Обучение с учителем. Регрессия. Регуляризация

Романова Елизавета, Горбачук Анна, Сидоренко Денис

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Кафедра статистического моделирования

Санкт-Петербург
2019г.

- Y — количественный отклик, $X_1, \dots, X_p$ — предикторы, $X = (X_1, \dots, X_p)$.
- Предполагаем существование зависимости $Y = f^*(X) + \varepsilon$.
- ε — ошибка, которая не зависит от X, $\mathrm{E}\varepsilon = 0$.
- $x_1, \dots, x_n$ — наблюдения, $x_i = (x_{i1}, \dots, x_{ip})^{\mathrm{T}}$, y_i — отклик i-го наблюдения.
- $(x_i, y_i)_{i=1}^{n}$ — обучающая выборка, участвует в оценке f^*.
- $(x'_i, y'_i)_{i=1}^{k}$ — тестовая выборка, не участвует в оценке f^*.
- Задача: найти такую функцию $\hat{f}$, что $y \approx \hat{f}(x)$ для любого наблюдения (x, y).

Регрессия

- $X_n = (x_i, y_i)_{i=1}^n$ — обучающая выборка, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$.
- $y_i = f^*(x_i) + \varepsilon_i$, $i = 1, \ldots, n$.
- Модель регрессии: параметрическое семейство функций $f(x, \theta)$, где $\theta \in \Theta \subset \mathbb{R}^p$ — вектор параметров модели.
- Средняя квадратичная ошибка (функционал качества, наиболее часто применяющийся в задачах регрессии):

$$Q(\theta, X_n) = \sum_{i=1}^{n} (f(x_i, \theta) - y_i)^2.$$

- Задача обучения по МНК — задача минимизации

$$Q(\theta, X_n) \to \min_{\theta \in \Theta}.$$

Проблема

- Минимизируем среднюю квадратичную ошибку на обучающей выборке:
$$\mathrm{MSE}_{\text{train}} = \frac{1}{n} Q(B, X) \to \min.$$
- Истинная цель — минимизировать ошибку на всем пространстве объектов, то есть минимизировать $\mathrm{MSE}_{\text{test}} = Q(X', B)/k$, где $X' = (x'_i, y'_i)_{i=1}^k$ — произвольная контрольная выборка.
- Нет гарантии, что оптимальные для $\mathrm{MSE}_{\text{train}}$ параметры будут минимизировать $\mathrm{MSE}_{\text{test}}$.
- Может возникнуть переобучение (переподгонка): $\mathrm{MSE}_{\text{test}} \gg \mathrm{MSE}_{\text{train}}$.

Bias–variance tradeoff

Пусть $(x', y') \in X'_k$ — объект данных из тестовой выборки, $y' = f^*(x') + \varepsilon$, $\mathrm{E}\,\varepsilon = 0$, $\mathrm{E}\,\varepsilon^2 = \sigma^2$.
Для математического ожидания квадрата ошибки предсказания на (x', y') справедливо

$$\mathrm{E}(\hat{f}(x') - y')^2 = \mathrm{D}\,\hat{f}(x') + (\mathrm{Bias}\hat{f}(x'))^2 + \sigma^2,$$

$\mathrm{D}\,\hat{f}$ — дисперсия оценки $\hat{f}$, $\mathrm{Bias}\hat{f}(x')$ — смещение оценки, σ^2 — неустранимая ошибка.

- MSE на контрольной выборке зависит от дисперсии оценки и квадрата ее смещения.
- Дисперсия оценки определяет, насколько изменится $\hat{f}$, если бы мы получали эту оценку по другому набору данных.
- Смещение $\hat{f}$ характеризует ошибку, возникающую при аппроксимации сложной функции f^* более простой моделью.
- Нужен метод обучения, который обеспечивает и низкую дисперсию, и низкое смещение.

Множественная линейная регрессия

- Пусть зависимость между ответами и признаками линейна.
- Пусть ответы и признаки центрированы.
- Модель множественной линейной регрессии:

$$y_i = f(x_i, B) + \varepsilon_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

- Задача – минимизировать функционал качества:

$$Q(B, X) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \to \min_{\beta_1, \dots, \beta_p}.$$

Введем матричные обозначения:

- $\mathbb{X} = [X_1, \ldots, X_p]$, где $X_i = (x_{1i}, \ldots, x_{ni})^{\mathrm{T}}$, $i = 1, \ldots, p$,
- $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$,
- $B = (\beta_1, \ldots, \beta_p)$ — вектор параметров модели.

Модель линейной регрессии в матричной форме:

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \mathcal{E} = \mathbb{X}B + \mathcal{E}.$$

Задача оптимизации принимает вид:

$$Q(B, X) = ||Y - \mathbb{X}B||^2 \to \min_B .$$

Решение МНК: $\hat{B} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}Y = \mathbb{X}^{-}Y$, $\hat{Y} = \mathbb{X}\hat{B}$.

Решение МНК: $\hat{B} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}Y = \mathbb{X}^{-}Y$, $\hat{Y} = \mathbb{X}\hat{B}$.
При плохой обусловленности матрицы вычисление обратной матрицы крайне нежелательно. Варианты обхода:

- Решать соответствующую нормальную систему (например, при помощи QR-разложения)

$$\mathbb{X}^{\mathrm{T}}Y = \mathbb{X}^{\mathrm{T}}\mathbb{X}B.$$

- Использовать SVD. Пусть $\mathbb{X} = \mathbb{V}\Lambda\mathbb{U}^{\mathrm{T}}$ — сингулярное разложение $\mathbb{X}$. Тогда вектор МНК-решения легко записать в виде

$$\hat{B} = \mathbb{X}^{-}Y = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j (V_j^{\mathrm{T}} Y).$$

Пусть $\mathbb{X} = \mathbb{V}\Lambda\mathbb{U}^{\mathrm{T}}$ — сингулярное разложение $\mathbb{X}$.

- Тогда псевдообратную к $\mathbb{X}$ матрицу легко записать в виде

$$\mathbb{X}^{-} = \mathbb{U}\Lambda^{-1}\mathbb{V}^{\mathrm{T}} = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j V_j^{\mathrm{T}}.$$

- Вектор МНК-решения:

$$\hat{B} = \mathbb{X}^{-} Y = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j (V_j^{\mathrm{T}} Y).$$

- Оценка вектора Y:

$$\hat{Y} = \mathbb{X}\hat{B} = \sum_{j=1}^{p} V_j (V_j^{\mathrm{T}} Y).$$

- Норма вектора коэффициентов:

$$||\hat{B}||^2 = \sum_{j=1}^{p} \frac{1}{\lambda_j} (V_j^{\mathrm{T}} Y)^2.$$

Мультиколлинеарность признаков

Избыточность.
Рассмотрим случай, когда матрица данных содержит несколько сильно коррелированных признаков (есть $\lambda_j \to 0$). Что будет происходить в таком случае с МНК-оценкой:

- Решение $\hat{B}$ неустойчиво,
- Решение неинтерпретируемо, $||\hat{B}|| \to \infty$,
- Ответы на контрольной выборке неустойчивы,
- На обучающей выборке все хорошо:

$$||\mathbb{X}\hat{B} - Y||^2 \to 0.$$

Способы решения проблемы:

- Отбор признаков.
- Преобразование признаков.
- Регуляризация.

Регуляризация

- $\mathrm{MSE}_{\mathrm{test}}$ зависит от дисперсии оценки $\hat{f}$ и ее смещения.
- Когда связь между откликом и предикторами (почти) линейна, оценки по МНК обладают (почти) нулевым смещением, но при этом могут иметь большую дисперсию.
- Ковариационная матрица МНК-оценки $\hat{B}$:

$$\mathrm{Cov}(\hat{B}) = \sigma^2(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}.$$

- Чем больше дисперсия оценки $\hat{B}$, тем больше дисперсия $\hat{f}$.
- Когда матрица $\mathbb{X}$ близка к вырожденной, дисперсия $\hat{B}$ становится большой и $\mathrm{MSE}_{\mathrm{test}}$ увеличивается.
- При $p > n$ или при полностью коллинеарных признаках оценки по МНК не имеют уникального решения.
- Введение небольшого смещения в оценке может привести к уменьшению дисперсии и тем самым уменьшению $\mathrm{MSE}_{\text{test}}$.

Регуляризация Тихонова

- МНК решает нормальную систему $\mathbb{X}^{\mathrm{T}}\mathbb{X}B = \mathbb{X}^{\mathrm{T}}Y$.
- При наличии сильно скоррелированных признаков матрица $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ близка к вырожденной.
- **Регуляризация Тихонова**: прибавляем к матрице $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ матрицу $\mathbb{T}^{\mathrm{T}}\mathbb{T}$ так, чтобы их сумма была хорошо обусловлена.
- Переходим к нормальной системе $(\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})B = \mathbb{X}^{\mathrm{T}}Y$.
- Решение системы соответствует минимизации функции

$$||Y - \mathbb{X}B||_2^2 + ||\mathbb{T}B||_2^2.$$

- Решение: $\hat{B}_T = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1}\mathbb{X}^{\mathrm{T}}Y$.
- $\mathrm{E}\,\hat{B}_T = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X}B$.
- $\mathrm{Cov}(\hat{B}_T) = \sigma^2(\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X}((\mathbb{X}^{\mathrm{T}}\mathbb{X} + \mathbb{T}^{\mathrm{T}}\mathbb{T})^{-1})^{\mathrm{T}}$.

- Гребневая регрессия — это частный случай регуляризации Тихонова с $\mathbb{T} = \sqrt{\tau}\mathbb{I}$.
- Вводим штраф за увеличение нормы вектора B и переходим к минимизации следующей функции:

$$Q_\tau(B) = ||\mathbb{X}B - Y||^2 + \tau||B||^2 \to \min_B,$$

где τ — неотрицательный параметр регуляризации.

- В развернутом виде задача оптимизации записывается так:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \tau\sum_{j=1}^{p}\beta_j^2 \to \min_B .$$

- Решение:

$$\hat{B}_\tau = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \tau\mathbb{I}_p)^{-1}\mathbb{X}^{\mathrm{T}}Y.$$

Гребневая регрессия (Ridge regression)

Параметр регуляризации

Задача гребневой регрессии:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \tau\sum_{j=1}^{p}\beta_j^2 \to \min_B .$$

- $\tau\sum_{j=1}^{p}\beta_j^2$ мало, когда $\beta_1, \ldots, \beta_p$ близки к нулю.
- Чем больше коэффициент регуляризации τ, тем устойчивее решение, но больше смещение.
- Когда $\tau = 0$, то гребневая регрессия совпадает с обычной регрессией, но при $\tau \to \infty$ коэффициенты регрессии стремятся к нулю.
- Необходимо выбрать хорошее ("Компромиссное") значение τ.

- Решение задачи гребневой регрессии:

$$\hat{B}_\tau = (\mathbb{X}^{\mathrm{T}}\mathbb{X} + \tau\mathbb{I}_p)^{-1}\mathbb{X}^{\mathrm{T}}Y.$$

- Подход на основе сингулярного разложения $\mathbb{X} = \mathbb{V}\Lambda\mathbb{U}^{\mathrm{T}}$ позволяет подбирать параметр τ, вычислив SVD только один раз.
- Решение гребневой регрессии через SVD:

$$\hat{B}_\tau = \mathbb{U}(\Lambda^2 + \tau\mathbb{I}_p)^{-1}\Lambda\mathbb{V}^{\mathrm{T}}Y = \sum_{j=1}^{p} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} U_j (V_j^{\mathrm{T}} Y).$$

- Оценка функции f^* для выборки X через SVD:

$$\mathbb{X}\hat{B}_\tau = \mathbb{V}\Lambda\mathbb{U}^{\mathrm{T}}\hat{B}_\tau = \mathbb{V}\mathrm{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right)\mathbb{V}^{\mathrm{T}}Y = \sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + \tau} V_j (V_j^{\mathrm{T}} Y).$$

Подбор параметра τ

Скользящий контроль:

- выбираем сетку значений τ;
- вычисляем ошибку кросс-проверки для каждого значения τ;
- выбираем τ с наименьшим значением ошибки кросс-проверки;
- перестраиваем модель со всеми наблюдениями с выбранным значением τ.

Эвристика (практические рекомендации, Воронцов К. В.):

- брать τ в отрезке $[0.1, 0.4]$, если столбцы матрицы $\mathbb{X}$ заранее стандартизованы.
- выбрать τ так, чтобы число обусловленности матрицы $\mathbb{X}^{\mathrm{T}}\mathbb{X} + \tau\mathbb{I}_p$ приняло заданное не слишком большое значение M_0, откуда следует рекомендация $\tau \approx \lambda_{max}/M_0$.

- Модель: $Y = \mathbb{X}B + \mathcal{E}$.
- Ошибки независимы и $\varepsilon_i \in N(0, \sigma^2)$.
- Пусть выборка $(x_i, y_i)_{i=1}^n$ из распределения с плотностью p.
- Функция правдоподобия выборки (совместное распределение независимой выборки): $L(X, Y, B) = \prod_{i=1}^n p(x_i, y_i, B)$.
- Пусть вектор параметров B имеет априорное распределение $\pi(B)$.
- По теореме Байеса при фиксированном X апостериорное распределение $q(B|X, Y)$ пропорционально $L(X, Y, B)\pi(B)$.
- Если $\pi(B) \stackrel{d}{=} N(\mathbf{0}, \frac{\sigma^2}{\tau}\mathbb{I})$, то оценка апостериорного максимума B совпадает с решением гребневой регрессии.

Выкладки

Оценка максимума апостериорной вероятности:

$$\arg\max_{\beta_1,\dots,\beta_p} \exp\left(-\sum_{j=1}^{n}\frac{\varepsilon_j^2}{2\sigma^2}\right)\exp\left(-\sum_{i=1}^{p}\frac{\tau\beta_i^2}{2\sigma^2}\right) =$$

$$= \arg\max_{\beta_1,\dots,\beta_p} \exp\left(-\sum_{j=1}^{n}\frac{(y_j-\sum_{i=1}^{p}\beta_i x_{ij})^2}{\sigma^2}\right)\exp\exp\left(-\sum_{i=1}^{p}\frac{\tau\beta_i^2}{2\sigma^2}\right) =$$

$$= \arg\max_{B}\exp\left(-\frac{||Y-\mathbb{X}B||^2}{2\sigma^2}-\frac{\tau||B||^2}{2\sigma^2}\right) =$$

$$= \arg\min_{B}(||Y-\mathbb{X}B||^2+\tau||B||^2).$$

Пришли к решению задачи гребневой регрессией с параметром регуляризации τ.

- Стандартные МНК-оценки инварианты относительно умножения признака на константу, то есть значение $X_j\hat{\beta}_j$ не зависит от масштаба j-го признака.
- Оценки МНК гребневой регрессии не обладают свойством инвариантности и могут существенно меняться.

Вывод: гребневую регрессию нужно использовать после стандартизации признаков.

Проблемы:

- в конечную модель входят все начальные признаки;
- если признаков много, то усложняется интерпретация.

- Рассмотрим метод, в котором в качестве штрафа за увеличение нормы вектора B используется его l_1-норма.
- Метод LASSO решает следующую задачу минимизации:

$$||\mathbb{X}B - Y||_2^2 + \tau||B||_1^2 \to \min_B,$$

где τ — неотрицательный параметр регуляризации.

- Задача оптимизации в развернутом виде:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \tau\sum_{j=1}^{p}|\beta_j| \to \min_{\beta_1,\dots,\beta_p}.$$

- Хотим применить теорему Куна-Таккера.
- Проблема: целевая функция не гладкая.

Теорема Куна-Таккера

Пусть $x \in \mathbb{R}^n$. Рассмотрим задачу

$$f(x) \to \min,$$
$$g_i(x) \leq 0, \quad i = 0, \ldots, m.$$

Функция Лагранжа:

$$\mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x).$$

Теорема Куна-Таккера

Пусть $f(x)$ выпукла и дифференцируема на допустимом множестве. Все ограничения регулярные (аффинные функции). Тогда x_* — оптимальное решение, тогда и только тогда, когда $\exists, \lambda_i$ такие, что

$$\frac{\partial f(x_*)}{\partial x_j} + \sum_{i=1}^{m} \lambda_i \frac{\partial g(x_*)}{\partial x_j} = 0 \quad j = 1, \ldots, n,$$
$$g_i(x_*) \leq 0, \quad \lambda_i \geq 0, \quad \lambda_i g_i(x_*) = 0, \quad i = 1, \ldots, m.$$

Решение задачи оптимизации

- Задачу lasso-оптимизации можно переписать в форме с ограничениями:

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \to \min_{\beta_1, \dots, \beta_p}, \\ \sum_{j=1}^{p} |\beta_j| \le æ, \end{cases}$$

где $æ = 1/\tau$.

- После замены $\beta_j = \beta_j^+ - \beta_j^-$, $|\beta_j| = \beta_j^+ + \beta_j^-$, переходим к задаче оптимизации ($2p$ переменных, $2p + 1$ ограничений):

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} (\beta_j^+ - \beta_j^-) x_{ij} \right)^2 \to \min_{\beta_1^+, \dots, \beta_p^+, \beta_1^-, \dots, \beta_p^-}, \\ \sum_{j=1}^{p} \beta_j^+ + \beta_j^- \le æ, \quad \beta_j^+ \ge 0, \quad \beta_j^- \ge 0. \end{cases}$$

- Выпуклая задача квадратичного программирования с линейными ограничениями-неравенствами.
- Чем меньше $æ$, тем больше j таких, что $\beta_j^+ = \beta_j^- = 0$.

Вероятностная интерпретация Лассо

- Модель: $Y = \mathbb{X}B + \mathcal{E}$.
- Ошибки независимы и $\varepsilon_i \in N(0, \sigma^2)$.
- Пусть выборка $(x_i, y_i)_{i=1}^n$ из распределения с плотностью p.
- Функция правдоподобия выборки (совместное распределение независимой выборки): $L(X, Y, B) = \prod_{i=1}^n p(x_i, y_i, B)$.
- Пусть вектор параметров B имеет априорное распределение $\pi(B)$.
- По теореме Байеса при фиксированном X апостериорное распределение $q(B|X, Y)$ пропорционально $L(X, Y, B)\pi(B)$.
- Если $\pi(B) = \prod_{j=1}^p g(\beta_j)$, где g — плотность распределения Лапласа $\mathrm{Laplace}(0, \tau)$, то оценка апостериорного максимума B совпадает с решением лассо регрессии.

Сравнение гребневой регрессии и Лассо

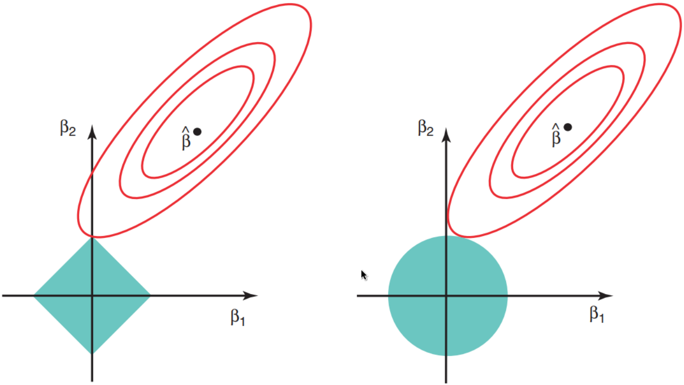

Рис.: Линии уровня квадратичной функции $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2$ и границы ограничений для Лассо (слева) и гребневой регрессии (справа) при $p = 2$.

Сравнение гребневой регрессии и Лассо

- Обычно Лассо подходит лучше в случае наличия большого количества лишних (незначимых) признаков,
- Для реальных данных обычно заранее не известно количество признаков, значимо влияющих на зависимую переменную,
- С помощью кросс-валидации можно определить какой подход лучше для конкретных данных.

Elastic net regularization

Решается задача оптимизации

$$||Y - \mathbb{X}B||_2^2 + \tau_1||B||_1^2 + \tau_2||B||_2^2 \to \min_B .$$

- ✓ Elastic net — это комбинация методов Lasso и Ridge:
  - Когда $\tau_1 = 0$: Ridge регрессия;
  - Когда $\tau_2 = 0$: Lasso регрессия;
- ✓ Elastic net в целом лучше, чем Lasso при наличии коррелированных признаков;
- ✓ В отличие от Ridge регрессии, когда $p > n$, Elastic net может учитывать более n переменных;
- ✓ При наличии группы релевантных и избыточных признаков Lasso обычно имеет тенденцию отказываться от всех, кроме одного признака из этой группы, в то время как Elastic net будет выбирать всю группу признаков.
- ✓ Elastic net можно свести к SVM, для которого разработано много быстрых решений.

Нелинейная модель регрессии $f(x,\theta)$, $\theta \in \mathbb{R}^k$. Решаем задачу минимизации функционала среднеквадратичного отклонения:

$$Q(\theta, X) = \sum_{i=1}^{n} (f(x_i, \theta) - y_i)^2$$

Метод Ньютона-Рафсона:

- Начальное приближение: $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)$,
- Итерационный процесс:

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

$Q'(\theta^t)$ — градиент, $Q''(\theta^t)$ — гессиан, h_t — величина шага (простейший вариант: $h_t = 1$).

Метод Ньютона-Рафсона

- Компоненты градиента:

$$\frac{\partial Q(\theta)}{\partial \theta_j} = 2\sum_{i=1}^{n}(f(x_i,\theta) - y_i)\frac{\partial f(x_i,\theta)}{\partial \theta_j}.$$

- Компоненты гессиана:

$$\frac{\partial^2 Q(\theta)}{\partial \theta_j \partial \theta_k} = 2\sum_{i=1}^{n}\frac{\partial f(x_i,\theta)}{\partial \theta_j}\frac{\partial f(x_i,\theta)}{\partial \theta_k} - 2\sum_{i=1}^{n}(f(x_i,\theta) - y_i)\frac{\partial^2 f(x_i,\theta)}{\partial \theta_j \partial \theta_k}.$$

- Линеаризация $f(x_i,\theta)$ в окрестности θ^t:

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

Метод Ньютона-Гаусса

Введем обозначения:

- $\mathbb{F}_t = (\frac{\partial f}{\partial \theta_j}(x_i, \theta^t))_{n \times p}$ — матрица первых производных,
- $f_t = (f(x_i, \theta^t)_{n \times 1}$ — вектор значений f.

Итерация метода Ньютона-Гаусса:

$$\theta^{t+1} := \theta^t - h_t \underbrace{(\mathbb{F}_t^{\mathrm{T}} \mathbb{F}_t)^{-1} \mathbb{F}_t^{\mathrm{T}} (f_t - Y)}_{\widetilde{B}}.$$

$\widetilde{B}$ — решение задачи множественной линейной регрессии

$$||\mathbb{F}_t B - (f_t - Y)||^2 \to \min_B.$$

Нелинейная регрессия сводится к серии линейных регрессий.