

Методы восстановления регрессии

К. В. Воронцов
vokov@forecsys.ru

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

3 апреля 2012

Метод наименьших квадратов

- X — объекты (часто \mathbb{R}^n); Y — ответы (часто \mathbb{R} , реже \mathbb{R}^m);
 $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;
 $y_i = y(x_i)$, $y: X \rightarrow Y$ — неизвестная зависимость;
- $a(x) = f(x, \alpha)$ — модель зависимости,
 $\alpha \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

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

где w_i — вес, степень важности i -го объекта.

$Q(\alpha^*, X^\ell)$ — остаточная сумма квадратов
(residual sum of squares, RSS).

Метод максимума правдоподобия

Модель данных с некоррелированным гауссовским шумом:

$$y(x_i) = f(x_i, \alpha) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad i = 1, \dots, \ell.$$

Метод максимума правдоподобия (ММП):

$$L(\varepsilon_1, \dots, \varepsilon_\ell | \alpha) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2} \varepsilon_i^2\right) \rightarrow \max_{\alpha};$$

$$-\ln L(\varepsilon_1, \dots, \varepsilon_\ell | \alpha) = \text{const}(\alpha) + \frac{1}{2} \sum_{i=1}^{\ell} \frac{1}{\sigma_i^2} (f(x_i, \alpha) - y_i)^2 \rightarrow \min_{\alpha};$$

Теорема

Решения МНК и ММП, совпадают, причём веса объектов обратно пропорциональны дисперсии шума, $w_i = \sigma_i^{-2}$.

Содержание

- 1 Непараметрическая регрессия**
 - Формула Надарая–Ватсона
 - Выбор ядра K и ширины окна h
 - Отсев выбросов
- 2 Многомерная линейная регрессия**
 - Решение задачи наименьших квадратов
 - Сингулярное разложение
 - Регуляризация (гребневая регрессия)
 - Лассо Тибширани
- 3 Метод главных компонент**
 - Постановка задачи
 - Основная теорема
 - Решение задачи наименьших квадратов
- 4 Нелинейная регрессия**
 - Нелинейная модель регрессии
 - Логистическая регрессия
 - Нелинейные преобразования признаков

Формула Надарая–Ватсона

Приближение константой $a(x) = \alpha$ в окрестности $x \in X$:

$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x) (\alpha - y_i)^2 \rightarrow \min_{\alpha \in \mathbb{R}}$$

где $w_i(x) = K\left(\frac{\rho(x, x_i)}{h}\right)$ — веса объектов x_i относительно x ;
 $K(r)$ — ядро, невозрастающее, ограниченное, гладкое;
 h — ширина окна сглаживания.

Формула ядерного сглаживания Надарая–Ватсона:

$$a_h(x; X^\ell) = \frac{\sum_{i=1}^{\ell} y_i w_i(x)}{\sum_{i=1}^{\ell} w_i(x)} = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)}.$$

Обоснование формулы Надарая–Ватсона

Теорема

Пусть выполнены следующие условия:

- 1) выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$ простая, из распределения $p(x, y)$;
- 2) ядро $K(r)$ ограничено: $\int_0^\infty K(r) dr < \infty$, $\lim_{r \rightarrow \infty} rK(r) = 0$;
- 3) зависимость $E(y|x)$ не имеет вертикальных асимптот:
 $E(y^2|x) = \int_Y y^2 p(y|x) dy < \infty$ при любом $x \in X$;
- 4) последовательность h_ℓ убывает, но не слишком быстро:
 $\lim_{\ell \rightarrow \infty} h_\ell = 0$, $\lim_{\ell \rightarrow \infty} \ell h_\ell = \infty$.

Тогда имеет место сходимость по вероятности:

$$a_{h_\ell}(x; X^\ell) \xrightarrow{P} E(y|x) \text{ в любой точке } x \in X,$$

в которой $E(y|x)$, $p(x)$ и $D(y|x)$ непрерывны и $p(x) > 0$.

- Ядро $K(r)$
 - существенно влияет на гладкость функции $a_h(x)$,
 - слабо влияет на качество аппроксимации.
- Ширина окна h
 - существенно влияет на качество аппроксимации.
- При неравномерной сетке $\{x_i\}$ — переменная ширина окна:

$$w_i(x) = K\left(\frac{\rho(x, x_i)}{h(x)}\right),$$

где $h(x) = \rho(x, x^{(k+1)})$, $x^{(k+1)}$ — k -й сосед объекта x .

- Оптимизация ширины окна по скользящему контролю:

$$\text{LOO}(h, X^\ell) = \sum_{i=1}^{\ell} \left(a_h(x_i; X^\ell \setminus \{x_i\}) - y_i \right)^2 \rightarrow \min_h.$$

Локально взвешенное сглаживание (LOWESS — LOcally WEighted Scatter plot Smoothing)

Основная идея:

чем больше величина ошибки $\varepsilon_i = |a_h(x_i; X^\ell \setminus \{x_i\}) - y_i|$, тем в большей степени прецедент (x_i, y_i) является выбросом, и тем меньше должен быть его вес $w_i(x)$.

Эвристика:

домножить веса $w_i(x)$ на коэффициенты $\gamma_i = \tilde{K}(\varepsilon_i)$,
где \tilde{K} — ещё одно ядро, вообще говоря, отличное от $K(r)$.

Рекомендация:

квартическое ядро $\tilde{K}(\varepsilon) = K_Q\left(\frac{\varepsilon}{6 \operatorname{med}\{\varepsilon_i\}}\right)$,
где $\operatorname{med}\{\varepsilon_i\}$ — медиана вариационного ряда ошибок.

Алгоритм LOWESS

Вход: X^ℓ — обучающая выборка;

Выход: коэффициенты γ_i , $i = 1, \dots, \ell$;

1: инициализация: $\gamma_i := 1$, $i = 1, \dots, \ell$;

2: **повторять**

3: **для всех** объектов $i = 1, \dots, \ell$

4: вычислить оценки скользящего контроля:

$$a_i := a_h(x_i; X^\ell \setminus \{x_i\}) = \frac{\sum_{j=1, j \neq i}^{\ell} y_j \gamma_j K\left(\frac{\rho(x_i, x_j)}{h(x_i)}\right)}{\sum_{j=1, j \neq i}^{\ell} \gamma_j K\left(\frac{\rho(x_i, x_j)}{h(x_i)}\right)};$$

5: **для всех** объектов $i = 1, \dots, \ell$

6: $\gamma_i := \tilde{K}(|a_i - y_i|)$;

7: **пока** коэффициенты γ_i не стабилизируются;

Многомерная линейная регрессия

$f_1(x), \dots, f_n(x)$ — числовые признаки;

Модель многомерной линейной регрессии:

$$f(x, \alpha) = \sum_{j=1}^n \alpha_j f_j(x), \quad \alpha \in \mathbb{R}^n.$$

Матричные обозначения:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad y_{\ell \times 1} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \alpha_{n \times 1} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}.$$

Функционал квадрата ошибки:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} (f(x_i, \alpha) - y_i)^2 = \|F\alpha - y\|^2 \rightarrow \min_{\alpha}.$$

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^T(F\alpha - y) = 0,$$

откуда следует *нормальная система* задачи МНК:

$$F^T F \alpha = F^T y,$$

где $F^T F$ _{$n \times n$} — ковариационная матрица набора признаков f_1, \dots, f_n .

Решение системы: $\alpha^* = (F^T F)^{-1} F^T y = F^+ y$.

Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2$,

где $P_F = F F^+ = F(F^T F)^{-1} F^T$ — проекционная матрица.

Сингулярное разложение

Произвольная $\ell \times n$ -матрица представима в виде *сингулярного разложения* (singular value decomposition, SVD):

$$F = VDU^T.$$

Основные свойства сингулярного разложения:

- 1 $\ell \times n$ -матрица $V = (v_1, \dots, v_n)$ ортогональна, $V^T V = I_n$, столбцы v_j — собственные векторы матрицы FF^T ;
- 2 $n \times n$ -матрица $U = (u_1, \dots, u_n)$ ортогональна, $U^T U = I_n$, столбцы u_j — собственные векторы матрицы $F^T F$;
- 3 $n \times n$ -матрица D диагональна, $D = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$, $\lambda_j \geq 0$ — собственные значения матриц $F^T F$ и FF^T .

Решение МНК через сингулярное разложение

Псевдообратная F^+ , вектор МНК-решения α^* ,
 МНК-аппроксимация целевого вектора $F\alpha^*$:

$$F^+ = (UDV^T VDU^T)^{-1}UDV^T = UD^{-1}V^T = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j v_j^T;$$

$$\alpha^* = F^+ y = UD^{-1}V^T y = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y);$$

$$F\alpha^* = P_F y = (VDU^T)UD^{-1}V^T y = VV^T y = \sum_{j=1}^n v_j (v_j^T y);$$

$$\|\alpha^*\|^2 = \|D^{-1}V^T y\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j} (v_j^T y)^2.$$

Проблема мультиколлинеарности

Если имеются $\lambda_j \rightarrow 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо:
 $\|\alpha\| \rightarrow \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»:
 $Q(\alpha^*) = \|F\alpha^* - y\|^2 \rightarrow 0$;
- **мультиколлинеарность влечёт переобучение.**

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \rightarrow \min$;
- Преобразование признаков: $f_1, \dots, f_n \rightarrow g_1, \dots, g_m, m \ll n$;
- Отбор признаков: $f_1, \dots, f_n \rightarrow f_{j_1}, \dots, f_{j_m}, m \ll n$.

Регуляризация (гребневая регрессия)

Штраф за увеличение нормы вектора весов $\|\alpha\|$:

$$Q_\tau(\alpha) = \|F\alpha - y\|^2 + \frac{1}{2\sigma} \|\alpha\|^2,$$

где $\tau = \frac{1}{\sigma}$ — неотрицательный *параметр регуляризации*.

Вероятностная интерпретация: априорное распределение вектора α — гауссовское с ковариационной матрицей σI_n .

Модифицированное МНК-решение (τI_n — «гребень»):

$$\alpha_\tau^* = (F^T F + \tau I_n)^{-1} F^T y.$$

Преимущество сингулярного разложения:

можно подбирать параметр τ , вычислив SVD только один раз.

Регуляризованный МНК через сингулярное разложение

Вектор регуляризованного МНК-решения α_τ^*
 и МНК-аппроксимация целевого вектора $F\alpha_\tau^*$:

$$\alpha_\tau^* = U(D^2 + \tau I_n)^{-1} D V^T y = \sum_{j=1}^n \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^T y);$$

$$F\alpha_\tau^* = V D U^T \alpha_\tau^* = V \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^T y = \sum_{j=1}^n \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^T y);$$

$$\|\alpha_\tau^*\|^2 = \|D^2(D^2 + \tau I_n)^{-1} D^{-1} V^T y\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j + \tau} (v_j^T y)^2.$$

$F\alpha_\tau^* \neq F\alpha^*$, но зато решение становится гораздо устойчивее.

Выбор параметра регуляризации τ

Контрольная выборка: $X^k = (x'_i, y'_i)_{i=1}^k$;

$$F'_{k \times n} = \begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix}, \quad y'_{k \times 1} = \begin{pmatrix} y'_1 \\ \dots \\ y'_k \end{pmatrix}.$$

Вычисление функционала Q на контрольных данных T раз потребует $O(kn^2 + knT)$ операций:

$$Q(\alpha_\tau^*, X^k) = \|F' \alpha_\tau^* - y'\|^2 = \left\| \underbrace{F' U}_{k \times n} \operatorname{diag} \left(\frac{\sqrt{\lambda_j}}{\lambda_j + \tau} \right) \underbrace{V^T y}_{n \times 1} - y' \right\|^2.$$

Зависимость $Q(\tau)$ обычно имеет характерный минимум.

Регуляризация сокращает «эффективную размерность»

Сжатие (shrinkage) или сокращение весов (weight decay):

$$\|\alpha_\tau^*\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j + \tau} (v_j^\top y)^2 < \|\alpha^*\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j} (v_j^\top y)^2.$$

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

$$\text{tr } F(F^\top F)^{-1} F^\top = \text{tr}(F^\top F)^{-1} F^\top F = \text{tr } I_n = n.$$

При использовании регуляризации:

$$\text{tr } F(F^\top F + \tau I_n)^{-1} F^\top = \text{tr } \text{diag} \left(\frac{\lambda_j}{\lambda_j + \tau} \right) = \sum_{j=1}^n \frac{\lambda_j}{\lambda_j + \tau} < n.$$

Лассо Тибширани — другой подход к регуляризации LASSO — Least Absolute Shrinkage and Selection Operator

$$\begin{cases} Q(\alpha) = \|F\alpha - y\|^2 \rightarrow \min_{\alpha}; \\ \sum_{j=1}^n |\alpha_j| \leq \varkappa; \end{cases}$$

Лассо приводит к отбору признаков! Почему?

После замены переменных

$$\begin{cases} \alpha_j = \alpha_j^+ - \alpha_j^-; \\ |\alpha_j| = \alpha_j^+ + \alpha_j^-; \end{cases} \quad \alpha_j^+ \geq 0; \quad \alpha_j^- \geq 0.$$

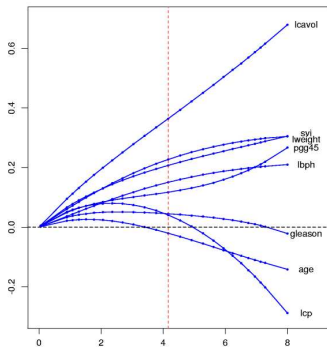
ограничения принимают канонический вид:

$$\sum_{j=1}^n \alpha_j^+ + \alpha_j^- \leq \varkappa; \quad \alpha_j^+ \geq 0; \quad \alpha_j^- \geq 0.$$

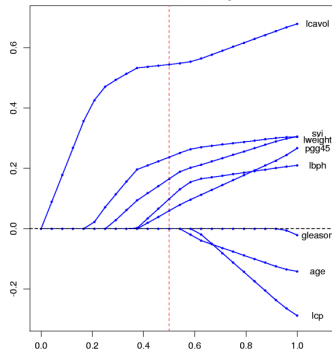
Чем меньше \varkappa , тем больше j таких, что $\alpha_j^+ = \alpha_j^- = 0$.

Сравнение гребневой регрессии и Лассо

Зависимость $\{\alpha_j\}$ от σ



Зависимость $\{\alpha_j\}$ от λ



Задача диагностики рака (prostate cancer, UCI)

T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. Springer, 2001.

Метод главных компонент: постановка задачи

$f_1(x), \dots, f_n(x)$ — исходные числовые признаки;
 $g_1(x), \dots, g_m(x)$ — новые числовые признаки, $m \leq n$;

Требование: старые признаки должны линейно
восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^m g_s(x) u_{js}, \quad j = 1, \dots, n, \quad \forall x \in X,$$

как можно точнее на обучающей выборке x_1, \dots, x_ℓ :

$$\sum_{i=1}^{\ell} \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 \rightarrow \min_{\{g_s(x_i)\}, \{u_{js}\}}$$

Матричные обозначения

Матрицы «объекты–признаки», старая и новая:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}; \quad G_{\ell \times m} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_\ell) & \dots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$U_{n \times m} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \quad \hat{F} = GU^T \stackrel{\text{ХОТИМ}}{\approx} F.$$

Найти: и новые признаки G , и преобразование U :

$$\sum_{i=1}^{\ell} \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 = \|GU^T - F\|^2 \rightarrow \min_{G,U},$$

Основная теорема метода главных компонент

Теорема

Если $m \leq \text{rk } F$, то минимум $\|GU^T - F\|^2$ достигается, когда столбцы U — это с.в. матрицы $F^T F$, соответствующие m максимальным с.з. $\lambda_1, \dots, \lambda_m$, а матрица $G = FU$.

При этом:

- 1 матрица U ортонормирована: $U^T U = I_m$;
- 2 матрица G ортогональна: $G^T G = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$;
- 3 $U\Lambda = F^T F U$; $G\Lambda = FF^T G$;
- 4 $\|GU^T - F\|^2 = \|F\|^2 - \text{tr } \Lambda = \sum_{j=m+1}^n \lambda_j$.

Связь с сингулярным разложением

Если взять $m = n$, то:

① $\|GU^T - F\|^2 = 0;$

② представление $\hat{F} = GU^T = F$ точное и совпадает с сингулярным разложением при $G = V\sqrt{\Lambda}$:

$$F = GU^T = V\sqrt{\Lambda}U^T; \quad U^TU = I_m; \quad V^TV = I_m.$$

③ линейное преобразование U работает в обе стороны:

$$F = GU^T; \quad G = FU.$$

Поскольку новые признаки некоррелированы ($G^TG = \Lambda$), преобразование U называется *декоррелирующим* (или преобразованием Карунена–Лоэва).

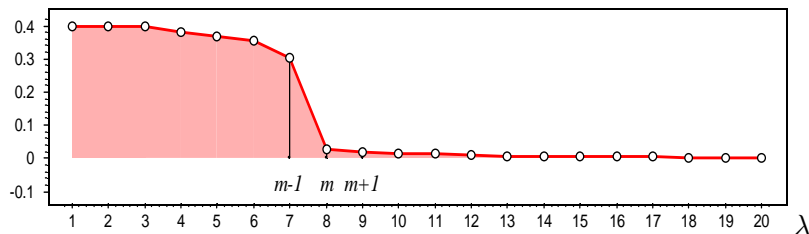
Эффективная размерность выборки

Упорядочим с.з. $F^T F$ по убыванию: $\lambda_1 \geq \dots \geq \lambda_n \geq 0$.

Эффективная размерность выборки — это наименьшее целое m , при котором

$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \leq \varepsilon.$$

Критерий «крутого склона»: находим m : $E_{m-1} \gg E_m$:



Решение задачи НК в новых признаках

Заменим F на её приближение GU^T :

$$\|G \underbrace{U^T \alpha}_{\beta} - y\|^2 = \|G\beta - y\|^2 \rightarrow \min_{\beta}.$$

Связь нового и старого вектора коэффициентов:

$$\alpha = U\beta; \quad \beta = U^T \alpha.$$

Решение задачи наименьших квадратов относительно β
(единственное отличие — m слагаемых вместо n):

$$\beta^* = D^{-1} V^T y = \sum_{j=1}^m \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y);$$

$$G\beta^* = VV^T y = \sum_{j=1}^m v_j (v_j^T y);$$

Нелинейная модель регрессии

Нелинейная модель регрессии $f(x, \alpha)$, $\alpha \in \mathbb{R}^p$.

Функционал среднеквадратичного отклонения:

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

Метод Ньютона–Рафсона.

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

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

$Q'(\alpha^t)$ — градиент функционала Q в точке α^t ,

$Q''(\alpha^t)$ — гессиан функционала Q в точке α^t ,

η_t — величина шага (можно полагать $\eta_t = 1$).

Метод Ньютона-Рафсона

Компоненты градиента:

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

Компоненты гессиана:

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

Не хотелось бы обращать гессиан на каждой итерации...

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

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

Метод Ньютона-Гаусса

Матричные обозначения:

$F_t = \left(\frac{\partial f}{\partial \alpha_j}(x_i, \alpha^t) \right)_{i=1, \ell}^{j=1, p}$ — $\ell \times p$ -матрица первых производных;

$f_t = (f(x_i, \alpha^t))_{i=1, \ell}$ — вектор значений f .

Формула t -й итерации метода Ньютона-Гаусса:

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

β — это решение задачи многомерной линейной регрессии

$$\|F_t \beta - (f^t - y)\|^2 \rightarrow \min_{\beta}.$$

Нелинейная регрессия сведена к серии линейных регрессий.

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

Логистическая регрессия (напоминание)

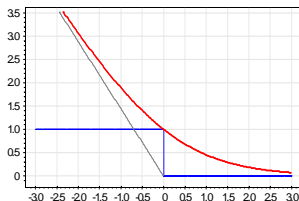
$Y = \{-1, +1\}$ — два класса, $x_i, w \in \mathbb{R}^n$.

Функционал аппроксимированного эмпирического риска:

$$Q(w) = - \sum_{i=1}^{\ell} \log \sigma(w^T x_i y_i) \rightarrow \min_w,$$

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

Логарифмическая функция потерь $\mathcal{L}(M_i) = \log(1 + e^{-M_i})$



$$M_i = w^T x_i y_i$$

Метода Ньютона-Рафсона

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

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

Элементы градиента — вектора первых производных $Q'(w^t)$:

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

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

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

где $\sigma_i = \sigma(y_i w^T x_i)$.

Матричные обозначения

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

$\Gamma_{\ell \times \ell} = \text{diag}(\sqrt{(1 - \sigma_i)\sigma_i})$ — диагональная матрица;

$\tilde{F} = \Gamma F$ — взвешенная матрица «объекты–признаки»;

$\tilde{y}_i = y_i \sqrt{(1 - \sigma_i)/\sigma_i}$, $\tilde{y} = (\tilde{y}_i)_{i=1}^{\ell}$ — взвешенный вектор ответов.

Тогда в методе Ньютона-Рафсона:

$$(Q''(w))^{-1} Q'(w) = -(F^T \Gamma^2 F)^{-1} F^T \Gamma \tilde{y} = -(\tilde{F}^T \tilde{F})^{-1} \tilde{F}^T \tilde{y} = -\tilde{F}^+ \tilde{y}.$$

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

$$Q(w) = \|\tilde{F}w - \tilde{y}\|^2 = \sum_{i=1}^{\ell} (1 - \sigma_i)\sigma_i \left(w^T x - \frac{y_i}{\sigma_i} \right)^2 \rightarrow \min_w.$$

Интерпретация

На каждом шаге метода Ньютона-Рафсона решается задача многомерной линейной регрессии:

$$Q(w) = \sum_{i=1}^{\ell} (1 - \sigma_i) \sigma_i \left(w^T x - \frac{y_i}{\sigma_i} \right)^2 \rightarrow \min_w.$$

Интерпретация:

- σ_i — вероятность правильного ответа на объекте x_i ;
- чем ближе x_i к границе, тем больше вес $(1 - \sigma_i) \sigma_i$;
- чем выше вероятность ошибки, тем больше y_i / σ_i .

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

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

Обобщение линейной модели регрессии

Пусть $\varphi_j: \mathbb{R} \rightarrow \mathbb{R}$ — некоторые нелинейные преобразования исходных признаков. Модель регрессии:

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

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

ИДЕЯ: будем по очереди уточнять функции φ_j по обучающей выборке $(f_j(x_i), z_i)_{i=1}^{\ell}$:

$$Q(\varphi_j, X^{\ell}) = \sum_{i=1}^{\ell} \left(\varphi_j(f_j(x_i)) - \underbrace{\left(y_i - \sum_{k=1, k \neq j}^n \varphi_k(f_k(x_i)) \right)}_{z_i = \text{const}(\varphi_j)} \right)^2 \rightarrow \min_{\varphi_j}.$$

Метод backfitting [Хасты, Тибширани, 1986]

Вход: F, y — матрица «объекты–признаки» и вектор ответов;

Выход: $\varphi_j(x)$ — все функции преобразования признаков.

1: нулевое приближение:

$\alpha :=$ решение задачи МЛР с признаками $f_j(x)$;

$\varphi_j(x) := \alpha_j f_j(x), j = 1, \dots, n$;

2: **повторять**

3: **для** $j = 1, \dots, n$

4: $z_i := y_i - \sum_{k=1, k \neq j}^n \varphi_k(f_k(x_i)), i = 1, \dots, \ell$;

5: $\varphi_j := \arg \min_{\varphi} \sum_{i=1}^{\ell} (\varphi(f_j(x)) - z_i)^2$; — одномерная регрессия

6: $Q_j := \sum_{i=1}^{\ell} (\varphi_j(f_j(x)) - z_i)^2$;

7: **пока** значения Q_j не стабилизируются

Резюме в конце лекции

- Непараметрическая регрессия — сглаживание функций
- Многомерная линейная регрессия — сингулярное разложение
- Гребневая регрессия — сингулярное разложение
- Метод главных компонент — ещё один способ избавления от близких к 0 сингулярных чисел
- Нелинейная регрессия — сводится к последовательности линейных