Лекция 3
Линейные модели в задаче регрессии

Габдуллин Р.А., Макаренко В.А.

МГУ им. М.В. Ломоносова

2 февраля 2021

Обучение с учителем

X – множество объектов,
Y – множество ответов,
$y : X \to Y$ – неизвестная зависимость.

Дано:
$\{x_1, x_2, \dots, x_\ell\} \subset X$ – обучающая выборка,
$y_i = y(x_i),\ i = 1, \dots, \ell$ – известные ответы.

Найти:
$a : X \to Y$ – решающая функция, приближающая y на всём X.

Описание объектов. Признаки

X – множество объектов,
$f_j : X \to F_j, \quad j = 1, \dots, n$ – признаки объектов (features),

Типы признаков:

Бинарные	Binary	$F_j = \{\text{true, false}\}$
Номинальные	Categorical	F_j – конечное мн-во
Порядковые	Ordinal	F_j – конечное упорядоченное мн-во
Количественные	Numerical	$F_j = \mathbb{R}$

$(f_1(x), f_2(x), \dots, f_n(x))$ – признаковое описание объекта $x \in X$.
Матрица «объекты-признаки» (feature data)

$$F = \|f_j(x_i)\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

Задача восстановления регрессии

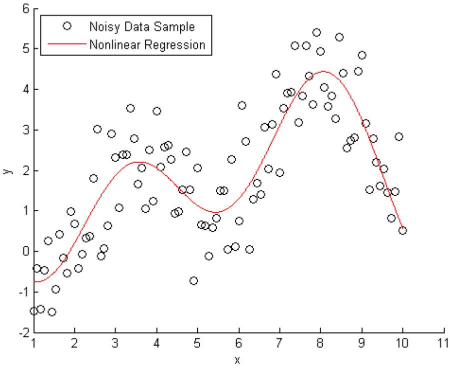

Рис.: Источник: datascience.stackexchange.com

- Вещественный ответ: $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$

Линейная модель регрессии

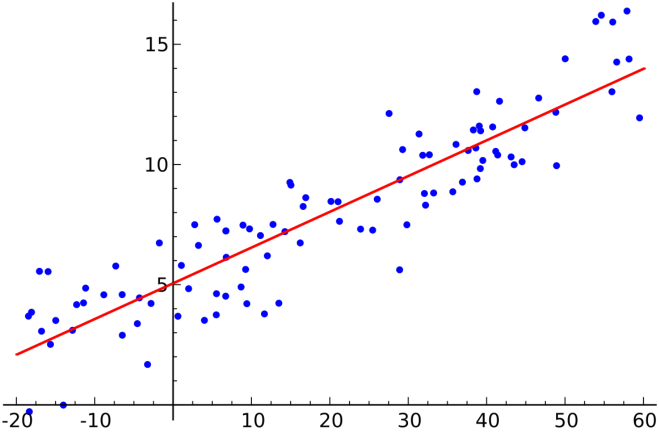

Рис.: Источник: Википедия

Линейная модель регрессии

- Семейство алгоритмов:

$$A = \{a(x, \theta) | \theta \in \mathbb{R}^{n+1}\},$$

$$a(x, \theta) = \theta_0 + \sum_{j=1}^{n} \theta_j f_j(x) = \sum_{j=0}^{n} \theta_j f_j(x),$$

если положить $f_0(x) \equiv 1$.

- Эмпирический риск:

$$Q(\theta, \mathbb{X}) = \sum_{i=1}^{\ell} w_i \cdot \Big(y_i - a(x_i, \theta)\Big)^2,$$

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

- Метод наименьших квадратов (МНК):

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, Q(\theta, \mathbb{X}).$$

Метод максимального правдоподобия

- Вероятностная модель:

$$y_i = a(x_i, \theta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2),$$

где $\{\varepsilon_i\}$ – независимые нормальные случайные величины.

- Функция правдоподобия ответов:

$$L(y_1, \dots, y_\ell | \theta) = \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{(y_i - a(x_i, \theta))^2}{2\sigma_i^2}\right).$$

- Метод максимального правдоподобия:

$$L(y_1, \dots, y_\ell | \theta) \to \max_\theta \iff -\ln L(y_1, \dots, y_\ell | \theta) \to \min_\theta,$$

$$\ln L(y_1, \dots, y_\ell | \theta) = \text{const} + \frac{1}{2} \cdot \sum_{i=1}^{\ell} \frac{(y_i - a(x_i, \theta))^2}{\sigma_i^2},$$

$$\sum_{i=1}^{\ell} w_i \cdot (y_i - a(x_i, \theta))^2 \to \min_\theta, \quad w_i = \frac{1}{\sigma_i^2}.$$

Аналитическое решение

- Многомерная линейная регрессия:

$$a(x,\theta) = \sum_{j=0}^{n} \theta_j f_j(x).$$

- Матричная запись:

$$F = \begin{pmatrix} f_0(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_0(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad y = \begin{pmatrix} y_1, \\ \dots, \\ y_\ell \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0, \\ \dots, \\ \theta_n \end{pmatrix},$$

$$y = F\theta,$$

$$Q(\theta, \mathbb{X}) = \|y - F\theta\|^2 \to \min_\theta .$$

Аналитическое решение

- Многомерная линейная регрессия:

$$a(x_i,\theta) = \sum_{j=0}^{n} \theta_j F_{ij}.$$

- Необходимое условие минимума:

$$\frac{\partial Q}{\partial \theta_j} = -2\sum_{i=1}^{\ell}(y_i - a(x_i,\theta))\cdot\frac{\partial a(x_i,\theta)}{\partial \theta_j} = -2\sum_{i=1}^{\ell}(y_i - a(x_i,\theta))\cdot F_{ij} = 0,$$

то есть

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

- Нормальная система уравнений:

$$F^T F\theta = F^T y.$$

- Решение нормальной системы уравнений:

$$\theta = (F^T F)^{-1} F^T y.$$

Численное решение

Градиентный спуск:

- Выбрать начальное приближение $\theta(0)$.
- Шаг в сторону антиградиента:

$$\theta(i+1) = \theta(i) - \alpha(i) \cdot \frac{\partial Q}{\partial \theta}\Big|_{\theta=\theta(i)} = \theta(i) - \alpha(i) \cdot 2F^T(F\theta(i) - y).$$

- Повторять до сходимости.

Варианты:

- Классический градиентный спуск: на каждой итерации делаем шаг в сторону антиградиента эмпирического риска по всей выборке.
- Стохастический градиентный спуск: на каждой итерации делаем шаг в сторону антиградиента эмпирического риска по одному наблюдению.
- Mini-batch градиентный спуск: на каждой итерации делаем шаг в сторону антиградиента эмпирического риска по части выборки.

Численное решение

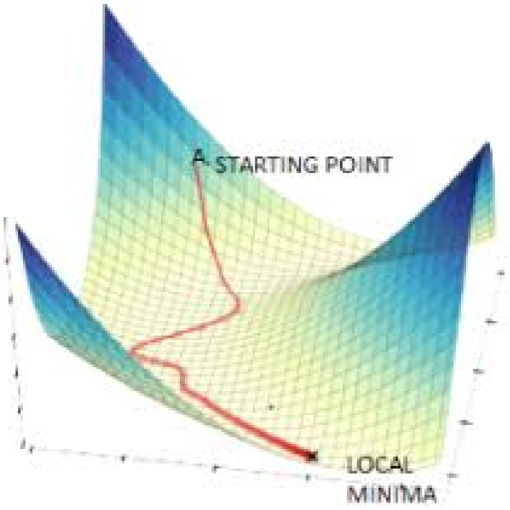

Рис.: Источник: datasciencecentral.com

Свойства оценок по МНК

Теорема (Гаусса-Маркова)

Пусть выполнены следующие условия:

- $y_i = a(x_i, \theta) + \varepsilon_i$.
- $\mathrm{rank}(F) = n + 1$.
- $\mathbb{E}\varepsilon_i = 0$.
- $\mathrm{cov}(\varepsilon) = \sigma^2 I$.

Тогда

$$\theta^* = \sum_{i=1}^{\ell} \cdot \Big(y_i - a(x_i, \theta)\Big)^2$$

являеся оптимальной оценкой в классе линейных оценок.

Проблема мультиколлинеарности

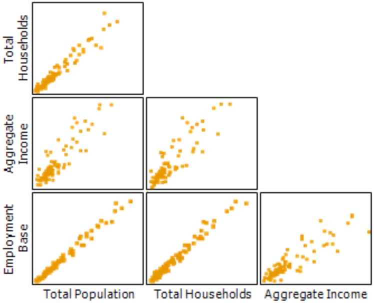

Рис.: Источник: medium.com

- Два или более признаков почти линейно зависимы.
- Решение получается неустойчивым.
- Неустойчивое решение ведет к переобучению.

Гребневая (Ridge) регрессия (L_2-регуляризация)

- Эмпирический риск:

$$Q(\theta, \mathbb{X}) = \|y - F\theta\|^2 + \lambda\|\theta\|^2, \quad \lambda > 0.$$

- Необходимое условие минимума:

$$\frac{\partial Q}{\partial \theta} = 2F^T(F\theta - y) + 2\lambda\theta = 2((F^TF + \lambda I)\theta - F^Ty) = 0,$$

где I – единичная матрица.

- Решение в явном виде:

$$\theta = (F^TF + \lambda I)^{-1}F^Ty.$$

Вероятностная интерпретация гребневой регрессии

- Вероятностная модель:

$$\theta \sim \mathcal{N}(0, \tau^2 I),$$

$$y_i = a(x_i, \theta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

где $\{\varepsilon_i\}$ – независимые нормальные случайные величины.

- Апостериорное распределение весов:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

- Максимум апостериорного распределения:

$$p(\theta|y) \to \max_{\theta} \iff -\ln p(y|\theta)p(\theta) \to \min_{\theta},$$

$$-\ln p(y|\theta)p(\theta) = \text{const} + \frac{1}{2} \cdot \left(\sum_{i=1}^{\ell} \frac{(y_i - a(x_i, \theta))^2}{\sigma^2} + \sum_{j=0}^{n} \frac{\theta_j^2}{\tau^2} \right),$$

то есть $\lambda = \frac{\sigma^2}{\tau^2}$.

Lasso-регрессия (L_1-регуляризация)

- Эмпирический риск:

$$Q(\theta, \mathbb{X}) = \|y - F\theta\|^2 + \lambda \sum_{j=0}^{n} |\theta_j|, \quad \lambda > 0.$$

- Вероятностная интерпретация: веса независимы и имеют одно и то же распределение Лапласа.

Ridge vs. Lasso

Эквивалентные задачи поиска условного минимума.

- Для *Ridge*-регрессии (L_2):

$$\|y - F\theta\|^2 \to \min_{\theta}, \quad \sum_{j=0}^{n} \theta_j^2 \leqslant \varkappa_1.$$

- Для *Lasso*-регрессси (L_1):

$$\|y - F\theta\|^2 \to \min_{\theta}, \quad \sum_{j=0}^{n} |\theta_j| \leqslant \varkappa_2.$$

Ridge vs. Lasso

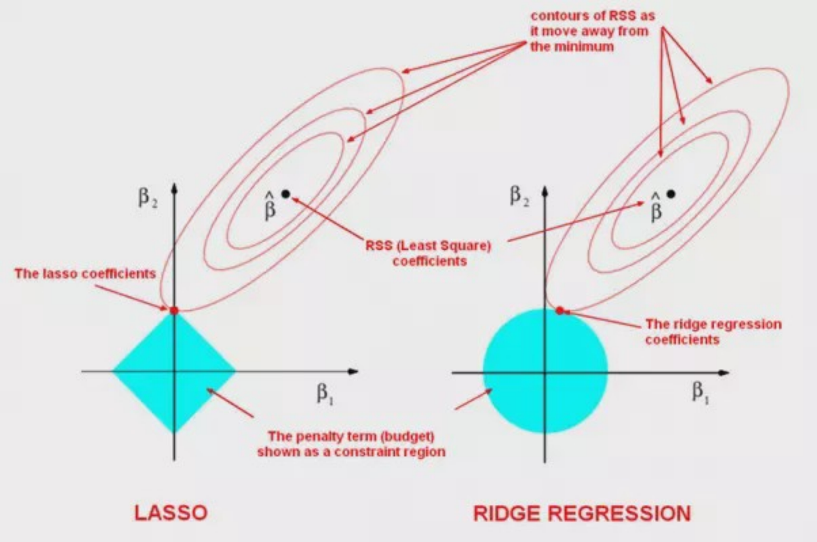

Рис.: Источник: medium.com

Ridge vs. Lasso

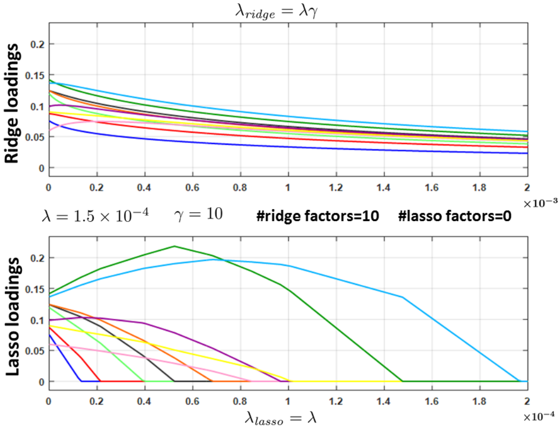

Рис.: Источник: arpm.co

Масштабирование вещественных признаков

Веса модели чувствительны к сдвиг-масштабным преобразованиям признаков:

$$\widetilde{f}_j(x) = \alpha_j + \beta_j f_j(x),$$

$$\theta_0 + \sum_{j=1}^{n} \theta_j \cdot \frac{\widetilde{f}_j(x) - \alpha_j}{\beta_j} = \left(\theta_0 - \sum_{j=1}^{n} \frac{\alpha_j \theta_j}{\beta_j}\right) + \sum_{j=1}^{n} \frac{\theta_j}{\beta_j} \cdot \widetilde{f}_j(x) =$$

$$= \widetilde{\theta}_0 + \sum_{j=1}^{n} \widetilde{\theta}_j . \widetilde{f}_j(x).$$

Часто признаки преобразовывают, приводя их к единой шкале:

$$\widetilde{f}_j(x) = \frac{f_j(x) - \mathbb{E} f_j(x)}{\sqrt{\mathbb{D} f_j(x)}} \quad \text{или} \quad \widetilde{f}_j(x) = \frac{f_j(x) - \min\limits_i f_j(x_i)}{\max\limits_i f_j(x_i) - \min\limits_i f_j(x_i)}.$$

Преобразование категориальных признаков

Пусть признак f_j может принимать одно из K возможных значений:

$$f_j(x) \in \{1, \ldots, K\}.$$

One-hot кодирование. Признак f_j «разбивается» на $K-1$ признаков:

$$\widetilde{f}_{j,k}(x) = [f_j(x) = k], \quad k = 1, \ldots, K-1.$$

Усложнение модели

Конструирование новых признаков на основе имеющихся:

- Применение функций к признакам (степени, логарифм, экспонента, ...).
- Добавление взаимодействий между признаками (перемножение, деление, ...).

Резюме лекции

- Линейная модель регрессии
  - МНК и ММП, их связь
  - Аналитическое решение
  - Численное решение. Градиентный спуск
  - Проблема мультиколлинеарности
  - L_1 и L_2-регуляризации
  - Вероятностный смысл регуляризации
  - Преобразование и конструирование признаков