Мат.модели машинного обучения. Обобщения линейных моделей регрессии и классификации

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ • 17 октября 2025

Содержание

1. **Нелинейная регрессия**
   - Нелинейная модель регрессии
   - Логистическая регрессия
   - Обобщённая аддитивная модель
2. **Обобщённая линейная модель**
   - Экспоненциальное семейство распределений
   - Максимизация правдоподобия для GLM
   - Логистическая регрессия как частный случай GLM
3. **Вероятностные модели классификации**
   - Вероятностные функции потерь
   - Вероятностный смысл регуляризации
   - Пример: кредитный скоринг

Нелинейная модель регрессии

Дано: обучающая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$
$y_i = y(x_i)$, $y : X \to Y$ — неизвестная регрессионная зависимость
Найти: параметры $\alpha \in \mathbb{R}^p$ модели регрессии $f(x, \alpha)$
Критерий: метод наименьших квадратов (МНК)

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} \big(f(x_i, \alpha) - y_i\big)^2 \to \min_\alpha$$

Метод Ньютона–Рафсона:

1. Начальное приближение $\alpha^0 = (\alpha_1^0, \ldots, \alpha_p^0)$.
2. Итерационный процесс

$$\alpha^{t+1} := \alpha^t - h_t \big(Q''(\alpha^t)\big)^{-1} \nabla Q(\alpha^t),$$

$\nabla Q(\alpha^t)$ — градиент функционала Q в точке α^t, вектор из $\mathbb{R}^p$
$Q''(\alpha^t)$ — гессиан функционала Q в точке α^t, матрица из $\mathbb{R}^{p \times p}$
h_t — величина шага (можно полагать $h_t = 1$).

Метод Ньютона-Рафсона

Компоненты градиента:

$$\frac{\partial Q(\alpha)}{\partial \alpha_j} = 2\sum_{i=1}^{\ell}\big(f(x_i,\alpha)-y_i\big)\frac{\partial f(x_i,\alpha)}{\partial \alpha_j}.$$

Компоненты гессиана:

$$\frac{\partial^2 Q(\alpha)}{\partial \alpha_j \partial \alpha_k} = 2\sum_{i=1}^{\ell}\frac{\partial f(x_i,\alpha)}{\partial \alpha_j}\frac{\partial f(x_i,\alpha)}{\partial \alpha_k} \underbrace{-2\sum_{i=1}^{\ell}\big(f(x_i,\alpha)-y_i\big)\frac{\partial^2 f(x_i,\alpha)}{\partial \alpha_j \partial \alpha_k}}_{\text{при линеаризации полагается } = 0}.$$

Не хотелось бы обращать гессиан на каждой итерации...

Линеаризация $f(x_i,\alpha)$ в окрестности текущего α^t:

$$f(x_i,\alpha) = f(x_i,\alpha^t) + \sum_{j=1}^{p}\frac{\partial f(x_i,\alpha_j^t)}{\partial \alpha_j}\big(\alpha_j-\alpha_j^t\big) + o\big(\alpha_j-\alpha_j^t\big).$$

Метод Ньютона-Гаусса

Матричные обозначения:
$F_t = \left(\frac{\partial f}{\partial \alpha_j}(x_i, \alpha^t)\right)_{\ell \times p}$ — матрица первых производных;
$f_t = \left(f(x_i, \alpha^t)\right)_{\ell \times 1}$ — вектор значений f.

Формула t-й итерации метода Ньютона–Гаусса:

$$\alpha^{t+1} := \alpha^t - h_t \underbrace{\left(F_t^\mathsf{T} F_t\right)^{-1} F_t^\mathsf{T} (f_t - y)}_{\beta}.$$

β — это решение задачи многомерной линейной регрессии

$$\|F_t \beta - (f_t - y)\|^2 \to \min_\beta.$$

Нелинейная регрессия сведена к серии линейных регрессий.

Скорость сходимости — как и у метода Ньютона–Рафсона, но для вычислений можно применять линейные методы.

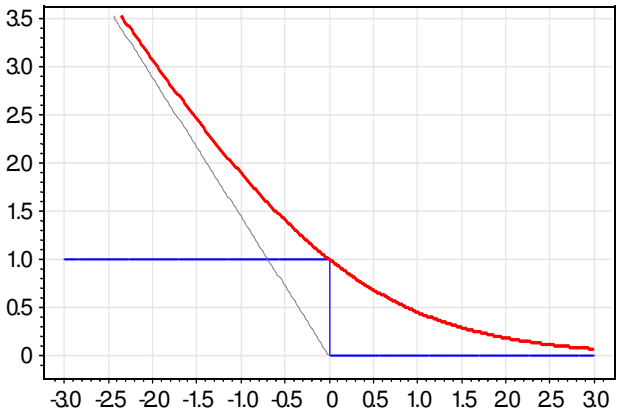
Задача классификации. Логистическая регрессия

$Y = \{-1, +1\}$ — два класса, $a(x, w) = \text{sign}(w^\mathsf{T} x)$, $x, w \in \mathbb{R}^n$.

Функционал аппроксимированного эмпирического риска:

$$Q(w) = \sum_{i=1}^{\ell} \big[M_i(w) < 0\big] \quad \leqslant \quad \sum_{i=1}^{\ell} \mathscr{L}\big(w^\mathsf{T} x_i y_i\big) \to \min_w,$$

где $\mathscr{L}(M) = \log\big(1 + e^{-M}\big)$ — логарифмическая функция потерь



$M_i = w^\mathsf{T} x_i y_i$

Метода Ньютона-Рафсона

Метода Ньютона-Рафсона для минимизации функционала $Q(w)$:

$$w^{t+1} := w^t - h_t\big(Q''(w^t)\big)^{-1}\nabla Q(w^t),$$

Элементы градиента — вектора первых производных $\nabla Q(w^t)$:

$$\frac{\partial Q(w)}{\partial w_j} = -\sum_{i=1}^{\ell}(1-\sigma_i)y_i f_j(x_i), \quad j = 1,\ldots,n.$$

Элементы гессиана — матрицы вторых производных $Q''(w^t)$:

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = \sum_{i=1}^{\ell}(1-\sigma_i)\sigma_i f_j(x_i) f_k(x_i), \quad j,k = 1,\ldots,n,$$

где $\sigma_i = \sigma(y_i w^{\mathsf{T}} x_i)$, $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция.

Снова сведение к задаче линейной регрессии

В матричных обозначениях $F = \big(f_j(x_i)\big)_{\ell \times n}$, $D = \operatorname{diag}\big((1-\sigma_i)\sigma_i\big)$

$$\big(Q''(w)\big)^{-1} \nabla Q(w) = -\big(F^\intercal D F\big)^{-1} F^\intercal \big(\tfrac{y_i}{\sigma_i}\big).$$

Это совпадает с МНК-решением задачи линейной регрессии со взвешенными объектами и модифицированными ответами:

$$Q(w) = \sum_{i=1}^{\ell} (1-\sigma_i)\sigma_i \left(w^\intercal x_i - \frac{y_i}{\sigma_i} \right)^2 \to \min_w .$$

Интерпретация:

- $\sigma_i = \mathrm{P}(y_i|x_i)$ — вероятность правильной классификации x_i
- чем ближе x_i к границе, тем больше вес $(1-\sigma_i)\sigma_i$
- чем выше вероятность ошибки, тем больше $\frac{1}{\sigma_i}$

ВЫВОД: на каждой итерации происходит более точная настройка на «наиболее трудных» объектах.

МНК с итерационным перевзвешиванием объектов

Метод IRLS — Iteratively Reweighted Least Squares

Вход: F, y — матрица «объекты–признаки» и вектор ответов;
Выход: w — вектор коэффициентов линейной комбинации.

1: $w := (F^\mathsf{T}F)^{-1}F^\mathsf{T}y$ — нулевое приближение, обычный МНК;
2: **для** $t := 1, 2, 3, \ldots$
3: $\quad \sigma_i = \sigma(y_i w^\mathsf{T} x_i)$ для всех $i = 1, \ldots, \ell$;
4: $\quad \gamma_i := \sqrt{(1-\sigma_i)\sigma_i}$ для всех $i = 1, \ldots, \ell$;
5: $\quad \tilde{F} := \operatorname{diag}(\gamma_1, \ldots, \gamma_\ell)F$;
6: $\quad \tilde{y}_i := y_i\sqrt{(1-\sigma_i)/\sigma_i}$ для всех $i = 1, \ldots, \ell$;
7: $\quad$ выбрать градиентный шаг h_t;
8: $\quad w := w + h_t(\tilde{F}^\mathsf{T}\tilde{F})^{-1}\tilde{F}^\mathsf{T}\tilde{y}$;
9: $\quad$ **если** $\{\sigma_i\}$ мало изменились **то** выйти из цикла;

Обобщённая аддитивная модель (Generalized Additive Model)

Регрессия с нелинейными преобразованиями признаков φ_j:

$$f(x,\alpha) = \sum_{j=1}^{n} \varphi_j(f_j(x),\alpha_j)$$

В частности, при $\varphi_j(f_j(x),\alpha_j) = \alpha_j f_j(x)$ это линейная модель

Идея 1: поочерёдно уточнять φ_j по выборке $\left(f_j(x_i), z_i\right)_{i=1}^{\ell}$:

$$\sum_{i=1}^{\ell}\Big(\varphi_j(f_j(x_i),\alpha_j) - \underbrace{\big(y_i - \sum_{k\neq j}\varphi_k(f_k(x_i),\alpha_k)\big)}_{z_i}\Big)^2 + \tau R(\alpha_j) \;\to\; \min_{\alpha_j}$$

Идея 2: постепенно уменьшать τ у регуляризатора гладкости

$$R(\alpha_j) = \int \left(\varphi_j''(\zeta,\alpha_j)\right)^2 d\zeta$$

В качестве φ_j использовать сплайны или ядерное сглаживание

Метод backfitting [Хасти, Тибширани, 1986]

Вход: F, y — матрица «объекты–признаки» и вектор ответов;
Выход: $\varphi_j(f_j, \alpha_j)$ — обучаемые преобразования признаков.

1: начальное приближение:
   $\alpha :=$ решение задачи МЛР с признаками $f_j(x)$;
   $\varphi_j(f_j, \alpha_j) := \alpha_j f_j(x), \;\; j = 1, \ldots, n$;
2: **повторять**
3: **для** $j = 1, \ldots, n$
4: $z_i := y_i - \sum\limits_{k=1, k\neq j}^{n} \varphi_k(f_k(x_i), \alpha_k), \;\; i = 1, \ldots, \ell$;
5: $\alpha_j := \arg\min\limits_{\alpha} \sum\limits_{i=1}^{\ell} \big(\varphi(f_j(x_i), \alpha) - z_i\big)^2 + \tau R(\alpha)$;
6: уменьшить коэффициент регуляризации τ;
7: **пока** $Q(\alpha, X^\ell)$ и/или $Q(\alpha, X^k)$ заметно уменьшаются;

Вероятностная постановка задачи регрессии

Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$

Найти: параметр w модели регрессии с гауссовским шумом:

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2),\ \mu_i = a(x_i, w) = \mathsf{E}y_i, \quad i = 1, \ldots, \ell.$$

Эквивалентная запись: $y_i = a(x_i, w) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

Критерий максимума правдоподобия эквивалентен МНК:

$$p(\varepsilon_1, \ldots, \varepsilon_\ell | w) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\varepsilon_i^2\right) \to \max_w$$

$$-\ln p(\varepsilon_1, \ldots, \varepsilon_\ell | w) = \text{const} + \frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\sigma_i^2}\big(a(x_i, w) - y_i\big)^2 \to \min_w$$

Что использовать вместо метода наименьших квадратов, если y_i не гауссовские, в частности, если y_i дискретнозначные?

Обобщённая линейная модель (Generalized Linear Model, GLM)

Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$
Найти: параметр w *обобщённой линейной модели* (GLM):

$$y_i \sim \mathrm{Exp}(\theta_i, \phi_i), \qquad \theta_i = g(\mathrm{E}y_i) = a(x_i, w) = \langle x_i, w \rangle,$$

вместо предположения о гауссовости y_i, теперь вводится Exp — *экспоненциальное семейство распределений* (exponential family) с параметром θ_i, параметром масштаба ϕ_i (scale) и параметрами-функциями $c(\theta)$, $h(y, \phi)$:

$$p(y_i|x_i) = \exp\left(\frac{y_i\theta_i - c(\theta_i)}{\phi_i} + h(y_i, \phi_i)\right)$$

Критерий максимума правдоподобия:

$$Q(w) = \ln \prod_{i=1}^{\ell} p(y_i|x_i) = \sum_{i=1}^{\ell} \frac{y_i\langle x_i, w\rangle - c\big(\langle x_i, w\rangle\big)}{\phi_i} \to \max_{w, \{\phi_i\}}$$

Экспоненциальное семейство распределений

Exp — *экспоненциальное семейство распределений*
с параметрами θ_i, ϕ_i и параметрами-функциями $c(\theta)$, $h(y, \phi)$:

$$p(y_i|\theta_i, \phi_i) = \exp\left(\frac{y_i\theta_i - c(\theta_i)}{\phi_i} + h(y_i, \phi_i)\right)$$

Свойства экспоненциальных распределений
Математическое ожидание и дисперсия с.в. $y_i \sim \mathrm{Exp}(\theta_i, \phi_i)$:

$$\mu_i = \mathsf{E}y_i = c'(\theta_i) \quad \Rightarrow \quad \theta_i = [c']^{-1}(\mu_i) = g(\mathsf{E}y_i)$$
$$\mathsf{D}y_i = \phi_i c''(\theta_i)$$

$g(\mu) = [c']^{-1}(\mu)$ — монотонная *функция связи* (link function)

Нормальная линейная модель — частный случай GLM:

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i = \langle x_i, w\rangle = \mathsf{E}y_i \qquad g(\mu_i) = \mu_i$$

Примеры распределений из экспоненциального семейства

Нормальное (гауссовское) распределение, $y_i \in \mathbb{R}$:

$$p(y_i|\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\Big(-\frac{1}{2\sigma_i^2}(y_i-\mu_i)^2\Big) =$$

$$= \exp\left(\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma_i^2} - \frac{y_i^2}{2\sigma_i^2} - \frac{1}{2}\ln(2\pi\sigma_i^2)\right);$$

$$\theta_i = g(\mu_i) = \mu_i, \qquad c(\theta_i) = \tfrac{1}{2}\mu_i^2 = \tfrac{1}{2}\theta_i^2, \qquad \phi_i = \sigma_i^2.$$

Распределение Бернулли, $y_i \in \{0,1\}$:

$$p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left(y_i \ln\tfrac{\mu_i}{1-\mu_i} + \ln(1-\mu_i)\right);$$

$$\theta_i = g(\mu_i) = \ln\tfrac{\mu_i}{1-\mu_i}, \qquad c(\theta_i) = -\ln(1-\mu_i) = \ln(1+e^{\theta_i}).$$

Примеры распределений из экспоненциального семейства

Биномиальное распределение, $y_i \in \{0, 1, \ldots, n_i\}$:

$$p(y_i|\mu_i, n_i) = C_{n_i}^{y_i} \left(\frac{\mu_i}{n_i}\right)^{y_i} \left(1 - \frac{\mu_i}{n_i}\right)^{n_i - y_i} =$$

$$= \exp\left(y_i \ln \frac{\mu_i}{n_i - \mu_i} + n_i \ln(n_i - \mu_i) + \ln C_{n_i}^{y_i} - n_i \ln n_i\right);$$

$$\theta_i = g(\mu_i) = \ln \frac{\mu_i}{n_i - \mu_i}, \qquad c(\theta_i) = -n_i \ln(n_i - \mu_i) = n_i \ln \frac{1 + e^{\theta_i}}{n_i}.$$

Пуассоновское распределение, $y_i \in \{0, 1, 2, \ldots\}$:

$$p(y_i|\mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \exp\left(\frac{y_i \ln(\mu_i) - \mu_i}{1} - \ln y_i!\right);$$

$$\theta_i = g(\mu_i) = \ln(\mu_i), \qquad c(\theta_i) = \mu_i = e^{\theta_i}, \qquad \phi_i = 1.$$

Примеры распределений из экспоненциального семейства

- нормальное (гауссовское)
- распределение Пуассона
- биномиальное и мультиномиальное
- геометрическое
- χ^2-распределение
- бета-распределение
- гамма-распределение
- распределение Дирихле
- распределение Лапласа с фиксированным матожиданием

Контр-примеры не экспоненциальных распределений:

- t-распределение Стьюдента, Коши, гипергеометрическое

Максимизация правдоподобия для GLM

Принцип максимума правдоподобия:

$$Q(w) = \ln \prod_{i=1}^{\ell} p(y_i|\theta_i, \phi_i) = \sum_{i=1}^{\ell} \frac{y_i\theta_i - c(\theta_i)}{\phi_i} \to \max_w,$$

где θ_i линейно зависит от w: $\theta_i = \langle x_i, w\rangle = \sum_{j=1}^{n} w_j f_j(x_i)$.

Метод Ньютона-Рафсона: $w^{t+1} := w^t + h_t\big(Q''(w^t)\big)^{-1}\nabla Q(w^t)$

Компоненты вектора градиента $\nabla Q(w)$:

$$\frac{\partial Q(w)}{\partial w_j} = \sum_{i=1}^{\ell} \frac{y_i - c'(\theta_i)}{\phi_i} f_j(x_i).$$

Компоненты матрицы Гессе $Q''(w)$:

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = -\sum_{i=1}^{\ell} \frac{c''(\theta_i)}{\phi_i} f_j(x_i) f_k(x_i).$$

Матричные обозначения

$F = \big(f_j(x_i)\big)_{\ell \times n}$ — матрица «объекты–признаки»;

$\tilde{F} = W_t F, \quad W_t = \operatorname{diag}\left(\sqrt{\frac{1}{\phi_i} c''(\theta_i)}\right)$ — веса объектов,

$\theta_i = \langle x_i, w^t \rangle$;

$\tilde{y} = \big(\tilde{y}_i\big)_{\ell \times 1}, \quad \tilde{y}_i = \frac{y_i - c'(\theta_i)}{\sqrt{\phi_i c''(\theta_i)}}$ — модифицированный вектор ответов.

Тогда метод Ньютона-Рафсона снова приводит к IRLS:

$$w^{t+1} := w^t - h_t \underbrace{\big(F^\intercal W_t W_t F\big)^{-1} F^\intercal W_t}_{(\tilde{F}^\intercal \tilde{F})^{-1} \tilde{F}^\intercal} \Bigg(\underbrace{\sqrt{\frac{\phi_i}{c''(\theta_i)}} \frac{y_i - c'(\theta_i)}{\phi_i}}_{\tilde{y}_i} \Bigg)_{\ell \times 1}$$

Это совпадает с МНК-решением линейной задачи регрессии со взвешенными объектами и модифицированными ответами:

$$Q(w) = \|\tilde{F} w - \tilde{y}\|^2 \to \min_w$$

МНК с итерационным перевзвешиванием объектов

IRLS — Iteratively Reweighted Least Squares

Вход: F, y — матрица «объекты–признаки» и вектор ответов;
Выход: w — вектор коэффициентов линейной комбинации.

1: $w := (F^\mathsf{T}F)^{-1}F^\mathsf{T}y$ — нулевое приближение, обычный МНК;
2: **для** $t := 1, 2, 3, \ldots$
3: $\quad \theta_i = \langle x_i, w^t \rangle$ для всех $i = 1, \ldots, \ell$;
4: $\quad \gamma_i := \sqrt{\frac{1}{\phi_i} c''(\theta_i)}$ для всех $i = 1, \ldots, \ell$;
5: $\quad \tilde{F} := \operatorname{diag}(\gamma_1, \ldots, \gamma_\ell) F$;
6: $\quad \tilde{y}_i := \frac{y_i - c'(\theta_i)}{\phi_i \gamma_i}$ для всех $i = 1, \ldots, \ell$;
7: $\quad$ выбрать градиентный шаг h_t;
8: $\quad w := w + h_t(\tilde{F}^\mathsf{T}\tilde{F})^{-1}\tilde{F}^\mathsf{T}\tilde{y}$;
9: $\quad$ **если** $\{\theta_i\}$ мало изменились **то** выйти из цикла;

Двухклассовая логистическая регрессия

Распределение Бернулли, $y_i \in \{0,1\}$: $p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$

$\theta_i = g(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$ $\qquad$ $\mathsf{E}y_i = \mu_i = g^{-1}(\theta_i) = \frac{1}{1+\exp(-\theta_i)} \equiv \sigma(\theta_i)$

Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{0,1\} \sim p(y_i|\mu_i)$

Найти: вероятностную модель $\mathsf{E}(y|x) = p(y{=}1|x) = \sigma\big(\langle x, w\rangle\big)$

Критерий: максимум log-правдоподобия (log-loss)

$$Q(w) = \sum_{i=1}^{\ell} \ln p(y_i|\mu_i) = \sum_{i=1}^{\ell} y_i \ln \mu_i + (1-y_i)\ln(1-\mu_i) \to \max_w$$

Удобная перекодировка: $y_i \in \{0,1\} \;\to\; \tilde{y}_i = 2y_i - 1 \in \{-1,1\}$

$$-\sum_{i=1}^{\ell} \ln p(\tilde{y}_i|x_i) = \sum_{i=1}^{\ell} \ln\big(1+\exp(-\underbrace{\langle w, x_i\rangle \tilde{y}_i}_{\text{margin}})\big) \;\to\; \min_w$$

Логистическая регрессия как частный случай GLM

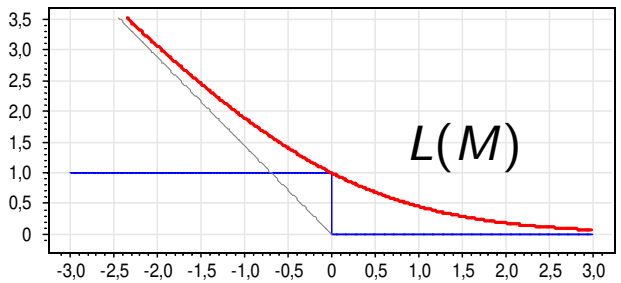
Всего лишь из двух предположений:

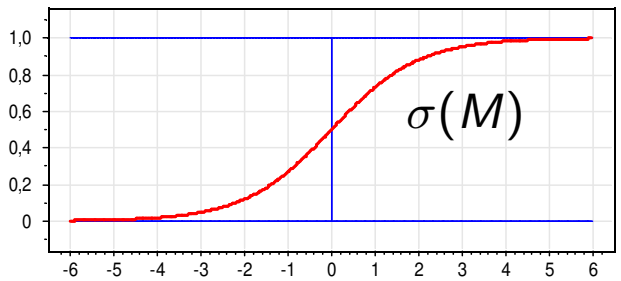
- y_i — бернуллиевские случайные величины с $\mathsf{E}y_i = \mu_i$
- параметр связан с линейной моделью: $\theta_i = g(\mu_i) = \langle x_i, w \rangle$

следуют важнейшие свойства логистической регрессии:

- логарифмическая функция потерь $\ln\bigl(1 + \exp(-\langle x_i, w\rangle \tilde{y}_i)\bigr)$;
- сигмоидная функция связи $\mathsf{P}(y_i|x_i) = \sigma(\langle x_i, w\rangle \tilde{y}_i)$;
- связь линейной модели с *отношением шансов* (odds ratio):

$$\langle x_i, w\rangle = \ln \frac{\mu_i}{1-\mu_i} = \ln \frac{\mathsf{P}(y_i=1|x_i)}{\mathsf{P}(y_i=0|x_i)}$$





Многоклассовая логистическая регрессия

Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in Y$, $2 \leqslant |Y| < \infty$

Найти: линейную модель классификации

$$a(x) = \arg\max_{y \in Y} \langle w_y, x \rangle, \quad x, w_y \in \mathbb{R}^n$$

и вероятность того, что объект x относится к классу y:

$$P(y|x, w) = \frac{\exp\langle w_y, x \rangle}{\sum_{z \in Y} \exp\langle w_z, x \rangle} = \underset{y \in Y}{\mathrm{SoftMax}} \langle w_y, x \rangle,$$

функция SoftMax: $\mathbb{R}^Y \to \mathbb{R}^Y$ переводит произвольный вектор в нормированный вектор дискретного распределения.

Критерий: максимум log-правдоподобия (log-loss)

$$Q(w) = \sum_{i=1}^{\ell} \log P(y_i|x_i, w) \;\to\; \max_w$$

Калибровка Платта (classifier with probabilistic output)

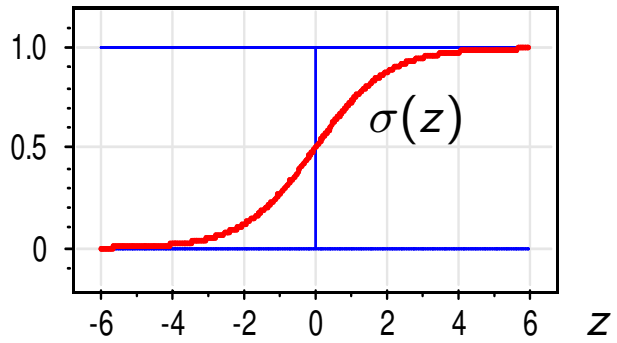
Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$;
ранее построенная модель классификации $a(x) = \text{sign}\, g(x, w)$

Найти: вероятностную модель классификации $\mathrm{P}(y|x)$

Модель условной вероятности:

$$\pi(x; a, b) = \mathrm{P}(y=1|x) = \sigma\big(a g(x, w) + b\big)$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция



Критерий: максимум log-правдоподобия для калибровки коэффициентов a, b по *контрольной* выборке:

$$\sum_{y_i=-1} \log\big(1 - \pi(x_i; a, b)\big) + \sum_{y_i=+1} \log \pi(x_i; a, b) \to \max_{a,b}$$

Задача классификации и принцип максимума правдоподобия

Дано: простая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$, порождаемая неизвестной плотностью $p(x, y)$ на в.п. $X \times Y$, $|Y| < \infty$

Найти: параметр w модели условной вероятности $P(y|x, w)$

$p(x, y; w) = P(y|x, w)p(x)$ — модель совместной плотности,
$p(x)$ — неизвестное и непараметризуемое распределение на X

Критерий: метод максимума правдоподобия (ММП)

$$p(X^\ell; w) = \prod_{i=1}^{\ell} p(x_i, y_i; w) = \prod_{i=1}^{\ell} P(y_i|x_i, w)\cancel{p(x_i)} \;\to\; \max_w$$

Максимум логарифма правдоподобия (log-likelihood, log-loss):

$$L(w) = \sum_{i=1}^{\ell} \ln P(y_i|x_i, w) \to \max_w$$

Связь правдоподобия и аппроксимации эмпирического риска

Максимизация логарифма правдоподобия,
$P(y|x,w)$ — модель *условной вероятности класса*:

$$L(w) = \sum_{i=1}^{\ell} \ln P(y_i|x_i,w) \to \max_w$$

Минимизация аппроксимированного эмпирического риска,
$g(x,w)$ — модель *разделяющей поверхности*, $Y = \{\pm 1\}$:

$$Q(w) = \sum_{i=1}^{\ell} \mathscr{L}\big(y_i g(x_i,w)\big) \to \min_w;$$

Эти два принципа эквивалентны, если положить

$$-\ln P(y_i|x_i,w) = \mathscr{L}\big(y_i g(x_i,w)\big).$$

$$\boxed{\text{модель } P(y|x,w)} \rightleftarrows \boxed{\text{модель } g(x,w) \text{ и } \mathscr{L}(M)}.$$

Вероятностный смысл регуляризации

Дано: простая выборка $(x_i, y_i)_{i=1}^{\ell} \sim p(x, y)$,
$p(w; \gamma)$ — *априорное распределение* параметров модели,
γ — вектор *гиперпараметров*

Найти: параметр w модели условной вероятности $P(y|x, w)$

Теперь случайна как выборка X^{ℓ}, так и модель $w \sim p(w; \gamma)$
Совместное правдоподобие данных и модели:

$$p(X^{\ell}, w) = p(X^{\ell}|w)\, p(w; \gamma)$$

Критерий *максимума апостериорной вероятности*
(Maximum a Posteriori Probability, MAP):

$$\ln p(X^{\ell}, w) = \underbrace{\sum_{i=1}^{\ell} \ln P(y_i|x_i, w)}_{\text{log правдоподобия}} + \underbrace{\ln p(w; \gamma)}_{\substack{\text{регуляризатор,} \\ \text{не зависит от } X^{\ell}}} \to \max_w$$

Примеры: априорные распределения Гаусса и Лапласа

Линейная модель $a(x,w) = \text{sign}\langle x,w\rangle$ или $\langle x,w\rangle$

Ограничения на параметры: $\mathsf{E}w_j = 0$, $\mathsf{E}w_j w_k = 0$, $\mathsf{D}w_j = C$

Распределение Гаусса и квадратичный (L_2) регуляризатор:

$$p(w;C) = \frac{1}{(2\pi C)^{n/2}} \exp\left(-\frac{\|w\|^2}{2C}\right), \quad \|w\|^2 = \sum_{j=1}^{n} w_j^2,$$

$$-\ln p(w;C) = \tfrac{1}{2C}\|w\|^2 + \cancel{\text{const}}$$

Распределение Лапласа и абсолютный (L_1) регуляризатор:

$$p(w;C) = \frac{1}{(2C)^{n}} \exp\left(-\frac{\|w\|}{C}\right), \quad \|w\| = \sum_{j=1}^{n} |w_j|,$$

$$-\ln p(w;C) = \tfrac{1}{C}\|w\| + \cancel{\text{const}}$$

C — гиперпараметр, $\tau = \frac{1}{C}$ — коэффициент регуляризации.

Скоринг — линейная вероятностная модель принятия решений

Пример. Кредитный скоринг:

- x_i — заёмщики
- $y_i = -1\,(\text{bad}), +1\,(\text{good})$

Бинаризация признаков $f_j(x)$:

$$b_{jk}(x) = \big[f_j(x) \text{ из } k\text{-го интервала}\big]$$

Линейная модель классификации:

$$a(x, w) = \text{sign} \sum_{j,k} w_{jk} b_{jk}(x).$$

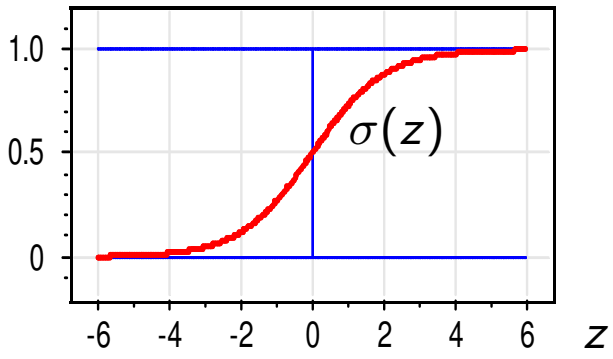
Вес признака w_{jk} равен его вкладу в общую сумму баллов (score).

признак j	интервал k	w_{jk}
Возраст	до 25	5
	25 - 40	10
	40 - 50	15
	50 и больше	10
Собственность	владелец	20
	совладелец	15
	съемщик	10
	другое	5
Работа	руководитель	15
	менеджер среднего звена	10
	служащий	5
	другое	0
Стаж	1/безработный	0
	1..3	5
	3..10	10
	10 и больше	15
Работа_мужа /жены	нет/домохозяйка	0
	руководитель	10
	менеджер среднего звена	5
	служащий	1

Оценивание рисков в скоринге

Логистическая регрессия не только определяет веса w, но и оценивает *апостериорные вероятности* классов:

$$\mathrm{P}(y|x) = \sigma\big(\langle w, x\rangle y\big) = \frac{1}{1 + e^{-\langle w,x\rangle y}}$$



Оценка *риска* (математического ожидания) потерь объекта x:

$$R(x) = \sum_{y \in Y} D_{xy}\, \mathrm{P}(y|x),$$

где D_{xy} — величина потери для объекта x с исходом y, причём если $y = -1$ (bad), то $D_{xy} > 0$; если $y = +1$ (good), то $D_{xy} < 0$

Оценка $R(x)$ говорит о том, сколько мы потеряем в среднем. Но сколько мы рискуем потерять в 1% худших случаев?

Методика VaR (Value at Risk)

Стохастическое моделирование: $N = 10^4$ раз

- для каждого x_i разыгрывается исход $y_i \sim P(y|x_i)$;
- вычисляется сумма потерь по портфелю $V = \sum_{i=1}^{\ell} D_{x_i y_i}$;

99%-квантиль эмпирического распределения потерь определяет величину резервируемого капитала

Резюме в конце лекции

- Нелинейная регрессия
  — сводится к последовательности линейных регрессий
  — метод Ньютона-Рафсона приводит к IRLS
- Логистическая регрессия
  — не регрессия, а классификация
  — метод Ньютона-Рафсона приводит к IRLS
- Обобщённая линейная модель (GLM)
  — мощно обобщает обычную и логистическую регрессию
  — метод Ньютона-Рафсона приводит к IRLS
- Обобщённая аддитивная регрессия (GAM, backfitting)
  — сводится к серии одномерных сглаживаний
- Вероятностный смысл *функции потерь*
  — это $-\ln$ вероятностной модели данных
- Вероятностный смысл *регуляризации*
  — это $-\ln$ априорного распределения параметров модели