Санкт-Петербургский государственный университет

Прикладная математика и информатика

Кафедра статистического моделирования

Ширинкина Дарья Андреевна

# Регуляризация в регрессии

Санкт-Петербург
2017

# Содержание

# 1 Задача регрессии

## 1.1 Задача регрессии на генеральном языке

Предполагаем, что существует неизвестная зависимость $f : \mathbb{R}^p \to \mathbb{R}$ между случайными величинами $\eta \in \mathbb{R}$ и $\xi \in \mathbb{R}^p$:

$$\eta = f(\xi) + \epsilon,$$

где $\epsilon$ — случайная величина (ошибка) такая, что $\xi$ и $\epsilon$ независимы. Будем считать также, что $\mathbb{E}\epsilon = 0$ и $\mathbb{E}\epsilon^2 = \sigma^2$.

**Задача:** хотим оценить функцию $f$.

## 1.2 Задача регрессии на выборочном языке

На практике имеем выборку размера $n$ из случайной величины $\xi$: $X = [X_1, \ldots, X_p]$, где $X_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$, $j = 1, \ldots, p$. На этой выборке известны результаты функции $f$, то есть имеем выборку $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ из случайной величины $\eta$ и выборку $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ из случайной величины $\epsilon$.

Пусть $x_i = (x_{i1}, \ldots, x_{ip})$, тогда запишем зависимость $f$ следующим образом:

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

Хотим по имеющейся выборке $X$ оценить функцию $f$.

## 1.3 Линейная регрессия и МНК

Будем называть выборку $(x_1, \ldots, x_n)^{\mathrm{T}}$, участвующую в оценке функции $f$, обучающей выборкой и при $i = 1, \ldots, n$

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

Будем называть выборку $(x'_1, \ldots, x'_k)^{\mathrm{T}}$, не участвующую в оценке функции $f$, тестовой выборкой и при $i = 1, \ldots, n$

$$y'_i = f(x'_i) + \varepsilon'_i.$$

Далее будем предполагать, что рассматриваемая зависимость — линейная. Считаем, что $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ и $X = [X_1, \ldots, X_p]$ — центрированы. Тогда модель многомерной линейной регрессии можно записать следующим образом:

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

где $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ — параметры функции $f$.

Процесс подбора оптимального параметра $\beta$ по обучающей выборке $X$ называют обучением. Для оценки параметра $\beta$ будем минимизировать среднюю квадратичную ошибку (МНК)

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

Решением метода наименьших квадратов (МНК) является вектор

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

и оценкой функции $f$, когда $f$ — линейная функция, является

$$\hat{f}(x_i) = \sum_{j=1}^{p} \hat{\beta}_j x_{ij}.$$

## 1.4 Проблема МНК

Метод наименьших квадратов минимизирует среднюю квадратичную ошибку для обучающей выборки, но это не гарантирует, что на тестовой выборке $\hat{f}$ будет также минимизировать среднюю квадратичную ошибку

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

Когда оценка $\hat{f}$ дает ошибку больше на новых объектах, не участвовавших в обучении, чем на обучающей выборке ($\mathrm{MSE}_{\mathrm{test}} \gg \mathrm{MSE}_{\mathrm{training}}$) говорят, что происходит переобучение.

По мере увеличения количества параметров модели значение $\mathrm{MSE}_{\mathrm{training}}$ будет уменьшаться, а $\mathrm{MSE}_{\mathrm{test}}$ нет. Это происходит потому что метод обучения будет сильно усложняться и пытаться объяснить то, что вызвано случайностью, а не свойствами функции $f$.

Рассмотрим от чего зависит $\mathrm{MSE}_{\mathrm{test}}$ на генеральном языке. Пусть

- $x'_i$ — реализация случайной величина из тестовой выборки;
- $y'_i = f(x'_i) + \varepsilon'_i$ — известное значение.

Тогда математическое ожидание квадрата ошибки для $x'_i$

$$\mathbb{E}(y'_i - \hat{f}(x'_i))^2 = Var(\hat{f}(x'_i)) + (Bias(\hat{f}(x'_i)))^2 + Var(\varepsilon'_i).$$

Таким образом, ошибка для объекта, не участвовавшего в обучении, зависит от дисперсии оценки функции $f$ и квадрата смещения. Дисперсия оценки определяет то количество, на которое изменится $\hat{f}$, если бы мы получали эту оценку с использованием другого набора данных. Мы хотим, чтобы оценка $\hat{f}$ не менялась сильно на разных обучающих выборках. Смещение $\hat{f}$ характеризует ошибку, возникающую при аппроксимации реальной сложной функции $f$ более простой моделью.

Как правило, при увеличении сложности метода (увеличение числа параметров) дисперсия будет увеличиваться, а смещение будет уменьшаться. За счет введения в оценку небольшого смещения, можно значительно уменьшить ее дисперсию и тем самым уменьшить $\mathrm{MSE}_{\mathrm{test}}$.

В случае многомерной линейной регрессии оценка $\hat{f}$ не имеет смещения, так как нет смещения у оценки $\hat{\beta}$. Но чем больше дисперсия оценки $\hat{\beta}$

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2(X^{\mathrm{T}}X)^{-1},$$

тем больше дисперсия $\hat{f}$. Когда матрица $X$ близка к вырожденной, дисперсия $\hat{\beta}$ становится большой и $\mathrm{MSE}_{\mathrm{test}}$ увеличивается.

# 2 Гребневая регрессия

## 2.1 Задача гребневой регрессии

Чтобы ввести смещение оценке параметра $\beta$ добавим в $\text{MSE}_{\text{training}}$ множитель, который будет ограничивать параметр $\beta$. Тогда задача минимизации будет выглядеть следующим образом:

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

где $\lambda \geq 0$ — неотрицательный параметр регуляризации (tuning parameter). Такая модель минимизирует среднюю квадратичную ошибку, но при этом второй множитель $\lambda\sum_{j=1}^{p}\beta_j^2$ мал, когда параметры $\beta_1, \ldots, \beta_p$ близки к нулю.

При $\lambda = 0$, то гребневая регрессия совпадает с обычной регрессией, но при $\lambda \to \infty$ коэффициенты регрессии стремятся к нулю. Таким образом, для каждого значения $\lambda$ будем получать свой оптимальный набор параметров $\beta_1, \ldots, \beta_p$, поэтому для уменьшения $\text{MSE}_{\text{test}}$ важно подобрать хорошее значение параметра $\lambda$.

## 2.2 Решение задачи минимизации и выбор параметра регуляризации

Решение гребневой регрессии можно выписать в явном виде:

$$\hat{\beta}_\lambda^R = (X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}Y. \tag{2.1}$$

Запишем модифицированное МНК решение гребневой регрессии через сингулярное разложение матрицы $X = VDU^{\mathrm{T}}$, где матрицы $V$ размера $n \times p$ и $U$ размера $p \times p$ — ортогональные, а матрица $D = \text{diag}(\lambda_1, \ldots, \lambda_p)$ — диагональная, на диагонали которой стоят собственные числа матрицы $X^{\mathrm{T}}X$.

Тогда решение МНК:

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

Решение гребневой регрессии:

$$\hat{\beta}_\lambda^R = U(D^2 + \lambda I_p)^{-1}DV^{\mathrm{T}}y = \sum_{j=1}^{p}\frac{\sqrt{\lambda_j}}{\lambda_j + \lambda}U_j(V_j^{\mathrm{T}}Y). \tag{2.2}$$

При этом оценка функции $f$ для выборки $X$ запишется следующим образом:

$$X\hat{\beta}_\lambda^R = VDU^{\mathrm{T}}\hat{\beta}_\lambda = V\text{diag}(\frac{\lambda_j}{\lambda_j + \lambda})V^{\mathrm{T}}Y = \sum_{j=1}^{p}\frac{\lambda_j}{\lambda_j + \lambda}V_j(V_j^{\mathrm{T}}Y). \tag{2.3}$$

Такой способ записи решения задачи минимизации удобен для выбора параметра $\lambda$. Выражение (2.1) можно переписать, используя (2.2) и (2.3). Таким образом, необходимо один раз произвести сингулярное разложение матрицы $X$, а затем несложным образом вычислять функционал (2.1) для интересующих значений параметра $\lambda$.

Как выбрать параметр $\lambda$:

1. выбираем сетку значений $\lambda$;

2. вычисляем ошибку кросс-проверки для каждого значения $\lambda$, используя запись (2.1) через сингулярное разложение;

3. выбираем $\lambda$ с наименьшим значением ошибки кросс-проверки;

4. перестраиваем модель со всеми наблюдениями с выбранным значением $\lambda$.

## 2.3 Свойства и проблемы

**Свойства:**

1. Рассмотрим вероятностную интерпретацию гребневой регрессии. Предполагая, что

   - $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ имеет априорное распределение $p(\beta)$;
   - $f(Y|X, \beta)$ — функция правдоподобия исходных данных;
   - линейная модель имеет независимые и нормально распределенные ошибки;
   - $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$ для некоторой плотности $g$.

   По теореме Байеса при фиксированном $X$ апостериорное распределение $p(\beta|X, Y)$ пропорционально $f(Y|X, \beta)p(\beta|X) = f(Y|X, \beta)p(\beta)$ и если $g$ — плотность $N(0, \lambda)$, то оценка апостериорного максимума $\beta$ совпадает с решением гребневой регрессии.

2. Так как решение гребневой регрессии можно выписать в явном виде (2.2), то покажем, что оценка $\beta$ имеет смещение. Пусть $X^{\mathrm{T}}X = \Sigma$ и $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, тогда

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

   Посчитаем математическое ожидание

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

   При $\lambda = 0$ оценка совпадает с МНК оценкой и не имеет смещения.

**Проблемы:**

1. Оценки МНК инварианты относительно умножения признака на константу, то есть значение $f(x_j)\hat{\beta}_j$ не зависит от масштаба $j$-го признака. В случае гребневой регрессии инвариант относительно масштаба теряется, оценки МНК гребневой регрессии могут сильно измениться при умножении заданного признака на константу. Поэтому гребневую регрессию нужно использовать после стандартизации признаков.

2. В конечную модель входят все начальные признаки, если признаков много, то усложняется интерпретация.

# 3 Лассо

## 3.1 Задача минимизации Лассо

Метод Лассо, как и гребневая регрессия, вводит смещение у оценки $\beta$, при этом в конечную модель Лассо могут входить не все начальные признаки, что решает проблему гребневой регрессии. Это происходит за счет другого множителя в задачи минимизации, который регулирует коэффициенты $\beta_1, \ldots, \beta_p$, однако из-за этого усложняется решение задачи. Задача минимизации Лассо:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \to \min_{\beta}, \tag{3.1}$$

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 $\lambda$ некоторые коэффициенты регрессии становятся равными нулю. Как и в гребневой регрессии необходимо выбрать хорошее значение $\lambda$.

## 3.2 Решение задачи минимизации и выбор параметра регуляризации

Задача минимизации (3.1) 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 $\lambda$ соответствует параметр $s$. Чем меньше $s$, тем больше нулевых значений коэффициентов $\beta$. Такую задачу можно решить, так как это задача квадратичного программирования.

Значение параметра $\lambda$ выбирается как в гребневой регрессии с помощью кросс-проверки.

## 3.3 Свойства

**Свойства:**

1. Рассмотрим вероятностную интерпретацию Лассо. Предполагая, что

   - $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ имеет априорное распределение $p(\beta)$;
   - $f(Y|X, \beta)$ — функция правдоподобия исходных данных;
   - линейная модель имеет независимые и нормально распределенные ошибки;
   - $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$ для некоторой плотности $g$.

По теореме Байеса при фиксированном $X$ апостериорное распределение $p(\beta|X,Y)$ пропорционально $f(Y|X,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta)$ и если $g$ — плотность распределения Лапласса с нулевым средним и параметром масштаба $\lambda$, то оценка апостериорного максимума $\beta$ является решением Лассо.

2. Лассо обнуляет некоторые коэффициенты $\beta_1, \ldots, \beta_p$. Задачу минимизации гребневой регрессии можно также рассмотреть в виде

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

Рассмотрим простой случай, когда $p = 2$. Тогда выражение $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2$ будет представлять собой эллипс, центр которого находится в точке $\hat{\beta}$ (см. рис. 1). Предположим, что центр эллипса не удовлетворяет ограничениям $\sum_{j=1}^{p}\beta_j^2 \le s$ и $\sum_{j=1}^{p}|\beta_j| \le s$, то есть лежит вне круга в случае гребневой регрессии и вне ромба в случае Лассо. Тогда решения задач минимизации будет лежать на границе возможных значений. Данная ситуация изображена на рис. 1, можно заметить, что для Лассо существует гораздо больше различных эллипсов, которые пересекались бы с ромбом (ограничениями) таким образом, чтобы один из коэффициентов был равен нулю.

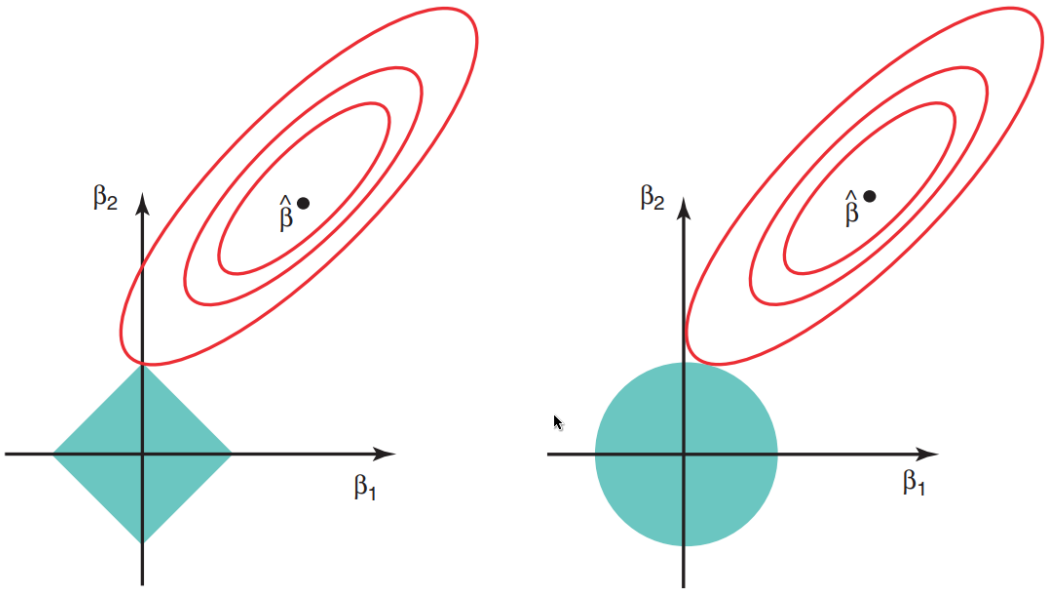


**Рис. 1.** Границы ошибки $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2$ и ограничений $\sum_{j=1}^{p}|\beta_j| \le s$ для Лассо (слева) и $\sum_{j=1}^{p}\beta_j^2 \le s$ для гребневой регрессии (справа).

# 4 Сравнение гребневой регрессии и Лассо

Рассмотрим простой случай, когда $n = p$ и $X$ — диагональная матрица с 1 на диагонали. Тогда задача минимизации для обычной регрессии принимает вид

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

Решение МНК:

$$\hat{\beta}_j = y_j.$$

Решение гребневой регрессии будет принимать вид:

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

Решение Лассо:

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

Рисунок 4 иллюстрирует разный подход к регулированию коэффициентов $\beta_1, \ldots, \beta_p$. Гребневая регрессия уменьшает каждый коэффициент с равной пропорцией. Лассо уменьшает значения коэффициентов на одинаковое значение, при этом если коэффициент по модулю меньше $\lambda/2$, то его значение становится равным нулю. В случае более общей матрицы $X$ ситуация немного сложнее, но принцип регуляризации сохраняется.

В целом нельзя выделить ни одну из моделей (Лассо или гребневая регрессия) как лучшую. С помощью кросс-проверки можно определить какой подход лучше.

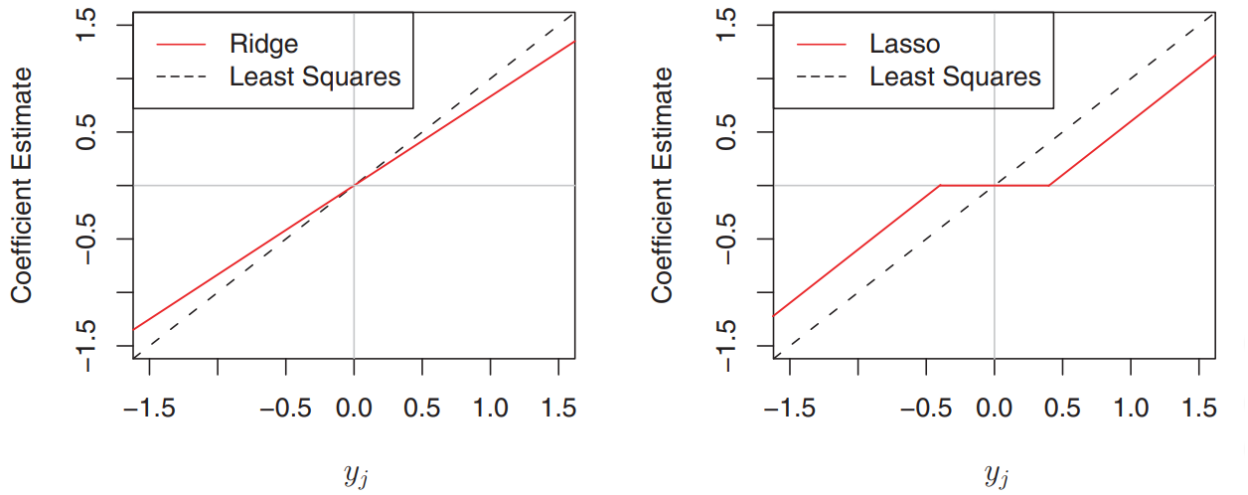


**Рис. 2.** Случай $n = p$. Слева сравнение МНК и гребневой регрессии. Справа сравнение МНК и Лассо.