiss(\lambda)$

$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\frac{1}{x!}\exp(x\ln\lambda),$$

натуральный параметр при этом $\theta = \ln\lambda$.

При этом

$$0 = \frac{\partial}{\partial\theta}\left(\int_{\mathbb{R}} f(x;\theta)dx\right) = \int_{\mathbb{R}} xf(x;\theta)dx - \int_{\mathbb{R}} m(\theta)f(x;\theta)dx = \mathbf{E}X - m(\theta),$$

$$0 = \frac{\partial^2}{\partial\theta^2}\left(\int_{\mathbb{R}} f(x;\theta)dx\right) = \int_{\mathbb{R}} x^2f(x;\theta)dx - m(\theta)\int_{\mathbb{R}} xf(x;\theta)dx - m'(\theta) = \mathbf{D}X - m'(\theta),$$

откуда $\mathbf{E}X = m(\theta)$, $\mathbf{D}X = m'(\theta)$.

1.2 Общая линейная модель

Мы будем решать задачу регрессии в общей линейной модели. Итак, пусть $x_{i,1}, ..., x_{i,k}$ — предикторы, y_i — зависимая переменная. Как мы уже обсуждали, в общей постановке мы хотим построить регрессию

$$\mathbf{E}(Y|X_\cdot = \vec{x}) = h(\vec{x}).$$

Мы будем считать, что h — функция от $\langle \vec{x}, \vec{a} \rangle$, где $\vec{a}$ — вектор параметров. Более того, будем считать, что h обратима, $h^{-1} = g$. Для задания модели мы должны знать три компоненты:

1. Случайная компонента $f(y; \theta)$: условное распределение y при заданном θ является заданным распределением, принадлежащему натуральному однопараметрическому семейству;
2. Систематическая компонента X: параметр $\eta = X\vec{a}$, где X — матрица предикторов, $\vec{a}$ — вектор параметров;
3. Функция связи $\eta = g(\mu)$, где $\mu = \mu(\theta)$ — математическое ожидание Y при заданном θ.

Канонической функцией связи будем называть $g(\mu) = \theta$.

Пример 4. Для нормальной модели $\mathcal{N}(\mu, 1)$ каноническая функция связи предполагает рассмотрение соотношения

$$\mu = \langle x, \vec{a} \rangle,$$

откуда $\mu = \langle x, \vec{b} \rangle$ для некоторого вектора параметров $\vec{b}$. Тем самым, мы предполагаем, что

$$\vec{Y} \sim \mathcal{N}(X\vec{b}, 1),$$

то есть используем обычную нормальную регрессию

Пример 5. Для бернуллиевской модели каноническая функция связи дает нам соотношение

$$\ln \frac{p}{1-p} = \langle x, \vec{a} \rangle,$$

откуда

$$p = \frac{\exp(\langle x, \vec{a} \rangle)}{1 + \exp(\langle x, \vec{a} \rangle)}.$$

Тем самым каноническая регрессия соответствует соотношению

$$\mathbf{E}(Y|\vec{X}) = \frac{\exp(\langle x, \vec{a} \rangle)}{1 + \exp(\langle x, \vec{a} \rangle)}$$

Пример 6. Для пуассоновской модели каноническая функция связи дает нам соотношение

$$\ln \lambda = \langle x, \vec{a} \rangle,$$

откуда

$$\lambda = \exp(\langle x, \vec{a} \rangle).$$

1.3 Оценка максимального правдоподобия

В рамках общей линейной модели мы получаем логарифм правдоподобия

$$\ln L(y_1, ..., y_n; \theta_1, ..., \theta_n) = \ln d(\theta_1) + ... + \ln d(\theta_n) + \sum_{i=1}^{n} \ln b(y_i) + y_1\theta_1 + ... + y_n\theta_n.$$

В нашем случае $\theta_i = \langle X_{i,\cdot}, \vec{a} \rangle$. Дифференцируя, имеем

$$\frac{\partial \ln L}{\partial a_j}(y_1, ..., y_n; X; \vec{a}) = \sum_{i=1}^{n} y_i X_{i,j} + \sum_{i=1}^{n} (\ln d)'(\theta_i) X_{i,j}.$$

Таким образом, критическая точка функции L определяется соотношением

$$h(\theta) = X^t(\vec{Y} + p(\theta)) = 0,$$

где $p(\theta) = ((\ln d)'(\theta_1), ..., (\ln d)'(\theta_n))$. Вне нормального случая это уравнение затруднительно решить в явном виде, поэтому предлагается использовать формулу Ньютона.

Рассмотрим некоторое $\vec{a}_0$, а в каждый момент будем определять $\vec{a}_{n+1}$ из соотношения

$$h(\vec{a}_0) + h'(\vec{a}_0)(\vec{a}_1 - \vec{a}_0) = 0,$$

заменяющего функцию h на ее линеаризацию в точке $\vec{a}_0$, где $h'(\vec{a}_0)$ — матрица $\partial h'(\vec{a}_0)_i / \partial a_j$.

Наша матрица производных есть

$$\frac{\partial^2 \ln L}{\partial a_i \partial a_j}(\vec{a}_n) = -X^t W X,$$

где W — диагональная матрица с

$$W_{i,i} = \mathbf{D}Y_i(\vec{a}_n) = -(\ln d)''(\langle X_{i,\cdot}, \vec{a}_n \rangle).$$

Таким образом, мы можем итеративно рассматривать

$$\vec{a}_{n+1} = \vec{a}_n - (X^t W X)^{-1} X^t (\vec{Y} + p(\theta)) = (X^t W X)^{-1} X^t W \vec{z},$$

где $\vec{z} = X\vec{a}_n - W^{-1}(\vec{Y} + p(\theta))$. Можно заметить, что a_{n+1} совпадает с решением задачи взвешенной регрессии

$$\min ||\vec{z} - X\vec{a}_{n+1}||_W$$

На каждом шаге мы хотим регрессировать $\vec{z}$ на пространство L относительно матрицы скалярного произведения W, заданной дисперсиями модели.

Указанный алгоритм включает в себя канонические регрессии для различных представителей натурального однопараметрического экспоненциального семейства.