

Занятие 6.1. Многомерная линейная регрессия

Гирдюк Дмитрий Викторович

21 октября 2023

СПбГУ, ПМ-ПУ, ДФС

Многомерная линейная регрессия

- Многомерная линейная регрессия

$$y = h(\mathbf{x}; \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^p \theta_j \tilde{x}_j = \boldsymbol{\theta}^T \mathbf{x}$$

где вектор $\mathbf{x} = [1, x_1, \dots, x_p]^T$ дополнен единицей.

- Сильный "inductive bias" и интерпретируемость.
- Нормализация (стандартизация) входных признаков:

$$\mathbf{Z} = \frac{\mathbf{X} - \overline{\mathbf{X}}}{S}$$

где $\overline{\mathbf{X}}$ – выборочное среднее, а S – корень из (несмещенной) выборочной дисперсии.

Метод наименьших квадратов i

- Используя выборку $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$, обучим модель методом наименьших квадратов:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^N \left(y^{(i)} - h(\mathbf{x}^{(i)}; \boldsymbol{\theta}) \right)^2 = \sum_{i=1}^N \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 = \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \\ &= \mathbf{y}^T \mathbf{y} - \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = \\ &= \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}\end{aligned}$$

- Необходимое условие минимума

$$\implies \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta}$$

$$\implies -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = 0$$

$$\implies \boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

- Значение лосса

$$\mathcal{L}(\boldsymbol{\theta}^*) = \|\mathbf{y} - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\|_2^2$$

где $P = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ есть проекционная матрица.

Метод наименьших квадратов iii

- Геометрический смысл МНК: опускание перпендикуляра из вектора откликов y на подпространство, образованное столбцами матрицы X .

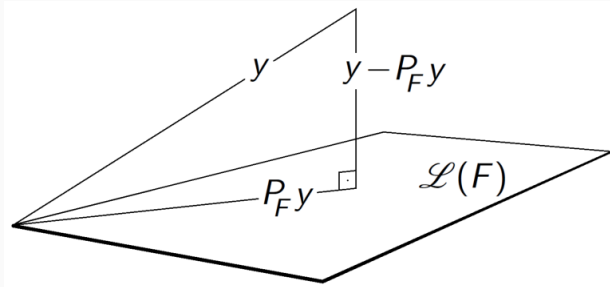


Рис. 1: Геометрическая интерпретация МНК [1]

- Вопрос: всегда ли решение единственно?
- Число обусловленности матрицы:

$$\mu(\mathbf{A}) = \|\mathbf{A}\| \|\mathbf{A}^{-1}\| = \frac{\lambda_{max}}{\lambda_{min}}$$

- Умножаем вектор на обратную матрицу \mathbf{A}^{-1} – относительная погрешность усиливается в $\mu(\mathbf{A})$ раз.
- Если матрица $\mathbf{X}^T \mathbf{X}$ плохо обусловлена, то решение $\boldsymbol{\theta}^*$ неустойчиво, норма $\|\boldsymbol{\theta}^*\|$ велика, возникает переобучение.

- Попытаться побороть это можно следующим образом: либо производим отбор признаков, либо эти признаки преобразуем в другие признаки, либо добавляем дополнительное условие на θ .
- Накладывание условий на вектор параметров называется регуляризацией.
- Замечание. В реальности никто не обращает матрицу $X^T X$ в лоб. Используют SVD или QR разложения этой матрицы, или вовсе используют итеративные алгоритмы (например, метод сопряженных градиентов) [2].

- Если линейная модель адекватно представляет данные, то график остатков $\hat{y}^{(i)} = y^{(i)} - \theta^T x^{(i)}$ будет выглядеть как облако точек достаточно симметрично распределенных относительно прямой $y = 0$.
- RSS (residual sum of squares)

$$RSS(\theta) = \sum_{i=1}^N \left(y^{(i)} - \hat{y}^{(i)} \right)^2$$

- RMSE (root mean squared error)

$$RMSE(\theta) = \sqrt{\frac{1}{N} RSS(\theta)}$$

- Коэффициент детерминации R^2

$$R^2 = 1 - \frac{\sum_{i=1}^N (y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^N (y^{(i)} - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

где $\bar{y} = \frac{1}{N} \sum_{i=1}^N y^{(i)}$.

- Вопрос: каково множество значений коэффициента детерминации и как его интерпретировать?

- Дополним рассмотренный функционал \mathcal{L} штрафом на увеличение L_2 нормы весов θ

$$\mathcal{L}_{\text{ridge}}(\theta) = \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_2^2$$

- Замечание:

$$\|\theta\|_2 = \sqrt{\sum_{j=1}^D |\theta_j|^2} = \sqrt{\theta^T \theta}$$

- Данная техника в общем случае называется L_2 -регуляризацией (weight decay).

- Аналогичное аналитическое решение

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{ridge}}}{\partial \boldsymbol{\theta}} &= -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} + 2\lambda \boldsymbol{\theta} \\ \implies -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} + 2\lambda \boldsymbol{\theta} &= 0 \\ \implies \boldsymbol{\theta} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \mathbf{X}^T \mathbf{y}.\end{aligned}$$

Подбор λ

- Тестировать разные значения (например, 0.001, 0.01, 0.1 и т.п.) на валидационной выборке (не путать с тренировочной).
Кросс-валидация (изучим в следующий раз).
- Можно начать с некоторого достаточно большого значения и постепенно снижать его.

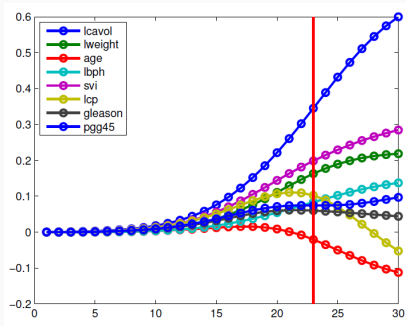


Рис. 2: Коэффициенты гребневой регрессии на данных по раку простаты по отношению к максимальному значению нормы $\|\theta\|_2$ [2].

- Лассо Тибширани (LASSO – Least Absolute Shrinkage and Selection Operator)

$$\mathcal{L}_{\text{lasso}}(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \mu\|\boldsymbol{\theta}\|_1$$

где $\|\boldsymbol{\theta}\|_1 = \sum_{j=1}^D |\theta_j|$ – L_1 норма.

- Целевая функция выше – лагранжиан. Эквивалентная оптимизационная задача выглядит так

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2, \quad \|\boldsymbol{\theta}\|_1 \leq C$$

- Основная идея – занулять коэффициенты у малозначимых признаков.

Почему коэффициенты зануляются? [2]

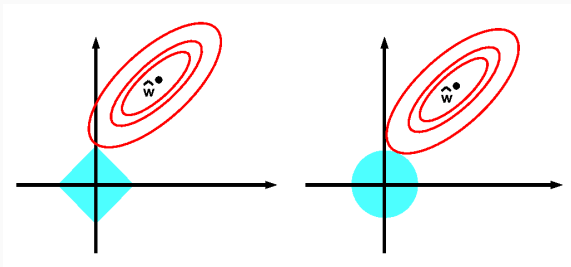
- Еще раз взглянем на оптимизационную задачу

$$\min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|_2^2, \quad \|\theta\|_1 \leq C$$

- Аналогично для гребневой регрессии

$$\min_{\theta} \|\mathbf{y} - \mathbf{X}\theta\|_2^2, \quad \|\theta\|_2^2 \leq C$$

- Вообще, зануление коэффициентов будет происходить и у $L_{p,p < 1}$ норм, однако для них оптимизационная задача теряет выпуклость.



- Алгоритмов для решения данной оптимизационной задачи хватает: метод покоординатного спуска, метод проксимального градиента, можно использовать вариацию градиентного спуска с учетом ограничений (projected gradient descent).
- У последнего подход следующий: сделаем замену переменных $\theta = \theta^+ - \theta^-$, где $\theta_j^+ = \max\{\theta_j, 0\}$ и $\theta_j^- = -\min\{\theta_j, 0\}$.
- Тогда $\|\theta\|_1 = \sum_{j=1}^D (\theta_j^+ + \theta_j^-) \leq C$. Причем $\theta^+ \geq 0$ и $\theta^- \geq 0$.

- Те же идеи, что и для гребневой регрессии

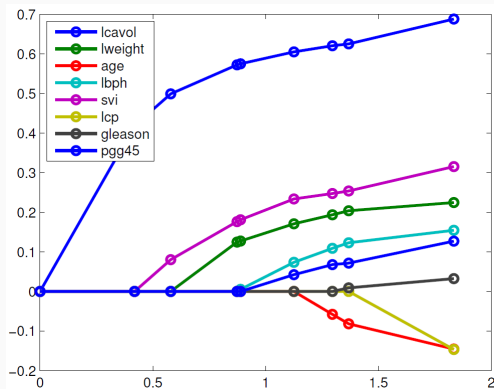


Рис. 3: Коэффициенты лассо Тибширани на данных по раку простаты по отношению к $\mu = \frac{1}{C}$ [2].

- Elastic net – комбинация L_1 и L_2 норм

$$\mathcal{L}_{\text{elastic_net}}(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \mu\|\boldsymbol{\theta}\|_1 + \lambda\|\boldsymbol{\theta}\|_2^2$$

- Групповое лассо: хотим обнуления коэффициентов у группы признаков. Например, после one-hot-encoding'а отбросить бинарные столбцы, относящиеся к одному категориальному признаку.

Линейная регрессия в scikit-learn

- scikit-learn предлагает кучу всего: LinearRegression, Ridge, Lasso, ElasticNet и другие обобщения регрессии с регуляризаторами.
- LinearRegression – обертка над `scipy.linalg.lstsq`, который, в свою очередь, использует небезызвестный LAPACK (QR-разложение по умолчанию).
- Для Ridge есть выбор из кучи оптимизационных алгоритмов на все случаи жизни: на основе сингулярного разложения (SVD, рассмотрим чуть позже), на основе разложения Холецкого, L-BFGS-B и другие.
- Lasso использует имплементацию ElasticNet с занулением коэффициента при L_2 норме. В scikit-learn решают задачу, используя собственную реализацию покоординатного спуска на Cython'е.

1. *Воронцов К.* Презентация по многомерной линейной регрессии из курса лекций Воронцова К.В. URL: <http://www.machinelearning.ru/wiki/images/a/a2/Voron-ML-regression-slides.pdf>.
2. *Murphy K. P.* **Probabilistic Machine Learning: An introduction.** MIT Press, 2022. URL: probml.ai.