Многомерная линейная регрессия

Гирдюк Дмитрий Викторович

02 ноября 2024

СПбГУ, ПМ–ПУ, ДФС

Многомерная линейная регрессия

- Многомерная линейная регрессия

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

  где вектор $\boldsymbol{x} = [1, x_1, \ldots, x_p]^T$ дополнен единицей.
- Сильный "inductive bias" и интерпретируемость.

Метод наименьших квадратов i

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

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

Метод наименьших квадратов ii

- Необходимое условие минимума

$$\begin{aligned} \implies & \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} \\ \implies & -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} = 0 \\ \implies & \boldsymbol{\theta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \end{aligned}$$

- Значение лосса

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

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

Метод наименьших квадратов iii

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

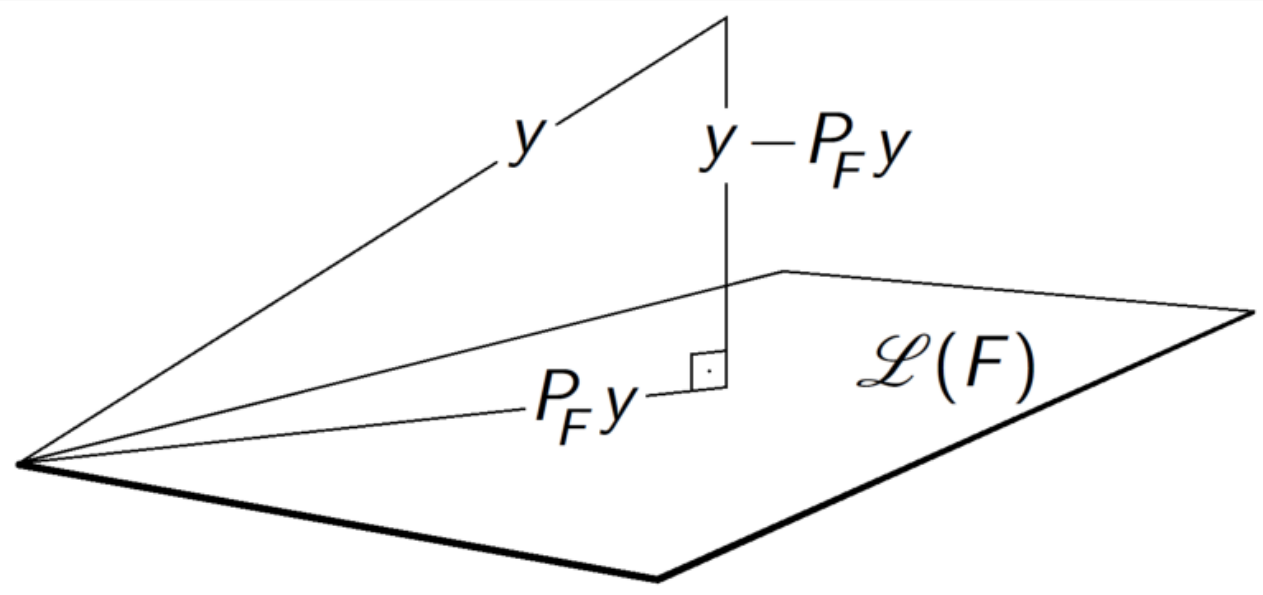


Рис. 1: Геометрическая интерпретация МНК [1]

Оценки качества модели i

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

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

- RMSE (root mean squared error)

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

Оценки качества модели ii

- Коэффициент детерминации R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y^{(i)} - \widehat{y}^{(i)}\right)}{\sum_{i=1}^{N} \left(y^{(i)} - \overline{y}\right)} = 1 - \frac{RSS}{TSS}$$

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

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

Проблемы

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

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

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

Решения

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

Гребневая регрессия

- Дополним рассмотренный функционал $\mathcal{L}$ штрафом на увеличение L_2 нормы весов $\boldsymbol{\theta}$

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

- Замечание:

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

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

МНК для гребневой регрессии

- Аналогичное аналитическое решение

$$\frac{\partial \mathcal{L}_{\text{ridge}}}{\partial \boldsymbol{\theta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} + 2\lambda\boldsymbol{\theta}$$
$$\Longrightarrow -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} + 2\lambda\boldsymbol{\theta} = 0$$
$$\Longrightarrow \boldsymbol{\theta} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_D\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

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.

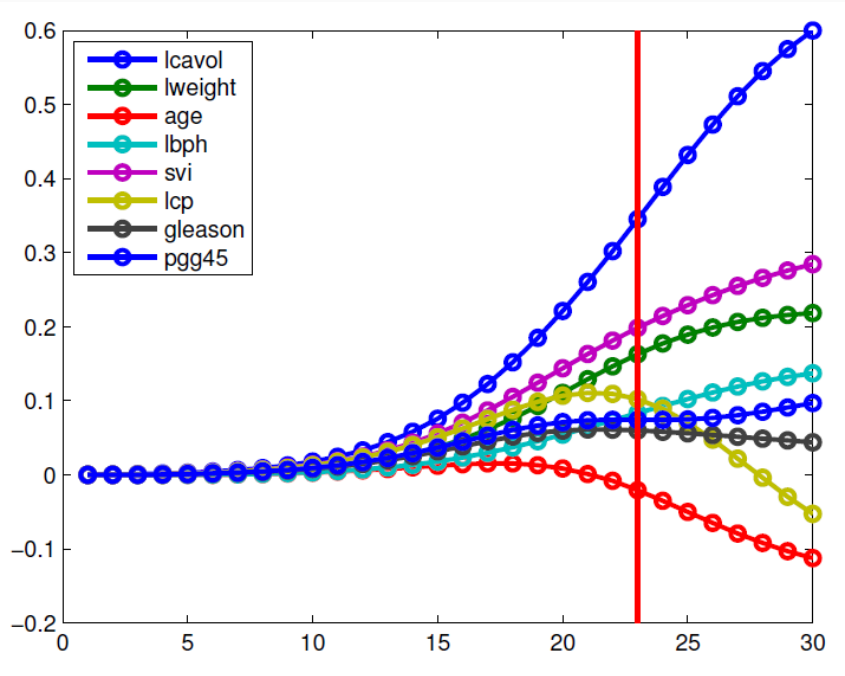


Рис. 2: Коэффициенты гребневой регрессии на данных по раку простаты по отношению к максимальному значению нормы $||\boldsymbol{\theta}||_2$ [2].

Лассо Тибширани

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

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

  где $||\boldsymbol{\theta}||_1 = \sum_{j=1}^{D} |\theta_j|$ – L_1 норма.
- Целевая функция выше – лагранжиан. Эквивалентная оптимизационная задача выглядит так

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

- Основная идея – зануллять коэффициенты у малозначимых признаков.

Почему коэффициенты зануляются? [2]

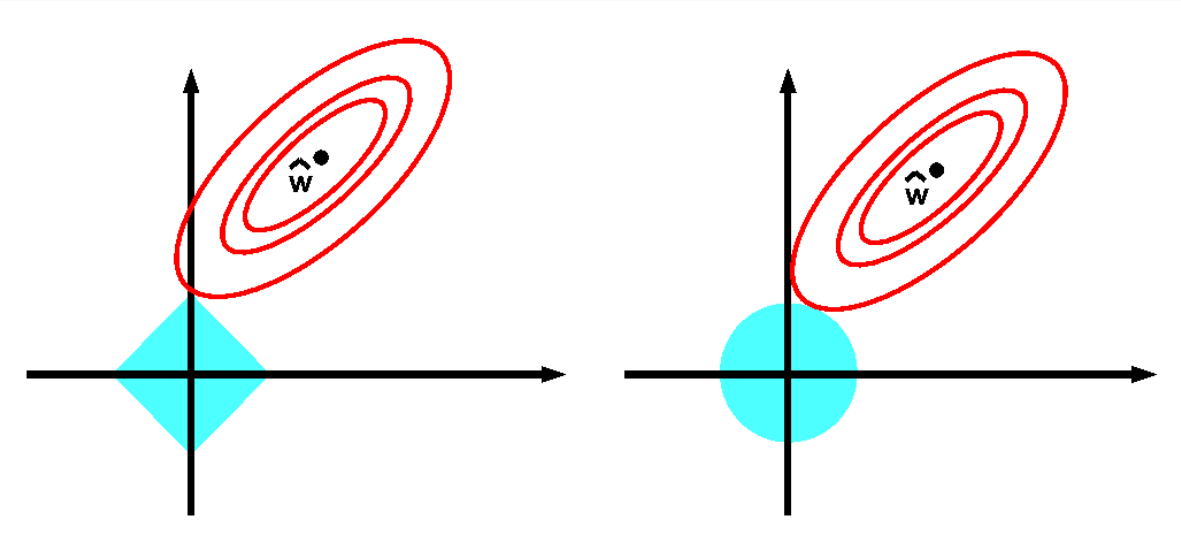
- Еще раз взглянем на оптимизационную задачу

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

- Аналогично для гребневой регрессии

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

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



Лассо Тибширани: решение

- 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 $\boldsymbol{\theta} = \boldsymbol{\theta}^+ - \boldsymbol{\theta}^-$, где $\theta_j^+ = \max\{\theta_j, 0\}$ и $\theta_j^- = -\min\{\theta_j, 0\}$.
- Тогда $||\boldsymbol{\theta}||_1 = \sum_{j=1}^{D} \left(\theta_j^+ + \theta_j^-\right) \leqslant C$. Причем $\boldsymbol{\theta}^+ \geq \mathbf{0}$ и $\boldsymbol{\theta}^- \geq \mathbf{0}$.

Подбор μ [2]

- Те же идеи, что и для гребневой регрессии

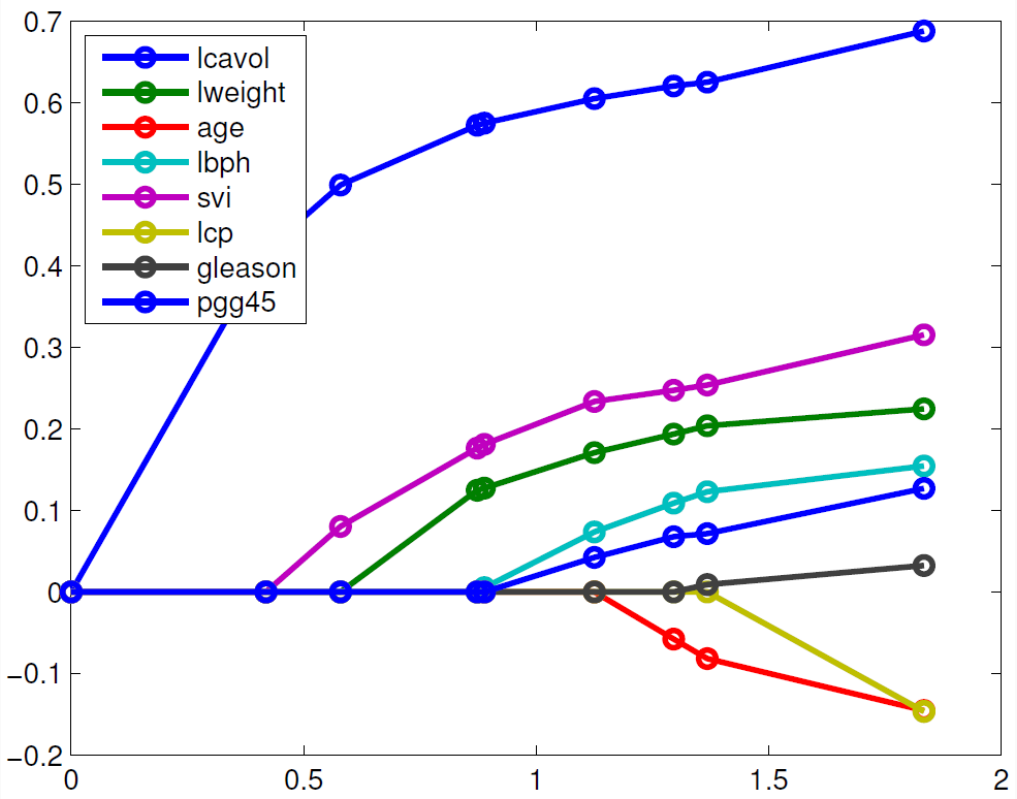


Рис. 3: Коэффициенты лассо Тибширани на данных по раку простаты по отношению к $\mu = \frac{1}{C}$ [2].

Еще про регуляризацию

- Elastic net – комбинация L_1 и L_2 норм

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

- Групповое лассо: хотим обнуления коэффициентов у группы признаков. Например, после one-hot-endocing'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'e.

Использованные источники i

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.