Регрессия, регуляризация, работа с признаками

Михайлов Дмитрий, Смирнов Иван

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

Санкт-Петербург, 2023

Идея

Имеется набор данных (обучающая выборка)

$$\boldsymbol{X} \in \mathbb{R}^{n \times p}, \quad \boldsymbol{y} \in \mathbb{R}^n$$

$\mathbf{x}_i \in \mathbb{R}^p$ — вектор-строки $\boldsymbol{X}$, $X_i \in \mathbb{R}^n$ — вектор-столбцы $\boldsymbol{X}$.

Задача: уметь предсказывать y (ответ) по новым $\mathbf{x}_i$ (объектам), установив некоторую зависимость на обучающей выборке.

Постановка задачи

"До эксперемента":

$\boldsymbol{\xi} \in \mathbb{R}^p$ — случайный вектор, $\eta, \varepsilon \in \mathbb{R}$ — случайные величины. Предполагаем, что η и $\boldsymbol{\xi}$ функционально зависимы:

$$\eta = \varphi(\boldsymbol{\xi}) + \varepsilon$$

Обычно $\mathrm{E}\varepsilon = 0$, $\mathrm{D}\varepsilon = \sigma^2$, $\boldsymbol{\xi} \perp \varepsilon$.

"После эксперемента":

$\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{L}(\boldsymbol{\xi})$; $y_1, \ldots, y_n \sim \mathcal{L}(\eta)$ — выборки, которые наблюдаем. Модель для всех $i \in 1:n$

$$y_i = \varphi(\mathbf{x}_i) + \varepsilon_i$$

Задача: найти функцию φ.

Этапы обучения модели

Модель

$$y_i = \varphi(\mathbf{x}_i) + \varepsilon_i$$

Задача: найти функцию φ.

1. Выбор модели регрессии (класс рассматриваемых $\varphi(\cdot)$)
   Линейная модель: $\varphi(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_i \mathbf{x}_i[j], \quad i \in 1:n$
2. Выбор функции потерь (loss function)
   Квадратичная функция потерь: $\sum_{i=1}^{n} (y_i - \varphi(\mathbf{x}_i, \boldsymbol{\beta}))^2$
3. Выбор метода обучения (training)
   МНК: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \varphi(\mathbf{x}_i, \boldsymbol{\beta}))^2$
4. Выбор метода проверки (test)
   MSE: $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i^{\text{test}} - \varphi(\mathbf{x}_i^{\text{test}}, \hat{\boldsymbol{\beta}}))^2$

Оптимизация. Матричный вид в общем случае

- $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ — матрица данных
- $\boldsymbol{y} \in \mathbb{R}^{n}$ — вектор ответов
- $\boldsymbol{\beta} \in \mathbb{R}^{d}$ — вектор параметров
- $\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}) := (\varphi(\mathbf{x}_1, \boldsymbol{\beta}), \dots, \varphi(\mathbf{x}_n, \boldsymbol{\beta}))^{\mathrm{T}}$ — функция от выборки и параметров
- $\mathcal{L}(\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}), \boldsymbol{y})$ — *некоторая* функция потерь

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

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}), \boldsymbol{y})$$

Выбор функции потерь

Какие варианты есть?

- $\|\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}) - \boldsymbol{y}\|_2^2$ — квадратичная ошибка (l_2-норма)
- $\|\boldsymbol{\varphi}(\boldsymbol{X}, \boldsymbol{\beta}) - \boldsymbol{y}\|_1$ — модуль ошибки (l_1-норма)

Как выбирать?

- Предположения о распределении остатков ε
- Простота функции $\mathcal{L}$ для оптимизации
- Точность данных/наличие выбросов

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

Решение задачи линейной регрессии — вектор $\hat{\boldsymbol{\beta}}$.
Классическая задача (с квадратичной функцией потерь):

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

Ищем наилучшую аппроксимацию в подпространстве столбцов $\boldsymbol{X}$.

МНК-оценка и её особенности I

Задача

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

Решение

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

- ✓ Явный вид решения
- ✓ Простота функции $\mathcal{L}$ для оптимизации
- ! Точность данных/наличие выбросов
- ! Конкретные предположения о распределении остатков ε
- ! Мультиколлинеарность
- ! $n \geqslant p$, иначе решений бесконечно много

МНК-оценка и её особенности II

Для оценок $\hat{\boldsymbol{\beta}}$ имеет место разложение

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

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

- $\mathrm{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, то есть оценка несмещённая
- $\mathrm{D}\hat{\boldsymbol{\beta}} = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ (в случае, когда $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$)
- В классе несмещённых оценок $\hat{\boldsymbol{\beta}}$ обладает наименьшей дисперсией ($\hat{\boldsymbol{\beta}}$ — BLUE)
- Если остатки распределены нормально, то $\hat{\boldsymbol{\beta}}$ — ОМП

Анализ остатков

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

- Распределение $\boldsymbol{\varepsilon}$ — нормальное, $\mathrm{E}\boldsymbol{\varepsilon} = \mathbf{0}$:
  - Классическая оценка по МНК — ОМП и BLUE
  - Работают стандартные критерии значимости
- Распределение $\boldsymbol{\varepsilon}$ неизвестно или с тяжёлыми хвостами:
  - Оценка по МНК уже не ОМП и не BLUE

МНК-оценка: вычислительный взгляд

Решение

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

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

  Отсюда $\hat{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y}$
- $\boldsymbol{D}^{-1} = \mathrm{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_n})$

Борьба с мультиколлинеарностью

Мультиколлинеарность ($\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ плохо обусловлена) влечёт

- Неустойчивость решения
- Высокая дисперсия у $\hat{\boldsymbol{\beta}} \Rightarrow$ высокая MSE

Возможные решения проблемы мультиколлинеарности

- Уменьшение числа признаков (отбор признаков)
- Регуляризация
- Преобразование признаков (АГК)

Регуляризация. Гребневая регрессия

Модель: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Вернёмся к разложению MSE

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

Несмещённая оценка может иметь большую дисперсию и MSE.
Возьмём оценку по МНК и сделаем её смещённой:

$$\hat{\boldsymbol{\beta}}_{\mathsf{ridge}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \quad \lambda > 0$$

Используем SVD и получим

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

Отделили знаменатель от нуля. Устойчивость вычислений повышается.

Гребневая регрессия. Дисперсия

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

Оценка по методу гребневой регрессии

$$\hat{\boldsymbol{\beta}}_{\mathsf{ridge}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \quad \lambda > 0$$

Смещение контролируется параметром λ.
Дисперсия:

- $\hat{\boldsymbol{\beta}}_{\mathsf{ridge}} = \sum_{i=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \lambda} U_j (V_j^{\mathrm{T}} \boldsymbol{y})$
- λ_j убывают
- $\frac{\sqrt{\lambda_j}}{\lambda_j + \lambda}$ штрафуют компоненты с наименьшей дисперсией
- $\mathrm{D}\hat{\boldsymbol{\beta}}$ уменьшается $\Rightarrow$ MSE $\downarrow$

Гребневая регрессия как задача оптимизации

Есть две эквивалентные формулировки:

Ridge Regression

$$\hat{\boldsymbol{\beta}}_{\mathsf{ridge}}(\lambda_1) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_2^2$$

$$\hat{\boldsymbol{\beta}}_{\mathsf{ridge}}(\lambda_2) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \text{ s.t. } \|\boldsymbol{\beta}\|_2^2 \leqslant \lambda_2$$

Регуляризация. Lasso

Есть две эквивалентные формулировки (явного вида нет):

Lasso

$$\hat{\boldsymbol{\beta}}_{\textsf{Lasso}}(\lambda_1) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1^2$$

$$\hat{\boldsymbol{\beta}}_{\textsf{Lasso}}(\lambda_2) = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \text{ s.t. } \|\boldsymbol{\beta}\|_1^2 \leqslant \lambda_2$$

Особенности:

- Уменьшение MSE
- Интерпретируемость результатов
- Выбор параметра: кросс-валидация

Ridge VS Lasso

Тут мы видим, что у lasso точка пересечения $(0; \beta_2^{lasso})$, а у ridge $(\beta_1^{ridge}; \beta_2^{ridge})$, соответственно lasso позволяет нам таким образом упростить модель и повысить её интерпретируемость.

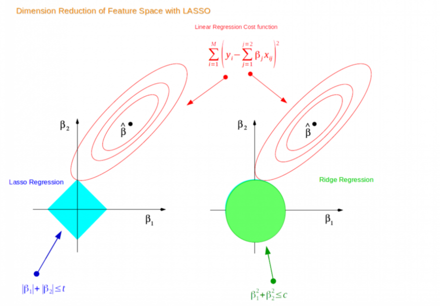
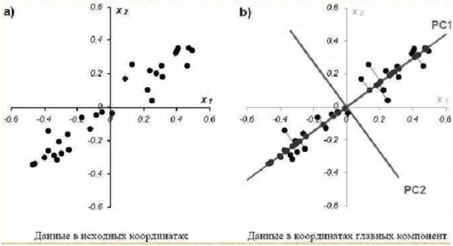


Figure 3: Why LASSO can reduce dimension of feature space? Example on 2D feature space. Modified from the plot used in 'The Elements of Statistical Learning' by Author.

Преобразование признаков. АГК

В методе главных компонент строится минимальное число новых признаков, по которым исходные признаки восстанавливаются линейным преобразованием с минимальными погрешностями.



Отбор признаков. Общий подход к выбору модели

Предположим, что есть некоторое семейство построенных моделей $\{\mathcal{M}_i\}_{i \in I}$.
Хотим выбрать лучшую модель $\mathcal{M}^*$ для предсказания.

Классические методы выбора модели

- Кросс-валидация
- Информационные критерии
  - $\mathrm{AIC}_i = 2p_i - 2\ln \mathcal{L}(\boldsymbol{X}; \mathcal{M}_i)$
  - $\mathrm{BIC}_i = p_i \ln n - 2\ln \mathcal{L}(\boldsymbol{X}; \mathcal{M}_i)$