Регуляризация в регрессии

Ширинкина Дарья Андреевна, гр. 622

Санкт-Петербургский государственный университет
Математико-механический факультет
Статистическое моделирование

Санкт-Петербург
2017г.

Задача

Пусть

- $x_1, \ldots, x_n \in \mathbb{R}^p$ — независимые одинаково распределенные случайные величины;
- $X = [X_1, \ldots, X_p]$, где $X_i = (x_{1i}, \ldots, x_{ni})^{\mathrm{T}}$, $i = 1, \ldots, p$.

Предполагаем существование неизвестной f такой, что

$$y_i = f(x_i) + \varepsilon_i,$$

где

- $Y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$;
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ — независимые случайные величины;
- ε_i и x_j независимы для $\forall i, j$;
- $\mathbb{E}\varepsilon_i = 0$, $i = 1, \ldots, n$ и $\mathbb{E}\varepsilon_i^2 = \sigma^2$.

Задача: Оценить функцию f.

Обозначения

Обучающая выборка:

- $x_1, \ldots, x_n$ — выборка, участвующая в оценке функции f (обучающая выборка);
- $y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$.

Тестовая выборка:

- $x'_1, \ldots, x'_k$ — выборка, по которой оценивается качество оценки функции f (тестовая выборка);
- $y'_i = f(x'_i) + \varepsilon'_i$, $i = 1, \ldots, k$.

Модель

Считаем, что $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ и X — центрированы.

Модель многомерной линейной регрессии:

$$y_i = f(x_i, \beta) + \varepsilon_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i.$$

Задача минимизации:

$$\mathrm{MSE}_{\text{training}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \to \min_{\beta}.$$

Решение МНК: $\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^T Y.$

Проблема

Насколько хорошо предсказывает $\hat{f}(x) = \sum_{i=1}^{p} \hat{\beta}_i x$?

Проблема: минимизируем $\mathrm{MSE}_{\mathrm{training}}$, но хотим минимизировать

$$\mathrm{MSE}_{\mathrm{test}} = \frac{1}{n} \sum_{i=1}^{n} (y_i' - \sum_{j=1}^{p} \beta_j x_{ij}')^2.$$

- Нет гарантии, что минимум $\mathrm{MSE}_{\mathrm{training}}$ будет соответствовать минимуму $\mathrm{MSE}_{\mathrm{test}}$.
- Когда $\mathrm{MSE}_{\mathrm{test}} \gg \mathrm{MSE}_{\mathrm{training}}$, говорят, что происходит переобучение.

Проблема

Пусть

- x_i' — реализация случайной величина из тестовой выборки;
- $y_i' = f(x_i') + \varepsilon_i'$ — известное значение.

$$\mathbb{E}(y_i' - \hat{f}(x_i'))^2 = Var(\hat{f}(x_i')) + (Bias(\hat{f}(x_i')))^2 + Var(\varepsilon_i').$$

- Как правило, при увеличении сложности метода (увеличение числа параметров) дисперсия будет увеличиваться, а смещение будет уменьшаться.
- Введение небольшого смещения в оценке может привести к значительному уменьшению дисперсии и тем самым уменьшению $\mathrm{MSE}_{\mathrm{test}}$.

Регуляризация

Регуляризация: вводим ограничения на коэффициенты β.

Для чего используем регуляризацию:

- можем уменьшить дисперсию оценки за счет введения смещения и тем самым уменьшить $\mathrm{MSE}_{\mathrm{test}}$ (особенно, когда $p > n$);
- можем производить отбор значимых признаков, делая коэффициенты при них равными нулю.

Гребневая регрессия

Задача минимизации:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 \to \min_{\beta},$$

где $\lambda \geq 0$ — неотрицательный параметр регуляризации (tuning parameter).

- $\lambda \sum_{j=1}^{p}\beta_j^2$ мало, когда $\beta_1, \ldots, \beta_p$ близки к нулю.
- Когда $\lambda = 0$, то гребневая регрессия совпадает с обычной регрессией, но при $\lambda \to \infty$ коэффициенты регрессии стремятся к нулю.
- Необходимо выбрать хорошее значение λ.

Способ решения оптимизационной задачи

Модифицированное МНК решение гребневой регрессии:

$$\hat{\beta}_\lambda^R = (X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}Y.$$

Решение через сингулярное разложение, где $X = VDU^{\mathrm{T}}$:

- **МНК**

$$\hat{\beta} = \sum_{j=1}^{p} \frac{1}{\sqrt{\lambda_j}} U_j(V_j^{\mathrm{T}}Y);$$

- **МНК гребневой регрессии**

$$\hat{\beta}_\lambda^R = U(D^2 + \lambda I_p)^{-1}DV^{\mathrm{T}}y = \sum_{j=1}^{p} \frac{\sqrt{\lambda_j}}{\lambda_j + \lambda} U_j(V_j^{\mathrm{T}}Y).$$

С помощью сингулярного разложения можно быстро выбирать параметр λ.

Выбор параметра регуляризации

Как выбрать параметр λ:

- выбираем сетку значений λ;
- вычисляем ошибку кросс-проверки для каждого значения λ;
- выбираем λ с наименьшим значением ошибки кросс-проверки;
- перестраиваем модель со всеми наблюдениями с выбранным значением λ.

Вероятностная интерпретация

- $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ имеет априорное распределение $p(\beta)$;
- $f(Y|X, \beta)$ — функция правдоподобия исходных данных.

При фиксированном X апостериорное распределение $p(\beta|X, Y)$ пропорционально

$$f(Y|X, \beta)p(\beta|X) = f(Y|X, \beta)p(\beta).$$

Предполагая, что

1. линейная модель имеет независимые и нормально распределенные ошибки;
2. $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$ для некоторой плотности g.

Если g — плотность $N(0, \lambda)$, то оценка апостериорного максимума β совпадает с решением гребневой регрессии.

Смещение оценки

Пусть $X^{\mathrm{T}}X = \Sigma$ и $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$.
Оценка гребневой регрессии через МНК оценку:

$$\hat{\beta}_\lambda^R = (I_p + \lambda\Sigma^{-1})\hat{\beta}.$$

Оценка гребневой регрессии имеет смещение:

$$\begin{aligned}\mathbb{E}\hat{\beta}_\lambda^R &= \mathbb{E}[(I_p + \lambda\Sigma^{-1})\hat{\beta}] = \\ &= (I_p + \lambda\Sigma^{-1})\beta.\end{aligned}$$

Если $\lambda = 0$, то оценка гребневой регрессии не имеет смещения.

Свойства

- Оценки МНК инварианты относительно умножения признака на константу, то есть значение $f(x_j)\hat{\beta}_j$ не зависит от масштаба j-го признака.
- Инвариант относительно масштаба теряется в случае гребневой регрессии, оценки МНК гребневой регрессии могут сильно измениться при умножении заданного признака на константу.

Вывод: гребневую регрессию нужно использовать после стандартизации признаков.

Проблема:

- в конечную модель входят все начальные признаки;
- если признаков много, то усложняется интерпретация.

Лассо регрессия

Задача минимизации:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \to \min_{\beta},$$

где $\lambda \geq 0$ — неотрицательный параметр регуляризации (tuning parameter).

- Как и в гребневой регрессии $\lambda \sum_{j=1}^{p}|\beta_j|$ мало, когда $\beta_1, \ldots, \beta_p$ близки к нулю.
- При увеличении параметра λ некоторые коэффициенты регрессии становятся равными нулю.
- Как и в гребневой регрессии необходимо выбрать хорошее значение λ.

Способ решения оптимизационной задачи

Задача:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \to \min_{\beta}$$

эквивалента задаче минимизации с ограничением:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 \to \min_{\beta}, \qquad \sum_{j=1}^{p}|\beta_j| \leq s,$$

где параметру λ соответствует параметр s.

- Чем меньше s, тем больше нулевых значений коэффициентов β.
- Значение параметра λ выбирается как в гребневой регрессии с помощью кросс-проверки.

Вероятностная интерпретация

- $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ имеет априорное распределение $p(\beta)$;
- $f(Y|X, \beta)$ — функция правдоподобия исходных данных.

При фиксированном X апостериорное распределение $p(\beta|X, Y)$ пропорционально

$$f(Y|X, \beta)p(\beta|X) = f(Y|X, \beta)p(\beta).$$

Предполагая, что

1. линейная модель имеет независимые и нормально распределенные ошибки;
2. $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$ для некоторой плотности g.

Если g — плотность распределения Лапласса с нулевым средним и параметром масштаба λ, то оценка апостериорного максимума β является решением Лассо.

Свойства

Почему Лассо обнуляет коэффициенты.

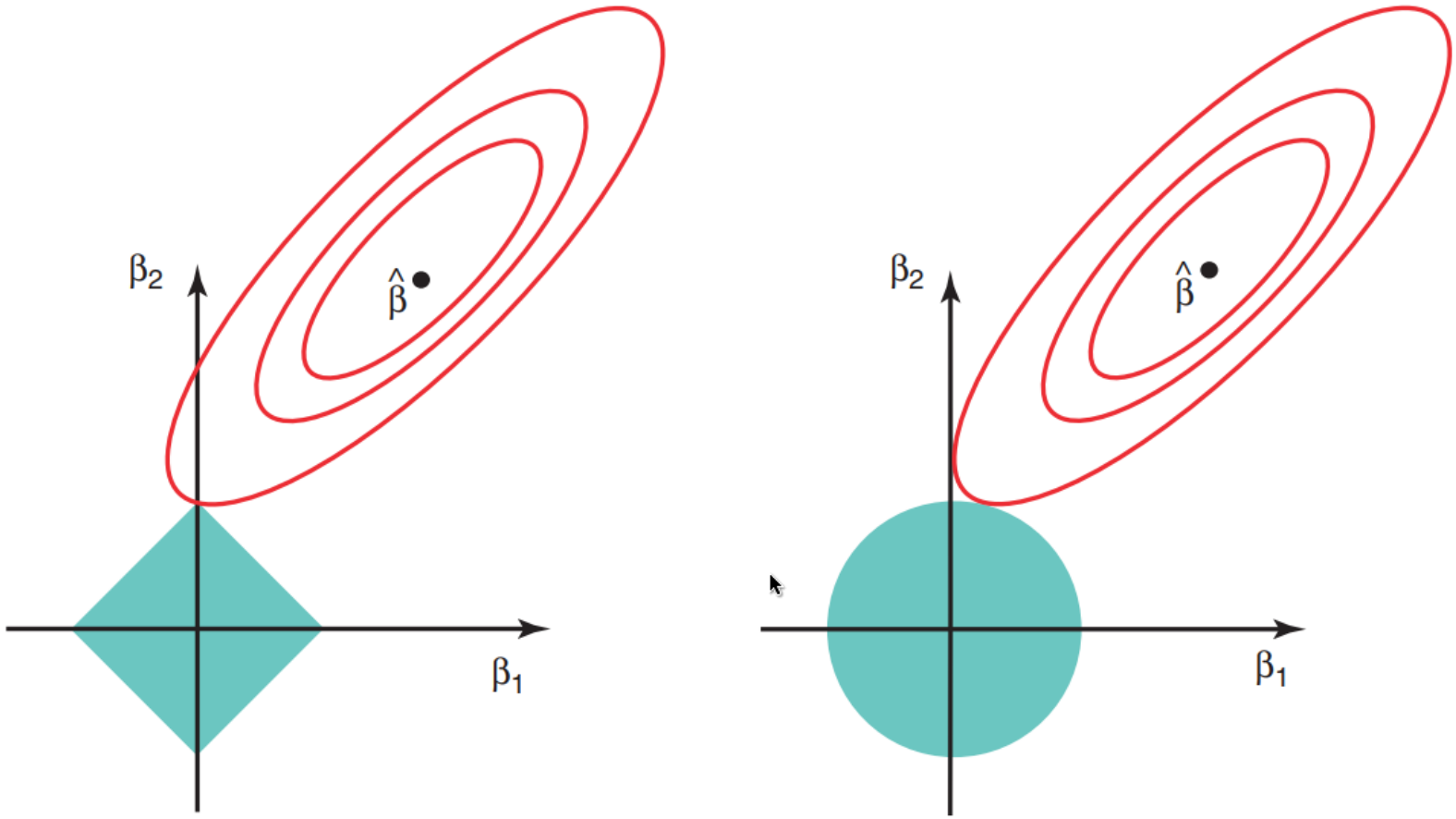


Рис.: Границы ошибки $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2$ и ограничений $\sum_{j=1}^{p}|\beta_j| \leq s$ для Лассо (слева) и $\sum_{j=1}^{p}\beta_j^2 \leq s$ для гребневой регрессии (справа).

Сравнение гребневой регрессии и Лассо

Рассмотрим простой случай, когда $n = p$ и X — диагональная матрица с 1 на диагонали.

МНК:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2.$$

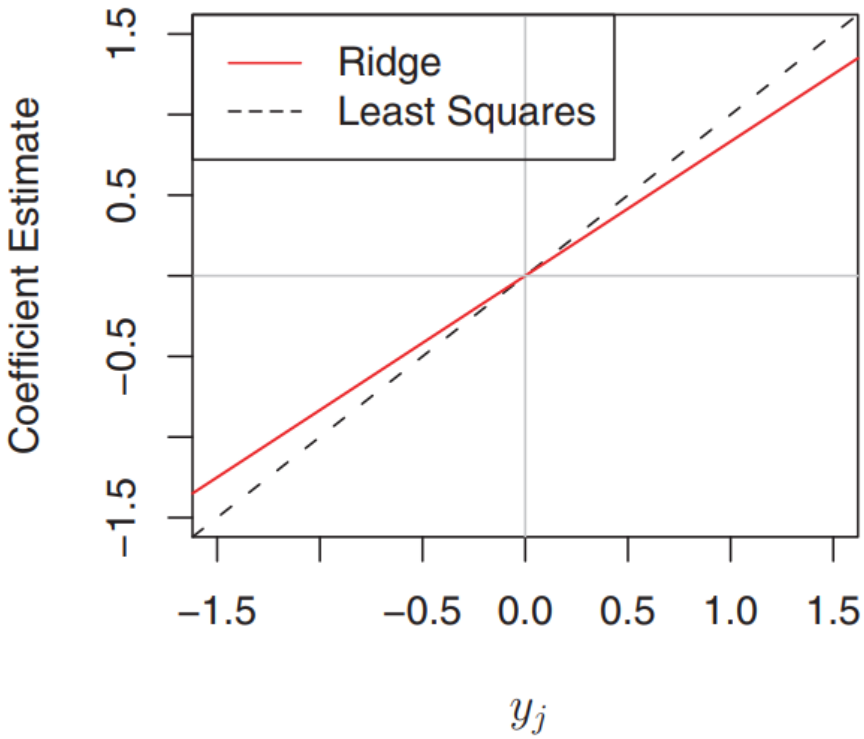
Решение гребневой регрессии:

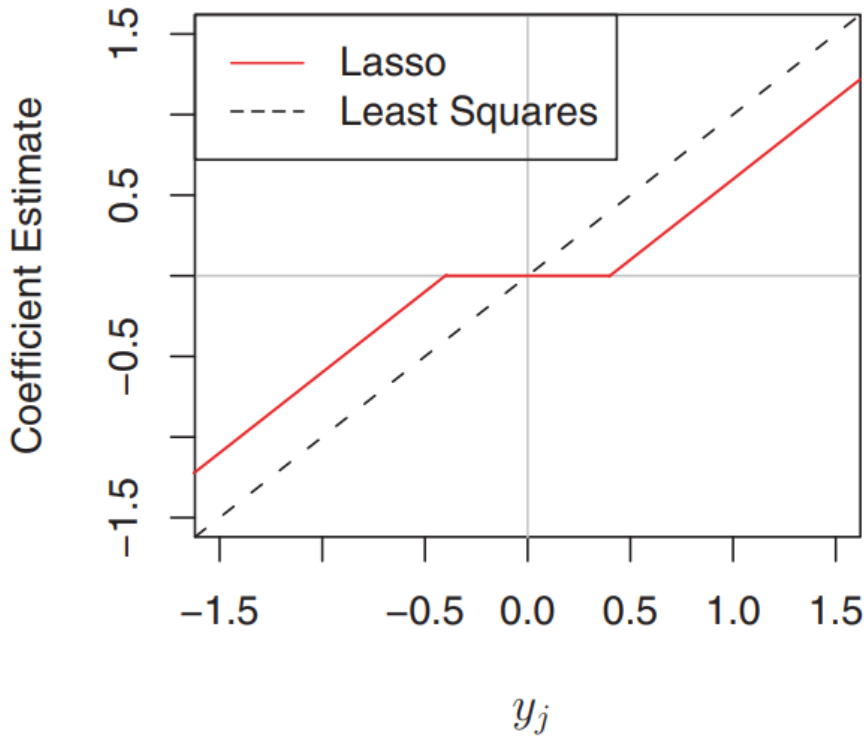
$$\hat{\beta}_\lambda^R = \frac{y_j}{1+\lambda}.$$

Решение Лассо:

$$\hat{\beta}_\lambda^L = \begin{cases} y_j - \lambda/2, & y_j > \lambda/2 \\ y_j + \lambda/2, & y_j < -\lambda/2 \\ 0, & |y_j| \leq \lambda/2. \end{cases}$$

Сравнение гребневой регрессии и Лассо





- Гребневая регрессия уменьшает каждый коэффициент с равной пропорцией;
- Лассо уменьшает значения коэффициентов на одинаковое значение;
- В Лассо, если коэффициент по модулю меньше $\lambda/2$, то его значение становится равным нулю.

Сравнение гребневой регрессии и Лассо

Нельзя выделить ни одну из моделей (Лассо или гребневая регрессия) как лучшую.

Можно ожидать, что

- Лассо будет иметь ошибку меньше, когда в модели мало значимых признаков (коэффициенты при таких признаках будут равны нулю);
- Гребневая регрессия будет иметь ошибку меньше, когда Y будет зависеть от признаков, которые имеют примерно равную значимость.

С помощью кросс-проверки можно определить какой подход лучше.