

Обобщённые линейные модели

GLM: от теории к алгоритмам

- 1 Введение
- 2 Экспоненциальное семейство
- 3 Структура GLM
- 4 IRLS: вывод алгоритма
- 5 Частные случаи
- 6 GLM как локальная регрессия
- 7 Одношаговые оценки
- 8 Заключение

Что такое GLM?

Обобщённые линейные модели (GLM)

- Гибкое расширение классической линейной регрессии
- Позволяют моделировать зависимые переменные из **экспоненциального семейства**
- Объединяют: линейную регрессию, логистическую регрессию, пуассоновскую регрессию и др.

Нелдер и Уэддерберн (1972) — систематическое изложение GLM

Экспоненциальное семейство (частный случай)

Определение

Плотность (или функция вероятности) имеет вид:

$$f(y | \theta) = h(y) \exp\{\theta y - A(\theta)\}$$

где:

- θ — канонический параметр (скаляр)
- $A(\theta) = \ln \int h(y) e^{\theta y} dy$ — нормирующая функция
- $h(y) \geq 0$ — базовая мера (не зависит от θ)

Theorem (Свойства экспоненциального семейства)

Для распределения $f(y | \theta) = h(y) \exp\{\theta y - A(\theta)\}$:

$$\mathbb{E}[Y] = A'(\theta)$$

$$\text{Var}(Y) = A''(\theta)$$

Обозначим $\mu = \mathbb{E}[Y]$. Тогда:

$$\mu = A'(\theta), \quad \text{Var}(Y) = A''(\theta) = V(\mu)$$

где $V(\mu)$ — функция дисперсии.

Три компонента GLM

- ❶ Случайная компонента: Y_i из экспоненциального семейства

$$f(y_i | \theta_i) = h(y_i) \exp\{\theta_i y_i - A(\theta_i)\}$$

- ❷ Систематическая компонента: линейный предиктор

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

- ❸ Функция связи: $g(\mu_i) = \eta_i$

Каноническая функция связи

Выбираем g так, чтобы $\theta_i = \eta_i$, то есть:

$$g(\mu_i) = \theta_i = (A')^{-1}(\mu_i)$$

Следствие: $g'(\mu) = \frac{1}{V(\mu)}$

Логарифмическое правдоподобие

Для одного наблюдения:

$$\ell_i(\boldsymbol{\beta}) = \theta_i y_i - A(\theta_i) + \ln h(y_i)$$

Поскольку $\theta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$:

$$\ell_i(\boldsymbol{\beta}) = (\mathbf{x}_i^\top \boldsymbol{\beta}) y_i - A(\mathbf{x}_i^\top \boldsymbol{\beta}) + \ln h(y_i)$$

Суммарная функция:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left[y_i (\mathbf{x}_i^\top \boldsymbol{\beta}) - A(\mathbf{x}_i^\top \boldsymbol{\beta}) + \ln h(y_i) \right]$$

Градиент (score-вектор)

$$u(\beta) = X^T(y - \mu)$$

где y — вектор наблюдений, μ — вектор математических ожиданий.

Матрица Гессе

$$H(\beta) = -X^T W X$$

где $W = \text{diag}(V(\mu_1), \dots, V(\mu_n))$ — диагональная матрица весов

Метод Ньютона-Рафсона

Итерационная формула:

$$\beta^{(t+1)} = \beta^{(t)} - \left(H^{(t)}\right)^{-1} u^{(t)}$$

Подставляя выражения:

$$\beta^{(t+1)} = \beta^{(t)} + \left(X^T W^{(t)} X\right)^{-1} X^T (y - \mu^{(t)})$$

Введём **скорректированный ответ** (working response):

$$z_i^{(t)} = \eta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{V(\mu_i^{(t)})}$$

Ключевое свойство:

$$X^T W^{(t)} (z^{(t)} - \boldsymbol{\eta}^{(t)}) = X^T (y - \boldsymbol{\mu}^{(t)})$$

где $z^{(t)}$ — вектор скорректированных ответов.

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} + (X^\top W^{(t)} X)^{-1} X^\top (y - \mu^{(t)}) \\ &= \beta^{(t)} + (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} (z^{(t)} - \eta^{(t)}) \\ &= (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} z^{(t)}\end{aligned}$$

Это взвешенная линейная регрессия!

- 1 Вход: $\beta^{(t)}$
- 2 $\eta^{(t)} = X\beta^{(t)}$
- 3 $\mu^{(t)} = A'(\eta^{(t)})$ (обратная функция связи)
- 4 $w_i^{(t)} = V(\mu_i^{(t)})$
- 5 $z_i^{(t)} = \eta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{w_i^{(t)}}$
- 6 $\beta^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)}$
- 7 Выход: $\beta^{(t+1)}$

Биномиальное распределение (логистическая регрессия)

- $\theta = \ln \frac{p}{1-p}$, $A(\theta) = \ln(1 + e^\theta)$
- $\mu = A'(\theta) = \frac{e^\theta}{1 + e^\theta} = p$
- $V(\mu) = \mu(1 - \mu)$

Элементы IRLS

$$w_i = \mu_i(1 - \mu_i), \quad z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i(1 - \mu_i)}$$

Пуассоновское распределение

- $\theta = \ln \lambda, A(\theta) = e^\theta$
- $\mu = A'(\theta) = e^\theta = \lambda$
- $V(\mu) = \mu$

Элементы IRLS

$$w_i = \mu_i, \quad z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}$$

Сравнительная таблица

	Биномиальное	Пуассоновское
$A(\theta)$	$\ln(1 + e^\theta)$	e^θ
$\mu = A'(\theta)$	$\frac{e^\theta}{1+e^\theta}$	e^θ
$V(\mu)$	$\mu(1 - \mu)$	μ
Функция связи	$\ln \frac{\mu}{1-\mu}$	$\ln \mu$
Вес w_i	$\mu_i(1 - \mu_i)$	μ_i
z_i	$\eta_i + \frac{y_i - \mu_i}{\mu_i(1 - \mu_i)}$	$\eta_i + \frac{y_i - \mu_i}{\mu_i}$

- Если бы мы наблюдали $g(Y_i)$, то использовали бы линейную регрессию
- Но мы наблюдаем Y_i , а не $g(Y_i)$
- Решение: **локальная линеаризация** в окрестности $\mu_i^{(t)}$

Разложение в ряд Тейлора

$$g(Y_i) \approx g(\mu_i^{(t)}) + g'(\mu_i^{(t)})(Y_i - \mu_i^{(t)}) = \eta_i^{(t)} + g'(\mu_i^{(t)})(Y_i - \mu_i^{(t)})$$

Скорректированный ответ и веса

Скорректированный ответ

$$z_i^{(t)} = \eta_i^{(t)} + g'(\mu_i^{(t)})(y_i - \mu_i^{(t)})$$

Дисперсия z_i

$$\text{Var}(z_i^{(t)}) \approx [g'(\mu_i^{(t)})]^2 \cdot V(\mu_i^{(t)})$$

Для канонической связи $g'(\mu) = 1/V(\mu)$

$$\text{Var}(z_i) = \frac{1}{V(\mu_i)} \Rightarrow \text{вес } w_i = V(\mu_i)$$

- 1 В точке $\mu^{(t)}$ строим касательное пространство
- 2 Проецируем данные на это пространство $\rightarrow z_i^{(t)}$
- 3 В касательном пространстве модель линейна
- 4 Веса учитывают неоднородность дисперсии
- 5 Решаем взвешенную линейную регрессию

IRLS = последовательность взвешенных линейных регрессий

Теорема об одношаговой оценке

Theorem (Одношаговая асимптотическая эффективность)

Пусть $\tilde{\beta}_n$ — $\sqrt{n}$ -состоятельная оценка:

$$\tilde{\beta}_n - \beta_0 = O_p(n^{-1/2})$$

Тогда одношаговая оценка

$$\hat{\beta}_{\text{one-step}} = \tilde{\beta}_n + \mathcal{I}(\tilde{\beta}_n)^{-1} u(\tilde{\beta}_n)$$

асимптотически эквивалентна МП-оценке:

$$\sqrt{n}(\hat{\beta}_{\text{one-step}} - \hat{\beta}_{MLE}) \xrightarrow{p} 0$$

и

$$\sqrt{n}(\hat{\beta}_{\text{one-step}} - \beta_0) \xrightarrow{d} N(0, \mathcal{I}(\beta_0)^{-1}),$$

где $\mathcal{I}(\beta)$ — информационная матрица Фишера.

GLM удовлетворяют всем условиям:

- **Регулярность:** экспоненциальное семейство
- **Невырожденность:** $X^T W X$ положительно определена
- **Гладкость:** $A(\theta)$, $g(\mu)$ бесконечно дифференцируемы
- **Начальные оценки:** метод моментов, эмпирические логиты

Один шаг IRLS асимптотически эквивалентен полной
МП-оценке!

Начальные оценки для GLM

Логистическая регрессия (эмпирические логиты)

$$\tilde{p}_i = \frac{y_i + 0.5}{2}, \quad \tilde{\eta}_i = \ln \frac{\tilde{p}_i}{1 - \tilde{p}_i}, \quad \tilde{\beta} = (X^T X)^{-1} X^T \tilde{\eta}$$

Пуассоновская регрессия (метод моментов)

$$\tilde{\beta} = (X^T X)^{-1} X^T \ln(y + \delta)$$

где $\delta > 0$ — малая константа.

- **Вычислительная эффективность:** один шаг IRLS вместо полной итерации
- **Асимптотический анализ:** основа для доказательства свойств МП-оценок
- **Робастность:** при больших n начальное приближение не критично
- **Компромисс:** при умеренных n полные итерации могут улучшить точность

- 1 **GLM** — единая структура для широкого круга задач
- 2 **IRLS** — итерационный алгоритм, сводящий задачу к взвешенной линейной регрессии:

$$\beta^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)}$$

- 3 **Локальная интерпретация** — GLM как последовательность линейных аппроксимаций в касательном пространстве
- 4 **Одношаговые оценки** — один шаг IRLS асимптотически эквивалентен полной МП-оценке

Логистическая регрессия — частный случай GLM с биномиальным распределением и логит-связью