Регрессия, регуляризация, отбор признаков

Михаил Козак, Павел Мехнин, Данил Шкурат

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

Семинар по статистическому и машинному обучению

Санкт-Петербург, 2022

Обучение с учителем

Обучение с учителем — это направление машинного обучения, объединяющее алгоритмы и методы построения моделей на основе совокупности прецедентов (обучающей выборки), содержащих пары «известный вход — известный выход».

Регрессия: формальная постановка

- X — множество объектов, заданное своими признаками (точки p-мерного пространства)
- Y — множество ответов (действительные числа)
- Предполагаем наличие неизвестной зависимости между объектами и ответами $y : X \to Y$
- Обучающая выборка из множества объектов $\{x_1, x_2, \dots, x_n\} \subset X$ и известных ответов (откликов) $y_i = y(x_i) \in Y, i = 1, \dots, n$

Задача:
Найти $y^* : X \to Y$ — отображение, приближающее неизвестную функцию y на всём множестве X, то есть восстановить зависимость, способную для любого объекта выдать достаточно точный отклик.

Задача регрессии как задача оптимизации

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

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

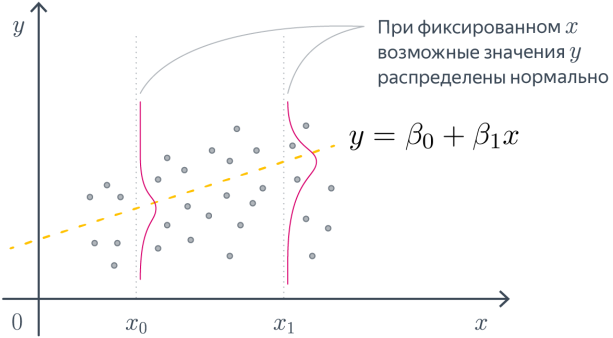
- Задача обучения по МНК — задача минимизации

$$Q(\beta, X_n) \to \min_{\beta \in B}.$$

Линейная регрессия: постановка

Модель: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

- $\boldsymbol{y} \in \mathbb{R}^n$ — вектор ответов, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ — вектор ошибок, $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}$
- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ — матрица данных
- $\boldsymbol{\beta} \in \mathbb{R}^p$ — вектор параметров модели



Решение задачи линейной регрессии — вектор $\hat{\boldsymbol{\beta}}$.

Задача оптимизации (с квадратичной функцией потерь):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

Полученную оценку называют оценкой по методу наименьших квадратов. Она имеет явный вид:

$$\hat{\beta}_{\text{МНК}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{X}^{-}\boldsymbol{y}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}_{\text{МНК}}$$

Теорема Гаусса–Маркова утверждает, что $\hat{\beta}_{\text{МНК}}$ имеет наименьшую дисперсию среди всех несмещённых оценок (best linear unbiased estimate — BLUE).

Достаточно быстро вычисляется посредством применения сингулярного разложения.

МНК-оценка: вычислительный взгляд

Сингулярное разложение: $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}$

- $\boldsymbol{V}$ и $\boldsymbol{U}$ — ортогональные, $\boldsymbol{D}$ — диагональная
- $\boldsymbol{V} = (V_1, V_2, \dots, V_n) \in \mathbb{R}^{n\times n}$, V_i — с. векторы $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$
- $\boldsymbol{U} = (U_1, U_2, \dots, U_n) \in \mathbb{R}^{p\times n}$, U_i — с. векторы $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$
- $\boldsymbol{D} = \mathrm{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$, $\lambda_j \geqslant 0$ — с. значения $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$

Подставим в формулу для $\hat{\beta}_{\text{МНК}}$ вместо матрицы $\boldsymbol{X}$ её сингулярное разложение и получим

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y},$$

где $\boldsymbol{D}^{-1} = \mathrm{diag}(1/\sqrt{\lambda_1}, \dots, 1/\sqrt{\lambda_n})$.

$$\hat{\beta}_{\text{МНК}} = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j (V_j^{\mathrm{T}}\boldsymbol{y})$$

Мультиколлинеарность признаков

Пусть матрица данных содержит несколько сильно коррелированных признаков (есть $\lambda_j \to 0$).
Что будет происходить в таком случае с МНК-оценкой?

- Решение $\hat{\beta}$ неустойчиво
- Решение неинтерпретируемо, $||\hat{\beta}|| \to \infty$
- Высокая дисперсия у $\hat{\beta} \to$ высокая MSE
- Ответы на контрольной выборке неустойчивы (переобучение)

Способы решения проблемы:

- Регуляризация
- Уменьшение числа признаков (отбор признаков)
- Преобразование признаков (анализ главных компонент)

Хорошая оценка $\hat{\boldsymbol{\beta}}$ должна иметь низкую среднеквадратическую ошибку

$$\mathbb{E}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^2 = \underbrace{\mathbb{D}\hat{\boldsymbol{\beta}}}_{\text{дисперсия}} + \underbrace{(\mathbb{E}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2}_{\text{смещение}}.$$

Несмещенная МНК-оценка не гарантирует минимизацию всей MSE, т.к. может иметь большую дисперсию (если матрица $\boldsymbol{X}$ близка к вырожденной).
Введение небольшого смещения в оценке может привести к уменьшению дисперсии и тем самым уменьшению $\mathrm{MSE}_{\text{test}}$.

Регуляризация. Гребневая регрессия

- Вводим штраф за увеличение нормы вектора $\boldsymbol{\beta}$ и переходим к минимизации следующей функции:

$$Q_\tau(\boldsymbol{\beta}) = ||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}||^2 + \tau||\boldsymbol{\beta}||^2 \to \min_{\boldsymbol{\beta}},$$

где τ — неотрицательный параметр регуляризации.

- В развернутом виде задача оптимизации записывается так:

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

- Решение:

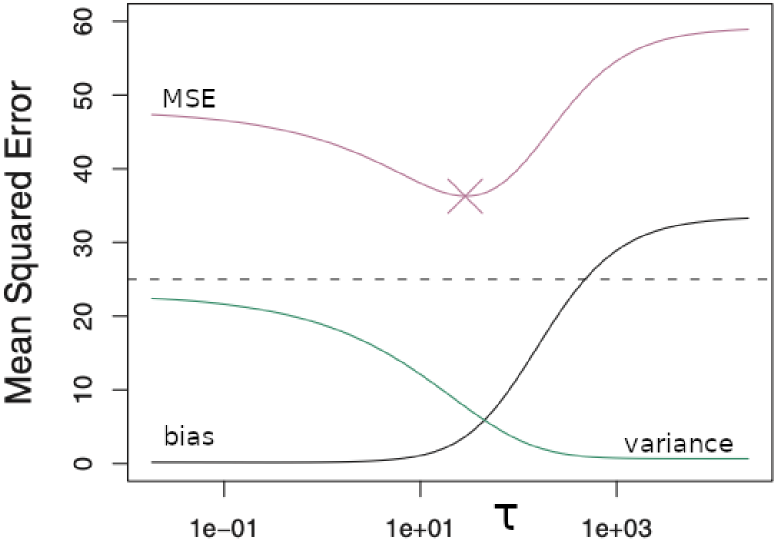
$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \tau\boldsymbol{I}_p)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$

Используем SVD и получим

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \sum_{i=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} U_j (V_j^{\mathrm{T}} \boldsymbol{y})$$

Отделили знаменатель от нуля. Устойчивость вычислений повышается.
Преимущество SVD разложения: можно подбирать параметр τ вычислив SVD только один раз.
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.

Гребневая регрессия. Выбор параметра



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

Скользящий контроль — вычислительно трудоёмкая процедура. Известна практическая рекомендация брать брать τ в отрезке $[0.1, 0.4]$, если столбцы матрицы $\boldsymbol{X}$ заранее стандартизованы.

Лассо регуляризация

- В качестве штрафа за увеличение нормы вектора $\boldsymbol{\beta}$ используется его l_1-норма.
- Метод LASSO решает следующую задачу минимизации:

$$||\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}||_2^2 + \tau||\boldsymbol{\beta}||_1^2 \to \min_{\boldsymbol{\beta}},$$

  где τ — неотрицательный параметр регуляризации.
- Задача оптимизации в развернутом виде:

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

Задачу lasso-оптимизации можно переписать в форме с ограничениями:

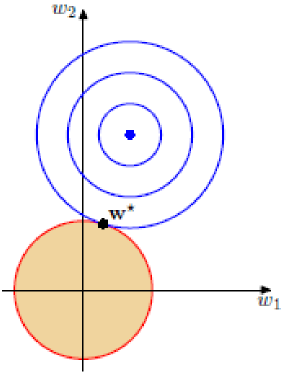
$$\begin{cases} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \to \min_{\beta_1,\ldots,\beta_p}, \\ \sum_{j=1}^{p}|\beta_j| \le æ, \end{cases}$$

где $æ = 1/\tau$.

Особенности:

- Уменьшение MSE
- Интерпретируемость результатов
- Быстрое вычисление $\hat{\beta}_{\text{Lasso}}(\tau)$
- Выбор параметра: кросс-валидация

Сравнение гребневой регрессии и Лассо



Синие линии уровня функционала качества (синяя точка — безусловный минимум, который достигается на МНК решении). Оранжевая зона — ограничения, задаваемые L2 и L1-регуляризаторами. Чёрная точка — минимум целевой функции при заданном ограничении.

Сравнение гребневой регрессии и Лассо

- Оба метода успешно решают проблему мультиколлинеарности
- Гребневая регрессия использует все признаки
- Лассо производит отбор признаков, что предпочтительнее, если среди признаков есть шумовые или измерения признаков связаны с ощутимыми затратами.
- С помощью кросс-валидации можно определить какой подход лучше для конкретных данных.

Elastic net regularization

Решается задача оптимизации

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \tau_1||\boldsymbol{\beta}||_1^2 + \tau_2||\boldsymbol{\beta}||_2^2 \to \min_{\boldsymbol{\beta}}.$$

- ✓ 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

Алгоритм полного перебора (Best Subset Selection)

Преимущества:

- ✓ простота реализации
- ✓ гарантированный результат
- ✓ эффективен, когда информативных признаком немного (≤ 5) и общее количество признаков также не велико $(\leq 20 \ldots 100)$

Недостатки:

- ✗ в общем случае очень долго: $O(2^p)$. Например, для p = 20: $2^p = 1,048,576$ моделей
- ✗ чем больше перебирается вариантов, тем больше перобучение

Решение: эвристические алгортмы сокращенного перебора

Жадные (greedy) алгоритмы

- Forward Stepwise Selection
  Начинаем со свободного члена, потом добавляем на каждом шаге предиктор, который максимально уменьшает ошибку.
  Подмножества получаются вложенные — для $p = 20$: $p(p+1)/2 = 210$ моделей.
- Backward Stepwise Selection
  Начинаем с полной регрессии и на каждом шаге убираем предиктор, который оказывает меньше всего влияния на ошибку.