

# Регрессия, регуляризация, отбор признаков

Елена Гоголева, Дэвид Капаца, Анастасия Мандрикова

Санкт-Петербургский государственный университет  
Прикладная математика и информатика  
Вычислительная стохастика и статистические модели

Семинар по статистическому и машинному обучению



Санкт-Петербург, 2021

## 1 Регрессия в общем виде

- Какая задача оптимизации решается?
- Выбор функции потерь

## 2 Линейная регрессия

- Многомерная постановка, МНК-оценка  $\hat{\beta}$
- Предположения о модели. Остатки
- Критерии
- Особенности вычисления  $\hat{\beta}$

## 3 Регуляризация

- Ridge Regression (гребневая регрессия)
- Lasso
- Некоторые модификации (Elastic Net, RegLAD)

## 4 Отбор признаков

- Best subset selection
- Forward/Backward selection

Имеется набор данных (обучающая выборка)

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

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

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

Какие минимальные требования наложить на все  $\mathbf{x}_i$  и  $y$ ?

## Гипотеза непрерывности

«близким» объектам  $\mathbf{x}_i$  соответствуют «близкие» ответы  $y_i$

## Генеральная постановка

$\xi \in \mathbb{R}^p$  — случайный вектор,  $\eta, \varepsilon \in \mathbb{R}$  — случайные величины.

Предполагаем, что  $\eta$  и  $\xi$  функционально зависимы:

$$\eta = \varphi(\xi) + \varepsilon$$

Обычно  $E\varepsilon = 0$ ,  $D\varepsilon = \sigma^2$ ,  $\xi \perp \varepsilon$ .

## Переход к выборке

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

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

**Задача:** найти функцию  $\varphi$ .

## Модель

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

**Задача:** найти функцию  $\varphi$ .

- ① Выбор модели регрессии (класс рассматриваемых  $\varphi(\cdot)$ )

Линейная модель:  $\varphi(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{j=1}^p \beta_j \mathbf{x}_i[j], \quad i \in 1 : n$

- 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:  $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i^{\text{test}} - \varphi(\mathbf{x}_i^{\text{test}}, \hat{\boldsymbol{\beta}}))^2$

# Задача регрессии как задача оптимизации

- $\mathbf{X} \in \mathbb{R}^{n \times p}$  — матрица данных (design matrix)
- $\mathbf{y} \in \mathbb{R}^n$  — вектор ответов
- $\boldsymbol{\beta} \in \mathbb{R}^d$  — вектор параметров
- $\varphi(\mathbf{X}, \boldsymbol{\beta}) := (\varphi(\mathbf{x}_1, \boldsymbol{\beta}), \dots, \varphi(\mathbf{x}_n, \boldsymbol{\beta}))^\top$  — функция от выборки и параметров
- $\mathcal{L}(\varphi(\mathbf{X}, \boldsymbol{\beta}), \mathbf{y})$  — некоторая функция потерь<sup>1</sup>

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

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

---

<sup>1</sup>здесь  $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$

# Выбор функции потерь

Какие варианты есть?

- $\|\varphi(\mathbf{X}, \boldsymbol{\beta}) - \mathbf{y}\|_2^2$  — квадратичная ошибка ( $l_2$ -норма)
- $\|\varphi(\mathbf{X}, \boldsymbol{\beta}) - \mathbf{y}\|_1$  — модуль ошибки ( $l_1$ -норма)
- $\frac{1}{n} \sum_{i \in 1:n} |\varphi(\mathbf{x}_i, \boldsymbol{\beta}) - y_i|$  — MAD (mean absolute deviation)<sup>2</sup>
- $\sum_{i \in 1:n} H_\delta(\varphi(\mathbf{x}_i, \boldsymbol{\beta}), y_i)$  — функция потерь Хубера
- ...

Как выбирать?

- Явный вид решения
- Простота функции  $\mathcal{L}$  для оптимизации
- Точность данных/наличие выбросов
- Конкретные предположения о распределении остатков  $\varepsilon$
- Инвариантность решения относительно масштаба/сдвига

---

<sup>2</sup>аналогично  $l_1$ , только ф-я нормирована на  $n$

Модель:  $y = X\beta + \varepsilon$

- $y \in \mathbb{R}^n$  — вектор ответов,  $\varepsilon \in \mathbb{R}^n$  — вектор ошибок,  $E\varepsilon = 0$
- $X \in \mathbb{R}^{n \times p}$  — матрица данных (design matrix)
  - детерминированная
  - случайная
- $\beta \in \mathbb{R}^p$  — вектор параметров

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

Классическая задача (с квадратичной функцией потерь):

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

Исходная модель не обязательно линейна: ищем наилучшую аппроксимацию в подпространстве столбцов  $X$ .

## Задача

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

## Решение

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

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

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

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

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

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

Модель:  $y = X\beta + \varepsilon$

- Распределение  $\varepsilon$  — нормальное,  $E\varepsilon = 0$ :
  - Классическая оценка по МНК — ОМП и BLUE
  - Работают стандартные критерии значимости
- Распределение  $\varepsilon$  неизвестно или с тяжёлыми хвостами:
  - Оценка по МНК уже не ОМП и не BLUE
  - Использование робастных функций потерь, итеративных методов
  - Проверка качества модели на тестовой выборке

## Вывод

Анализ остатков необходим!

## Решение

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

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

Отсюда  $\hat{\beta} = \mathbf{U} \mathbf{D}^{-1} \mathbf{V}^T \mathbf{y}$

- $\mathbf{D}^{-1} = \text{diag}(1/\sqrt{\lambda_1}, \dots, 1/\sqrt{\lambda_n})$

Что, если  $\lambda_j \approx 0$ ?

Плохо.

# Когда $\lambda_i \approx 0$ ? Мультиколлинеарность

- $\lambda_i$  — собственные числа  $\mathbf{X}^T \mathbf{X}$
- **Факт:** Если существует  $\mathbf{v} \in \mathbb{R}^p$  такой, что  $\mathbf{X}\mathbf{v} \approx \mathbf{0}$ , то некоторые  $\lambda_i \approx 0$
- Когда  $\mathbf{X}\mathbf{v} \approx \mathbf{0}$ ?<sup>3</sup> Происходит ли такое на практике?
- Влияние  $\lambda_i$  на  $\hat{\beta}$  можно объяснить так:

## Определение

$\mu(S) = \|S\| \|S^{-1}\| = \lambda_{max}/\lambda_{min}$  называется **числом обусловленности** матрицы  $S$

При умножении  $S^{-1}\mathbf{v} = z$  происходит увеличение погрешности в  $\mu(S)$  раз:

$$\frac{\|\delta z\|}{\|z\|} \leq \mu(S) \frac{\|\delta v\|}{\|v\|}$$

<sup>3</sup>распишите произведение как линейную комбинацию

Мультиколлинеарность ( $X^T X$  плохо обусловлена) влечёт

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

Возможные решения проблемы мультиколлинеарности

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

# Практика

## Регрессия

# Регуляризация. Гребневая регрессия

Модель:  $y = X\beta + \varepsilon$

Вернёмся к разложению MSE

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

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

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

Используем SVD и получим

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

Отделили знаменатель от нуля. Устойчивость вычислений повышается.

# Гребневая регрессия. Дисперсия

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

## Оценка по методу гребневой регрессии

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

Смещение контролируется параметром  $\lambda$ .

Что с дисперсией?

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

Есть две эквивалентные формулировки:

## Ridge Regression

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

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

**Явное решение** — уже видели.

**Демо:** <https://www.desmos.com/calculator/3fp4awzeyp>

# Гребневая регрессия. Выбор параметра

$$\hat{\beta}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

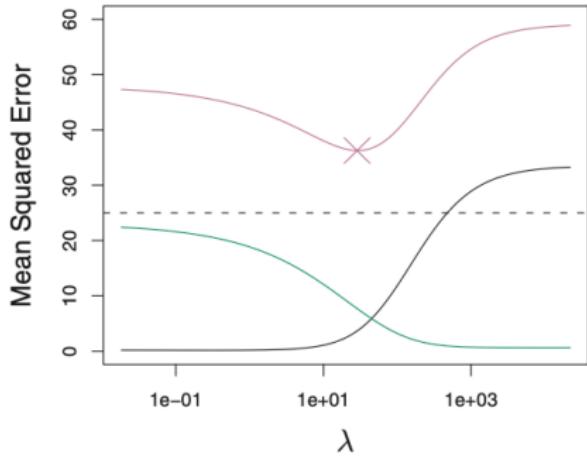
Выбор параметра  
осуществляется с помощью  
кросс-валидации.

На рисунке:

MSE

Дисперсия оценки  $\hat{\beta}_{\text{ridge}}$

Смещение оценки  $\hat{\beta}_{\text{ridge}}$



Есть две эквивалентные формулировки (явного вида нет):

## Lasso

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

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

## Особенности:

- Уменьшение MSE
- Интерпретируемость результатов
- Быстрое вычисление  $\hat{\boldsymbol{\beta}}_{\text{Lasso}}(\lambda)$
- Случай  $p > n$
- Выбор параметра: кросс-валидация

- Elastic net (совмещение Lasso и Ridge)
- Использование других норм ( $l_q$ )
- Робастные функции потерь

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
  - leave-one-out CV
  - $k$ -fold CV
- Информационные критерии и  $R^2$ 
  - AIC
  - BIC
  - $R^2$
  - adj. $R^2$

- Best Subset Selection  $p = 20$ :  $2^p = 1,048,576$  моделей
- Жадные (greedy) методы
  - Forward Stepwise Selection  $p = 20$ :  $p(p + 1)/2 + 1 = 211$
  - Backward Stepwise Selection

# Практика

## Регуляризация и отбор параметров

- ESL (Elements of Statistical Learning) — Hastie, Tibshirani, Friedman
- ISLR (An Introduction to Statistical Learning) — James, Witten, Hastie, Tibshirani
- Лекции Н.Э. и А.И.
- Лекции Воронцова по ML
- Лекции Larry Wasserman — Statistical Learning
- All of Statistics — Larry Wasserman