Линейные модели

Методы анализа данных

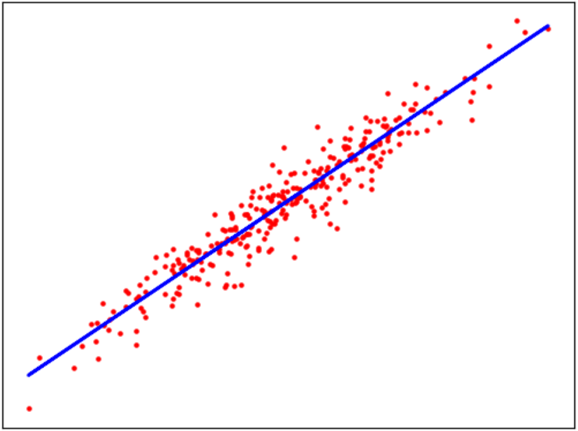
Лекция 2

Москва, МФТИ, 2020

План

- Линейные модели
- Обучение модели линейной регрессии
- Градиентный спуск для линейной регрессии
- Стохастический градиентный спуск
- Метрики для линейной регрессии
- Переобучение
- Метод максимизации правдоподобия
- Регуляризация
- Общие советы

Линейная зависимость



Линейная модель

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x^j \tag{1}$$

Обозначения:

- w_0 - свободный коэффициент
- x^j - признаки
- w_j - веса признаков

$$a(x) = \sum_{j=1}^{d+1} w_j x^j = w \cdot x \tag{2}$$

Функция потерь

Возможные меры ошибки:

- $|a(x) - y|$
- $(a(x) - y)^2$

$$Q(a, x) = \frac{1}{\ell} \sum_{j=1}^{\ell} (a(x_i) - y_i)^2 \tag{3}$$

$$Q(w, x) = \frac{1}{\ell} \sum_{j=1}^{\ell} (w \cdot x_i - y_i)^2 \tag{4}$$

Обучение линейной регрессии

Задача обучения линейной регрессии сводится к задаче минимизации ошибки:

$$\min_{w} \frac{1}{\ell} \sum_{j=1}^{\ell} (w \cdot x_i - y_i)^2 \tag{5}$$

Переход к матричной форме записи

$$X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{l1} & \dots & x_{ld} \end{pmatrix} \tag{6}$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix} \tag{7}$$

$$Q(\omega, X) = \frac{1}{l} ||X\omega - y||^2 \to \min_{\omega} \tag{8}$$

Недостатки аналитического метода решения оптимизационной задачи

Аналитическое решение:

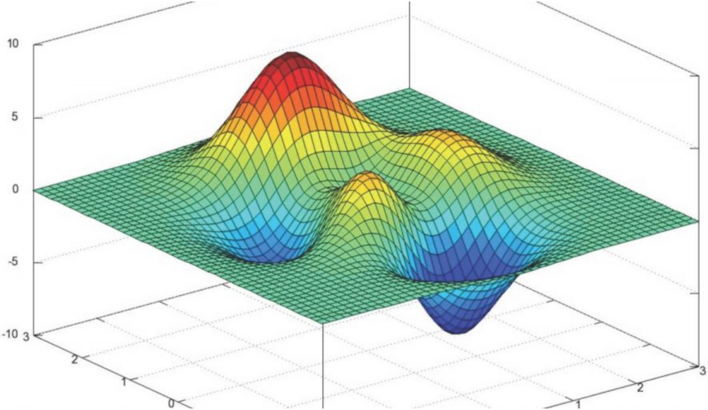
$$w_* = (X^T X)^{-1} X^T y \tag{9}$$

- Алгоритмическая сложность вычислений обратной матрицы
- Возможная плохая обусловленность матрицы, от которой нужно взять обратную

Оптимизационный подход

- В соответствии с вышеуказанными недостатками способом решения задачи является оптимизационный подход
- Одним из основных методов оптимизации, используемых в машинном обучении является метод стохастического градиентного спуска(SGD).

Возможный вид функции потерь



Градиентный спуск

$$Q(w, x) = \min_{w} \frac{1}{\ell} ||Xw - y||^2 \tag{10}$$

Шаг обычного градиентного спуска:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1}, x) \tag{11}$$

Градиент функции потерь:

$$\nabla_w Q(w, x) = \frac{2}{\ell} X^T (Xw - y) \tag{12}$$

Частная производная по j-компоненте:

$$\frac{\partial Q}{\partial w_j} = \frac{2}{\ell} \sum_{i=1}^{\ell} (x_i^j < w \cdot x_i > - y_i) \tag{13}$$

Критерий останова:

$$||w^t - w^{t-1}|| < \epsilon \tag{14}$$

- Как мы увидели на предыдущем слайде для вычисления частной производной по j-компоненте функции потерь, необходимо вычислять сумму по всем объектам обучающей выборки. Это существенно замедляет процесс обучения.
- В стохастическом методе градиентного спуска градиент функции качества вычисляется только на одном случайно выбранном объекте обучающей выборки:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1}, x_i) \tag{15}$$

Преимущества SGD

- Время работы
- Эффективное использование памяти
- Применимость в онлайн-обучении

Метрики качества

- Среднеквадратическая ошибка:

$$\text{MSE}(a, X) = \frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2 \tag{16}$$

- Средняя абсолютная ошибка:

$$\text{MAE}(a, X) = \frac{1}{l} \sum_{i=1}^{l} |a(x_i) - y_i| \tag{17}$$

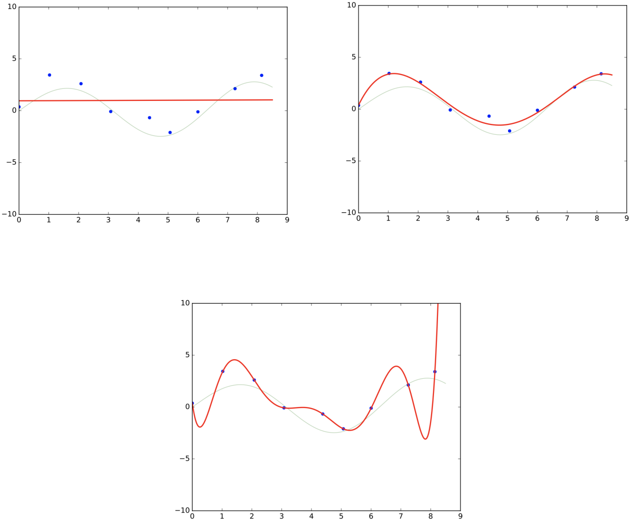
- Коэффициент детерминации:

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{l} (a(x_i) - y_i)^2}{\sum_{i=1}^{l} (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{l} \sum_{i=1}^{l} y_i \tag{18}$$

Переобучение и неодообучение

- Недообучение — ситуация, когда алгоритм плохо описывает и обучающую выборку, и новые данные
- Переобучение - ситуация, когда алгоритм хорошо описывает обучающую выборку, а новые данные плохо

Иллюстрация проблемы переобучения



Симптомы переобучения

- Большие значения весов:

$$a(x) = \omega_0 + \omega_1 x + \omega_2 x^2 + \ldots + \omega_9 x^9 \tag{19}$$

$$a(x) = 0.5 + 12458922x + 43983740x^2 + \ldots + 2740x^9 \tag{20}$$

- Мультиколлинеарность признаков:

$$\alpha_1 x_i^1 + \ldots + \alpha_d x_i^d = 0 \tag{21}$$

$$\langle \alpha, x_i \rangle = 0 \tag{22}$$

$$\omega_* = \mathrm{argmin}_\omega \frac{1}{l} \sum_{i=1}^{l} (\langle \omega, x_i \rangle - y_i)^2 \tag{23}$$

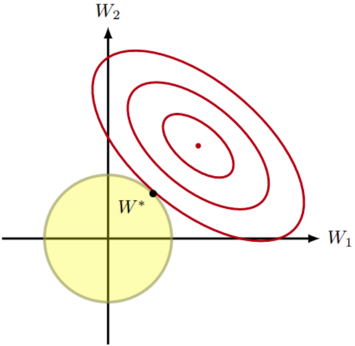
$$\omega_1 = \omega_* + t\alpha \tag{24}$$

$$\langle \omega_* + t\alpha, x \rangle = \langle \omega_*, x \rangle + t\langle \alpha, x \rangle = \langle \omega_*, x \rangle \tag{25}$$

Регуляризация

$$Q(\omega, X) + \lambda||\omega||^2 \to \min_{\omega} \quad (26)$$

$$\begin{cases} Q(\omega, X) \to \min_{\omega} \\ ||\omega||^2 \leq C \end{cases} \quad (27)$$



Выявление проблемы переобучения

- Отложенная выборка
- Кросс-валидация

Борьба с переобучением

- Упрощение модели
- Регуляризация модели

Метод максимизации правдоподобия

- Пусть есть случайная величина X:

$$X \sim F(x, \theta), \quad X^n = (X_1, \dots, X_n) \tag{28}$$

- Функция правдоподобия:

$$L(X^n, \lambda) = \prod_{i=1}^{n} P(X = X_i, \theta) \tag{29}$$

$$\ln L(X^n, \lambda) = \sum_{i=1}^{n} \ln P(X = X_i, \theta) \tag{30}$$

- Оценка максимального правдоподобия

$$\hat{\lambda}_{\text{ОМП}} = \text{argmax}_{\lambda} \ln L(X^N, \lambda) \tag{31}$$

$$X \sim F(x, \theta), L(X^n, \lambda) = \prod_{i=1}^{n} f(X = X_i, \theta), \quad \hat{\lambda}_{\text{ОМП}} = \text{argmax}_{\lambda} L(X^N, \lambda)$$

Свойство метода максимизации правдоподобия

- Получаемые оценки при увеличении объема выборки начинают стремиться к истинным значениям:

$$\hat{\lambda}_{\text{ОМП}} \to \theta \quad \text{при} \quad n \to \infty \tag{32}$$

- С ростом объема выборки, оценки максимального правдоподобия все лучше описываются нормальным распределением с средним, равным истинному значению параметра, и дисперсией, равной величине, обратной к информации Фишера

$$\hat{\lambda}_{\text{ОМП}} \sim N(\theta, I^{-1}(\theta)) \quad \text{при} \quad n \to \infty \tag{33}$$

Модель шума: нормальное распределение

- Значение отклика:

$$y = a(x) + \varepsilon \tag{34}$$

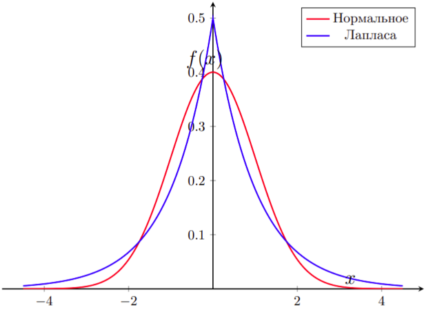
- Если этот случайный шум имеет нормальное распределение с нулевым средним, оказывается, что задача минимизации среднеквадратичной ошибки дает оценку максимального правдоподобия для регрессионной функции:

$$a_* = \mathrm{argmin}_a \frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2 \tag{35}$$

Модель шума: распределение Лапласа

$$f(x) = \frac{1}{\alpha} e^{-\alpha|x|} \tag{36}$$

$$a_* = \text{argmin}_a \frac{1}{l} \sum_{i=1}^{l} |a(x_i) - y_i| \tag{37}$$



Регуляризация

- L_2-регуляризации

$$\omega_* = \mathrm{argmin}_\omega \left(\frac{1}{l} \sum_{i=1}^{l} (\langle \omega, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} \omega_j^2 \right) \qquad (38)$$

- L_1-регуляризации

$$\omega_* = \mathrm{argmin}_\omega \left(\frac{1}{l} \sum_{i=1}^{l} (\langle \omega, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} |\omega_j| \right) \qquad (39)$$

Сравнение L-1 и L-2 регуляризаторов

- Пусть матрица признаков является единичной:

$$X = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{40}$$

- В случае отсутвия регуляризаторов решение оптимизационной задачи выглядит следующим образом:

$$\omega_* = \mathrm{argmin}_\omega \sum_{i=1}^{l} (\omega_i - y_i)^2 \tag{41}$$

- В результате получится следующий вектор весов:

$$\omega_{*j} = y_j \tag{42}$$

Вид вектора весов в решении пробного примера

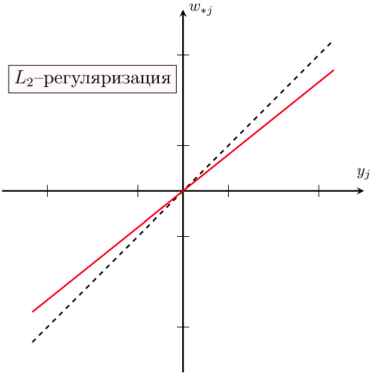
- Для L-2 регуляризатора:

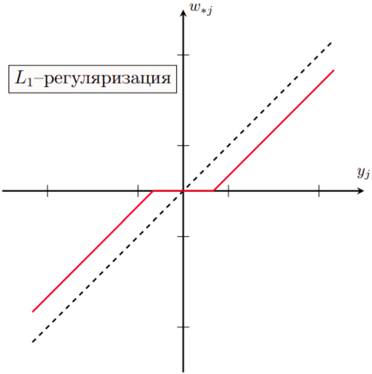
$$\omega_{*j} = \frac{y_j}{1 + \lambda} \tag{43}$$

- Для L-1 регуляризатора:

$$\omega_{*j} = \begin{cases} y_j - \lambda/2, & y_j > \lambda/2 \\ y_j + \lambda/2, & y_j < -\lambda/2 \\ 0, & |y_j| \leq \lambda/2 \end{cases} \tag{44}$$

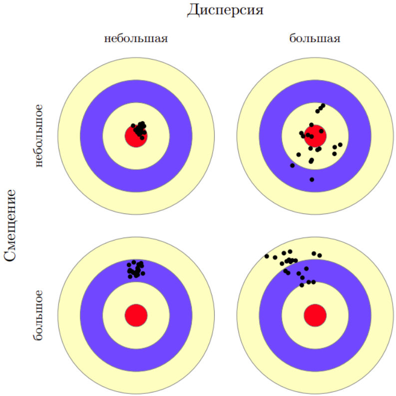
Регуляризация





Смещение и дисперсия

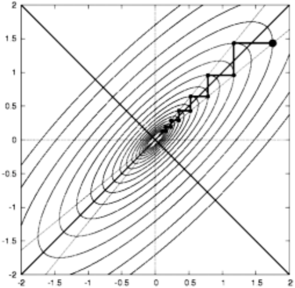
$$\mathbb{E}(a_*(x) - y)^2 = \underbrace{(\mathbb{E}a_*(x) - a(x))^2}_{\text{Квадрат смещения}} + \underbrace{Da_*(x)}_{\text{Дисперсия оценки}} + \underbrace{\sigma^2}_{\text{Шум}}$$



Масштабирование признаков

$$\omega_1^2 + \omega_2^2 \to \min_{\omega} \tag{45}$$

$$\omega_1^2 + 100\omega_2^2 \to \min_{\omega} \tag{46}$$



Масштабирование выборки

- Стандартизация

$$\mu_j = \frac{1}{l}\sum_{i=1}^{l} x_i^j, \quad \sigma_j = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(x_i^j - \mu_j)^2} \tag{47}$$

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j} \tag{48}$$

- Масшатабирование на отрезок [0:1]

$$m_j = \min(x_1^j, \ldots, x_l^j), \quad M_j = \max(x_1^j, \ldots, x_l^j) \tag{49}$$

$$x_i^j := \frac{x_i^j - m_j}{M_j - m_j} \tag{50}$$

Практические советы

- Масштабирование признаков
- Отбор признаков
- Контоль перееобучения