

Интеллектуальный анализ данных (Data Mining)

Шорохов С.Г.

кафедра математического моделирования и искусственного интеллекта

Лекция 7. Регрессия





Пусть даны независимые переменные (признаки) X_1, X_2, \dots, X_d и зависимая переменная (отклик) Y , тогда целью **регрессии** является прогнозирование значения Y на основе значений X_1, X_2, \dots, X_d , т.е. цель состоит в том, чтобы определить функцию регрессии f , такую, что

$$Y = f(\mathbf{X}) + \varepsilon = f(X_1, X_2, \dots, X_d) + \varepsilon,$$

где ε – случайная ошибка, которая предполагается независимой от многомерной случайной величины $\mathbf{X} = (X_1, X_2, \dots, X_d)^T \in \mathbb{R}^d$, причем $\mathbb{E}[\varepsilon] = 0$.

Выражение для Y состоит из двух слагаемых, одно из которых зависит от переменных X_1, X_2, \dots, X_d , а другое зависит от ошибки, независимой от переменных X_1, X_2, \dots, X_d . Слагаемое ошибки соответствует неустранимой неопределенности, присущей Y , а также, возможно, влиянию ненаблюдаемых, скрытых (латентных) переменных. Таким образом, функция регрессии f может быть построена как условное математическое ожидание

$$f(x_1, \dots, x_d) = \mathbb{E}[Y \mid X_1 = x_1, \dots, X_d = x_d]$$



В **линейной регрессии** функция регрессии f предполагается линейной по признакам $\mathbf{X} = (X_1, X_2, \dots, X_d)^T$, т.е.

$$f(\mathbf{X}) = \beta + \omega_1 X_1 + \omega_2 X_2 + \dots + \omega_d X_d = \beta + \sum_{i=1}^d \omega_i X_i = \beta + \boldsymbol{\omega}^T \mathbf{X},$$

где β – истинное (неизвестное) смещение (bias), ω_i – истинный (неизвестный) коэффициент регрессии или вес для признака X_i , $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_d)^T$ – d -мерный вектор истинных весов.

Линейная функция f определяет гиперплоскость $f(\mathbf{X}) = 0$ в пространстве признаков \mathbb{R}^d , причем вектор весов $\boldsymbol{\omega}$ ортогонален (перпендикулярен) гиперплоскости $f(\mathbf{X}) = 0$, а смещение β задает точки пересечения гиперплоскости с осями координат пространства признаков. Функция регрессии f полностью определяется $d + 1$ параметрами β и ω_i для $i = 1, \dots, d$.

Наиболее распространенным подходом к оценке (прогнозированию) параметров линейной регрессии (коэффициентов смещения β и регрессии $\boldsymbol{\omega}$) является **метод наименьших квадратов**.



Задача оценки параметров линейной регрессии заключается в подборе таких значений коэффициентов смещения b и регрессии $\mathbf{w} = (w_1, \dots, w_d)^T$, чтобы значения функции регрессии $\hat{y} = b + \mathbf{w}^T \mathbf{x}$ были максимально близки к имеющимся значениям отклика. Суть **метода наименьших квадратов** (МНК) заключается в выборе в качестве «меры близости» суммы квадратов отклонений значений функции регрессии от значений отклика.

Пусть обучающие данные \mathbf{D} содержат d -мерные векторы значений признаков \mathbf{x}_i и значения откликов y_i (для $i = 1, \dots, n$), тогда требуется определить параметры b и \mathbf{w} , минимизирующие сумму квадратов остаточных ошибок (SSE или Sum of Squared Errors)

$$\min_{b, \mathbf{w}} SSE = \min_{b, \mathbf{w}} \sum_{i=1}^n \varepsilon_i^2 = \min_{b, \mathbf{w}} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min_{b, \mathbf{w}} \sum_{i=1}^n (y_i - b - \mathbf{w}^T \mathbf{x}_i)^2$$

Основные подходы к минимизации SSE:

- продифференцировать SSE по неизвестным параметрам, приравнять производные к нулю и решить полученную систему уравнений
- использовать численные методы минимизации



В **парной** (bivariate) **регрессии** обучающие данные \mathbf{D} содержат один признак $\mathbf{X} = (x_1, x_2, \dots, x_n)^T$ вместе с откликом $\mathbf{Y} = (y_1, y_2, \dots, y_n)^T$, а линейная функция регрессии зависит от двух скалярных параметров b и w : $\hat{y} = b + w x$.

Остаточная ошибка для точки x_i равна $\varepsilon_i = y_i - \hat{y}_i = y_i - b - w x_i$ и параметры парной регрессии b , w определяются из задачи минимизации:

$$\min_{b, w} SSE = \min_{b, w} \sum_{i=1}^n \varepsilon_i^2 = \min_{b, w} \sum_{i=1}^n (y_i - b - w x_i)^2.$$

Дифференцируем SSE по b и приравниваем результат к нулю:

$$\frac{\partial}{\partial b} SSE = -2 \sum_{i=1}^n (y_i - b - w x_i) = 0 \Rightarrow b = \frac{1}{n} \sum_{i=1}^n y_i - w \frac{1}{n} \sum_{i=1}^n x_i.$$

Отсюда получаем следующее выражение для коэффициента смещения b

$$b = \mu_{\mathbf{Y}} - w \mu_{\mathbf{X}}, \mu_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n x_i, \mu_{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^n y_i.$$



Дифференцируем SSE по w и получаем:

$$\frac{\partial}{\partial w} SSE = -2 \sum_{i=1}^n x_i (y_i - b - w x_i) = 0 \Rightarrow \sum_{i=1}^n x_i y_i - b \sum_{i=1}^n x_i - w \sum_{i=1}^n x_i^2 = 0 \Rightarrow$$

$$\sum_{i=1}^n x_i y_i - \mu_Y \sum_{i=1}^n x_i + w \mu_X \sum_{i=1}^n x_i - w \sum_{i=1}^n x_i^2 = 0 \Rightarrow$$

$$w = \frac{\sum_{i=1}^n x_i y_i - n \mu_X \mu_Y}{\sum_{i=1}^n x_i^2 - n \mu_X^2}$$

Коэффициент регрессии w также может быть выражен через ковариацию \mathbf{X} и \mathbf{Y} и дисперсию \mathbf{X} :

$$w = \frac{\sum_{i=1}^n (x_i - \mu_X) (y_i - \mu_Y)}{\sum_{i=1}^n (x_i - \mu_X)^2} = \frac{\sigma_{\mathbf{XY}}}{\sigma_{\mathbf{X}}^2} = \frac{\text{cov}(\mathbf{X}, \mathbf{Y})}{\text{var}(\mathbf{X})}$$



Итак, в парной регрессии

$$\hat{y} = b + w x$$

оценка параметров регрессии b и w по обучающим данным

$$\mathbf{D} = \left\{ \mathbf{X} = (x_1, x_2, \dots, x_n)^T, \mathbf{Y} = (y_1, y_2, \dots, y_n)^T \right\}$$

производится по формулам

$$b = \mu_{\mathbf{Y}} - \frac{\sigma_{\mathbf{XY}}}{\sigma_{\mathbf{X}}^2} \mu_{\mathbf{X}}, w = \frac{\sigma_{\mathbf{XY}}}{\sigma_{\mathbf{X}}^2},$$

где

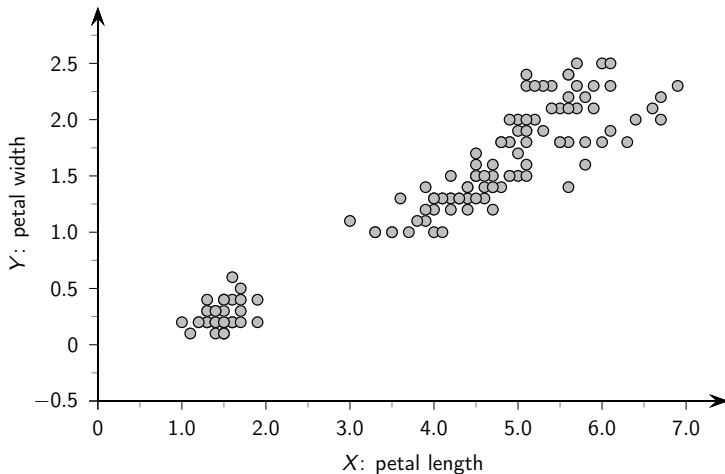
$$\mu_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n x_i, \mu_{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^n y_i,$$

$$\sigma_{\mathbf{XY}} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\mathbf{X}})(y_i - \mu_{\mathbf{Y}}),$$

$$\sigma_{\mathbf{X}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\mathbf{X}})^2.$$



В наборе данных Ирисы длина лепестка (petal length) X рассматривается как независимая переменная и ширина лепестка (petal width) Y как отклик. Исследуем зависимость ширины лепестка Y от длины лепестка X .





Средние значения для переменных X и Y равны

$$\mu_X = \frac{1}{150} \sum_{i=1}^{150} x_i = \frac{563.8}{150} = 3.7587$$

$$\mu_Y = \frac{1}{150} \sum_{i=1}^{150} y_i = \frac{179.8}{150} = 1.1987$$

Дисперсии X и Y и ковариация X и Y равны

$$\sigma_X^2 = \frac{1}{150} \sum_{i=1}^{150} (x_i - \mu_X)^2 = 3.0924$$

$$\sigma_Y^2 = \frac{1}{150} \sum_{i=1}^{150} (y_i - \mu_Y)^2 = 0.5785$$

$$\sigma_{XY} = \frac{1}{150} \sum_{i=1}^{150} (x_i - \mu_X) \cdot (y_i - \mu_Y) = 1.2877$$



Предполагая линейную связь между откликом Y и независимой переменной X , вычислим следующие коэффициенты регрессии (наклона) и смещения

$$w = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{1.2877}{3.0924} = 0.4164$$

$$b = \mu_Y - w \mu_X = 1.1987 - 0.4164 \cdot 3.7587 = -0.3665$$

Таким образом, полученная функция линейной регрессии имеет вид

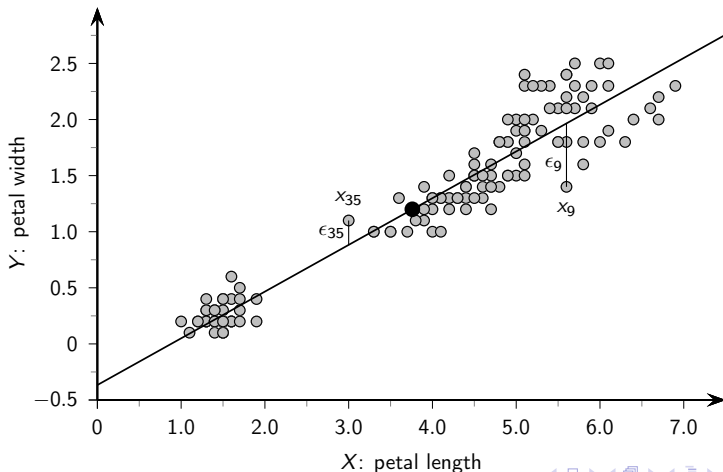
$$\hat{y} = -0.3665 + 0.4164 x$$

Сумма квадратов остаточных ошибок SSE вычисляется следующим образом:

$$SSE = \sum_{i=1}^{150} \varepsilon_i^2 = \sum_{i=1}^{150} (y_i - \hat{y}_i)^2 = 6.343$$



Изобразим на плоскости линию регрессии, отражающую зависимость ширины лепестка Y от длины лепестка X . Сплошной черный кружок показывает среднюю точку, остаточная ошибка показана для двух точек x_9 и x_{35} .





В **множественной регрессии** (multiple regression) несколько независимых признаков X_1, X_2, \dots, X_d и один отклик Y . Обучающая выборка $\mathbf{D} \in \mathbb{R}^{n \times d}$ содержит n точек $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ в d -мерном пространстве вместе с соответствующими наблюдаемыми значениями откликов y_i .

Вместо того, чтобы рассматривать смещение b отдельно от весов w_i , можно ввести новый атрибут X_0 , значение которого всегда равно единице ($x_{i0} = 1$).

Тогда прогнозируемое значение отклика для расширенной $(d + 1)$ -мерной точки $\tilde{\mathbf{x}}_i$ можно записать как

$$\hat{y}_i = w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + \dots + w_d x_{id} = \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_i,$$

где $\tilde{\mathbf{w}} = (w_0, w_1, w_2, \dots, w_d)^T$, $\tilde{\mathbf{x}}_i = (x_{i0}, x_{i1}, x_{i2}, \dots, x_{id})^T$. Таким образом, для обучающей выборки \mathbf{D} вектор прогнозируемых значений откликов равен

$$\hat{\mathbf{Y}} = \tilde{\mathbf{D}} \tilde{\mathbf{w}},$$

где $\tilde{\mathbf{D}}$ – дополненная обучающая выборка, состоящая из расширенных точек $\tilde{\mathbf{x}}_i = (1, x_{i1}, x_{i2}, \dots, x_{id})^T$.



Задача множественной регрессии состоит в том, чтобы найти наиболее подходящую линейную функцию регрессии $f(\tilde{\mathbf{x}}) = \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}$, определяемую расширенным вектором весов $\tilde{\mathbf{w}}$, которая минимизирует ошибку SSE:

$$\begin{aligned} SSE &= \sum_{i=1}^n \varepsilon_i^2 = \|\varepsilon\|^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) = \\ &= \mathbf{Y}^T \mathbf{Y} - 2\mathbf{Y}^T \hat{\mathbf{Y}} + \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} = \mathbf{Y}^T \mathbf{Y} - 2\mathbf{Y}^T (\tilde{\mathbf{D}}\tilde{\mathbf{w}}) + (\tilde{\mathbf{D}}\tilde{\mathbf{w}})^T (\tilde{\mathbf{D}}\tilde{\mathbf{w}}) = \\ &= \mathbf{Y}^T \mathbf{Y} - 2\tilde{\mathbf{w}}^T (\tilde{\mathbf{D}}^T \mathbf{Y}) + \tilde{\mathbf{w}}^T (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}) \tilde{\mathbf{w}} \end{aligned}$$

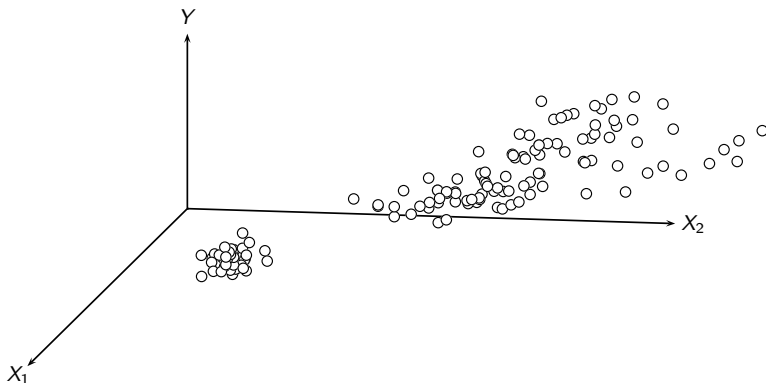
Дифференцируя SSE по $\tilde{\mathbf{w}}$ и приравнявая результат к нулю, получим, что оптимальный вектор весов $\tilde{\mathbf{w}}$ множественной регрессии задается формулой

$$\tilde{\mathbf{w}} = (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T \mathbf{Y},$$

где $\tilde{\mathbf{D}} = \begin{pmatrix} \mathbf{1} & \mathbf{D} \end{pmatrix}$ – дополненная обучающая выборка (матрица размерами $n \times (d+1)$), \mathbf{Y} – вектор значений откликов для точек \mathbf{D} .



Рассматривая независимые признаки длину чашелистика X_1 (sepal length) и длину лепестка X_2 (petal length), а также ширину лепестка (petal width) как отклик Y , исследуем множественную регрессию в наборе данных Iris (количество точек $n = 150$).





Имеем $X_0 = \mathbf{1}_{150}$ и $\tilde{\mathbf{D}} \in \mathbb{R}^{150 \times 3}$ (всего три признака X_0, X_1, X_2), тогда

$$\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} = \begin{pmatrix} 150.0 & 876.50 & 563.80 \\ 876.5 & 5223.85 & 3484.25 \\ 563.8 & 3484.25 & 2583.00 \end{pmatrix}, \tilde{\mathbf{D}}^T \mathbf{Y} = \begin{pmatrix} 179.80 \\ 1127.65 \\ 868.97 \end{pmatrix}$$

$$(\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} = \begin{pmatrix} 0.793 & -0.176 & 0.064 \\ -0.176 & 0.041 & -0.017 \\ -0.017 & 0.064 & 0.009 \end{pmatrix}$$

Дополненный вектор весов $\tilde{\mathbf{w}}$ вычисляется как

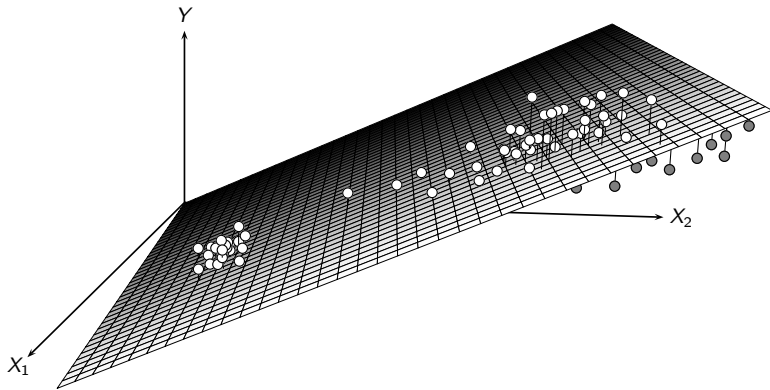
$$\tilde{\mathbf{w}} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T \mathbf{Y} = \begin{pmatrix} -0.014 \\ -0.082 \\ 0.45 \end{pmatrix}$$

Тогда $b = w_0 = -0.014$ и уравнение множественной регрессии имеет вид

$$\hat{y} = -0.014 - 0.082 x_1 + 0.45 x_2$$



На рисунке показана построенная гиперплоскость и остаточная ошибка для каждой точки. Положительные ошибки (т.е. $\varepsilon_i > 0$ или $\hat{y}_i > y_i$) белые, а отрицательные ошибки (т.е. $\varepsilon_i < 0$ или $\hat{y}_i < y_i$) серые. Значение ошибки SSE для модели множественной регрессии равно 6.18.





Если размерность d пространства признаков высока, то задача вычисления обратной матрицы к матрице $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$ размерности $(d+1) \times (d+1)$ (нецентрированной матрице рассеяния) является вычислительно сложной.

Для того, чтобы облегчить вычисление обратной матрицы $(\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1}$, можно использовать т.н. ортогонализацию Грама-Шмидта для матрицы $\tilde{\mathbf{D}}$, в результате которой получаем QR-разложение $\tilde{\mathbf{D}} = \mathbf{Q} \mathbf{R}$, где по построению \mathbf{Q} – матрица размерами $n \times (d+1)$ с ортогональными столбцами вида

$$\mathbf{Q} = \begin{pmatrix} | & | & \dots & | \\ U_0 & U_1 & \dots & U_d \\ | & | & & | \end{pmatrix}$$

и \mathbf{R} – верхнетреугольная матрица размерами $(d+1) \times (d+1)$ вида

$$\mathbf{R} = \begin{pmatrix} 1 & p_{10} & p_{20} & \dots & p_{d0} \\ 0 & 1 & p_{21} & \dots & p_{d1} \\ 0 & 0 & 1 & \dots & p_{d2} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$



Тогда для $\tilde{\mathbf{D}}$ имеем представление

$$\underbrace{\begin{pmatrix} | & | & & | \\ X_0 & X_1 & \dots & X_d \\ | & | & & | \end{pmatrix}}_{\tilde{\mathbf{D}}} = \underbrace{\begin{pmatrix} | & | & & | \\ U_0 & U_1 & \dots & U_d \\ | & | & & | \end{pmatrix}}_{\mathbf{Q}} \cdot \underbrace{\begin{pmatrix} 1 & p_{10} & \dots & p_{d0} \\ 0 & 1 & \dots & p_{d1} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}}_{\mathbf{R}}$$

и в силу ортогональности столбцов матрицы \mathbf{Q} имеем

$$\mathbf{Q}^T \mathbf{Q} = \mathbf{\Delta} = \begin{pmatrix} \|U_0\|^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \|U_d\|^2 \end{pmatrix}$$

Отсюда выводим уравнение для определения вектора весов $\tilde{\mathbf{w}}$:

$$\begin{aligned} \tilde{\mathbf{w}} &= (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T \mathbf{Y} \Rightarrow \tilde{\mathbf{D}}^T \tilde{\mathbf{D}} \tilde{\mathbf{w}} = \tilde{\mathbf{D}}^T \mathbf{Y} \Rightarrow \mathbf{R}^T (\mathbf{Q}^T \mathbf{Q}) \mathbf{R} \tilde{\mathbf{w}} = \mathbf{R}^T \mathbf{Q}^T \mathbf{Y} \Rightarrow \\ &\Rightarrow \mathbf{R}^T \mathbf{\Delta} \mathbf{R} \tilde{\mathbf{w}} = \mathbf{R}^T \mathbf{Q}^T \mathbf{Y} \Rightarrow \mathbf{\Delta} \mathbf{R} \tilde{\mathbf{w}} = \mathbf{Q}^T \mathbf{Y} \Rightarrow \mathbf{R} \tilde{\mathbf{w}} = \mathbf{\Delta}^{-1} \mathbf{Q}^T \mathbf{Y} \end{aligned}$$

Система уравнений $\mathbf{R} \tilde{\mathbf{w}} = \mathbf{\Delta}^{-1} \mathbf{Q}^T \mathbf{Y}$ решается обратной подстановкой.



Алгоритм основан на QR-факторизации, которая выражает матрицу $\tilde{\mathbf{D}}$ как произведение двух матриц: ортогональной матрицы \mathbf{Q} и верхней (или правой) треугольной матрицы \mathbf{R} .

Multiple-Regression (\mathbf{D}, \mathbf{Y}):

- 1 $\tilde{\mathbf{D}} \leftarrow \begin{pmatrix} \mathbf{1} & \mathbf{D} \end{pmatrix}$ // дополненные входные данные с $X_0 = \mathbf{1} \in \mathbb{R}^n$
- 2 $\{\mathbf{Q}, \mathbf{R}\} \leftarrow \text{QR-факторизация}(\tilde{\mathbf{D}})$ // $\mathbf{Q} = (U_0 \ U_1 \ \dots \ U_d)$
- 3 $\Delta \leftarrow \begin{pmatrix} \|U_0\|^2 & 0 & \dots & 0 \\ 0 & \|U_1\|^2 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \|U_d\|^2 \end{pmatrix}$ // квадраты норм по диагонали
- 4 $\Delta^{-1} \leftarrow \begin{pmatrix} \frac{1}{\|U_0\|^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\|U_1\|^2} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \frac{1}{\|U_d\|^2} \end{pmatrix}$ // обратные квадраты норм
- 5 $\mathbf{R}\tilde{\mathbf{w}} \leftarrow \Delta^{-1}\mathbf{Q}^T\mathbf{Y}$ // веса $\tilde{\mathbf{w}}$ находятся обратной подстановкой
- 6 $\hat{\mathbf{Y}} \leftarrow \mathbf{Q}\Delta^{-1}\mathbf{Q}^T\mathbf{Y}$ // прогноз отклика без определения весов $\tilde{\mathbf{w}}$



Найдем зависимость ширины лепестка Y от длины чашелистника X_1 и длины лепестка X_2 для набора данных Ирисы с $n = 150$ точками.

Ортогонализация Грама–Шмидта приводит к следующей QR-факторизации:

$$\underbrace{\begin{pmatrix} | & | & | \\ X_0 & X_1 & X_2 \\ | & | & | \end{pmatrix}}_{\tilde{D}} = \underbrace{\begin{pmatrix} | & | & | \\ U_0 & U_1 & U_2 \\ | & | & | \end{pmatrix}}_{Q} \cdot \underbrace{\begin{pmatrix} 1 & 5.843 & 3.759 \\ 0 & 1 & 1.858 \\ 0 & 0 & 1 \end{pmatrix}}_{R},$$

где $Q \in \mathbb{R}^{150 \times 3}$ и матрицы Δ и Δ^{-1} равны

$$\Delta = \begin{pmatrix} 150.0 & 0 & 0 \\ 0 & 102.17 & 0 \\ 0 & 0 & 111.35 \end{pmatrix}, \quad \Delta^{-1} = \begin{pmatrix} 0.00667 & 0 & 0 \\ 0 & 0.00979 & 0 \\ 0 & 0 & 0.00898 \end{pmatrix}$$



Используем обратную подстановку для определения $\tilde{\mathbf{w}}$:

$$\mathbf{R}\tilde{\mathbf{w}} = \mathbf{\Delta}^{-1}\mathbf{Q}^T\mathbf{Y} \text{ или } \begin{pmatrix} 1 & 5.843 & 3.759 \\ 0 & 1 & 1.858 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 1.1987 \\ 0.7538 \\ 0.4499 \end{pmatrix}$$

Обратная подстановка начинается с веса w_2 , потом определяется w_1 и, наконец, w_0 :

$$w_2 = 0.4499,$$

$$w_1 + 1.858 w_2 = 0.7538 \Rightarrow$$

$$w_1 = 0.7538 - 0.8358 = -0.082,$$

$$w_0 + 5.843 w_1 + 3.759 w_2 = 1.1987 \Rightarrow$$

$$w_0 = 1.1987 + 0.4786 - 1.6911 = -0.0139$$

Модель множественной регрессии в наборе Ирисы построена в виде:

$$\hat{y} = -0.0139 - 0.082 x_1 + 0.4499 x_2$$



Вместо использования подхода на основе QR-факторизации для точного решения задачи множественной регрессии можно использовать алгоритм стохастического градиентного спуска (SGD). Градиент целевой функции SSE по весам $\tilde{\mathbf{w}}$ задается как

$$\nabla_{\tilde{\mathbf{w}}} = \frac{\partial}{\partial \tilde{\mathbf{w}}} SSE = -\tilde{\mathbf{D}}^T \mathbf{Y} + (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}) \tilde{\mathbf{w}}$$

Стартуя с начального вектора весов $\tilde{\mathbf{w}}^{(0)}$, мы обновляем веса согласно следующей итеративной процедуре:

$$\tilde{\mathbf{w}}^{(t+1)} = \tilde{\mathbf{w}}^{(t)} - \eta \nabla_{\tilde{\mathbf{w}}} = \tilde{\mathbf{w}}^{(t)} + \eta \tilde{\mathbf{D}}^T \left(\mathbf{Y} - \tilde{\mathbf{D}} \tilde{\mathbf{w}}^{(t)} \right),$$

где $\tilde{\mathbf{w}}^{(t)}$ – оценка вектора весов на шаге t , $\eta > 0$ – шаг обучения. Вектор весов обновляется по одной (случайной) точке набора \mathbf{D} на каждой итерации, т.е.

$$\tilde{\mathbf{w}}^{(t+1)} = \tilde{\mathbf{w}}^{(t)} - \eta \nabla_{\tilde{\mathbf{w}}} (\tilde{\mathbf{x}}_k) = \tilde{\mathbf{w}}^{(t)} + \eta \left(y_k - \tilde{\mathbf{x}}_k \tilde{\mathbf{w}}^{(t)} \right) \tilde{\mathbf{x}}_k$$



Входными данными для алгоритма множественной регрессии при помощи SGD являются матрица входных данных \mathbf{D} , вектор откликов \mathbf{Y} для точек набора \mathbf{D} , шаг обучения $\eta > 0$, требуемая точность $\varepsilon > 0$.

Multiple Regression: SGD ($\mathbf{D}, \mathbf{Y}, \eta, \varepsilon$):

```
1  $\tilde{\mathbf{D}} \leftarrow (\mathbf{1} \quad \mathbf{D})$  // создаем дополненные входные данные
2  $t \leftarrow 0$  // инициализируем счетчик шагов/итераций
3  $\tilde{\mathbf{w}}^{(0)} \leftarrow$  случайный вектор в  $\mathbb{R}^{d+1}$  // начальный вектор весов
4 repeat
5   foreach  $k = 1, 2, \dots, n$  (в случайном порядке) do
6      $\nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) \leftarrow -(y_k - \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}}^{(t)}) \cdot \tilde{\mathbf{x}}_k$  // вычислить градиент в  $\tilde{\mathbf{x}}_k$ 
7      $\tilde{\mathbf{w}}^{(t+1)} \leftarrow \tilde{\mathbf{w}}^{(t)} - \eta \cdot \nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k)$  // обновить оценку для весов
8    $t \leftarrow t + 1$ 
9 until  $\|\tilde{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}^{(t-1)}\| \leq \varepsilon$ 
```



Рассматривается множественная регрессия для набора данных Ирисы с признаками длины чашелистника X_1 и длины лепестка X_2 и шириной лепестка Y в качестве отклика.

Используя точный подход, получаем модель множественной регрессии в виде

$$\hat{y} = -0.0139 - 0.082 x_1 + 0.4499 x_2$$

Используя SGD, получаем следующую модель для $\eta = 0.001$ и $\varepsilon = 0.0001$:

$$\hat{y} = -0.031 - 0.078 x_1 + 0.45 x_2$$

Результаты подхода SGD по сути такие же, как и для точного метода, с небольшой разницей в коэффициенте смещения.

Значение ошибки SSE для точного метода составляет 6.179, тогда как для SGD ошибка составляет 6.181.



Для линейной регрессии вектор $\hat{\mathbf{Y}}$ лежит в линейном подпространстве, порожденном вектор-столбцами дополненной матрицы данных $\tilde{\mathbf{D}}$. Часто данные бывают зашумлены или не определены, поэтому вместо того, чтобы подгонять модель к данным точно, целесообразно использовать модель, более устойчивую к ошибкам в данных.

Регуляризация модели – это метод добавления некоторых дополнительных ограничений к условиям модели (обычно в форме штрафа за сложность модели) с целью повысить качество модели. Например, регуляризация может накладывать ограничение на L_2 -норму вектора весов $\tilde{\mathbf{w}}$.

Для этого в **гребневой** (ridge) **регрессии** к ошибке $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ добавляется слагаемое регуляризации ($\|\tilde{\mathbf{w}}\|^2$) и решается задача минимизации функции

$$J(\tilde{\mathbf{w}}) = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \alpha \|\tilde{\mathbf{w}}\|^2 = \|\mathbf{Y} - \tilde{\mathbf{D}}\tilde{\mathbf{w}}\|^2 + \alpha \|\tilde{\mathbf{w}}\|^2$$

Коэффициент $\alpha \geq 0$ управляет балансом между квадратом нормы вектора весов и квадратом ошибки прогноза в процессе минимизации. Таким образом, гребневая регрессия – это линейная регрессия с регуляризацией L_2 .



Для построения точного решения дифференцируем функцию $J(\tilde{\mathbf{w}})$ по $\tilde{\mathbf{w}}$ и приравняем результат к нулю, чтобы получить вектор весов в виде

$$\tilde{\mathbf{w}} = (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} + \alpha \mathbf{I})^{-1} \tilde{\mathbf{D}}^T \mathbf{Y},$$

где \mathbf{I} – единичная $(d+1) \times (d+1)$ -матрица. Матрица $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} + \alpha \mathbf{I}$ всегда является обратимой (невырожденной) для $\alpha > 0$, даже если матрица $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$ не обратима (вырождена).

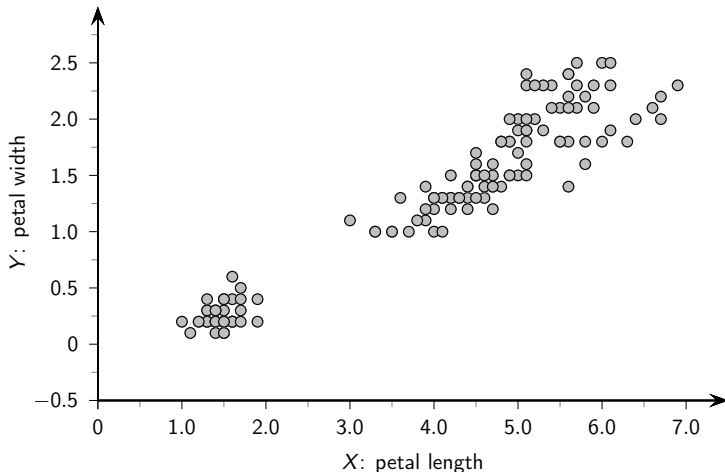
Если число λ_i является собственным значением матрицы $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$, то $\lambda_i + \alpha$ является собственным значением матрицы $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} + \alpha \mathbf{I}$. Поскольку матрица $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$ неотрицательно определенная, она имеет неотрицательные собственные значения. Даже если $\lambda_i = 0$, то соответствующее собственное значение матрицы $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} + \alpha \mathbf{I}$ равно $\lambda_i + \alpha = \alpha > 0$.

Регуляризованная таким образом регрессия называется гребневой (ridge) регрессией, потому что она добавляет «гребень» вдоль главной диагонали матрицы $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$, т.е. решение зависит от $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} + \alpha \mathbf{I}$.

Если выбирается положительное $\alpha > 0$, то гребневая регрессия гарантирует существование точного решения.



Рассматриваем длину лепестка X (petal length) как признак и ширину лепестка (petal width) как переменную отклика Y и исследуем гребневую регрессию в наборе данных Ирисы (с количеством точек $n = 150$).





Нецентрированная матрица рассеяния (uncentered scatter matrix) равна

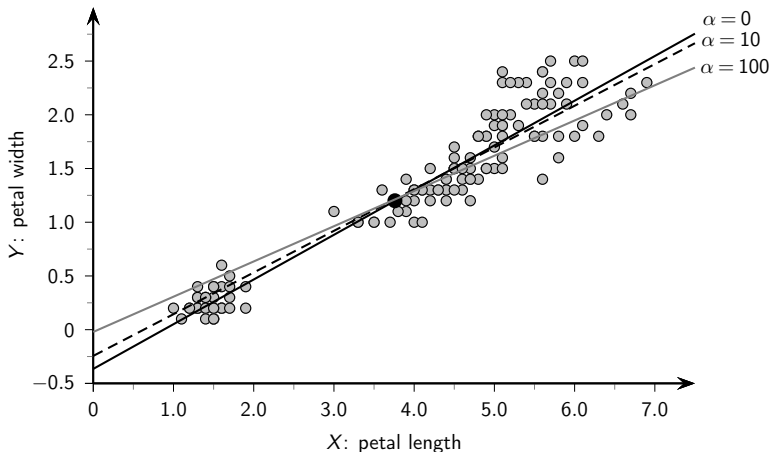
$$\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} = \begin{pmatrix} 150.0 & 563.8 \\ 563.8 & 2583.0 \end{pmatrix}$$

Получим различные линии наилучшего соответствия для различных значений параметра регуляризации α :

$$\begin{aligned} \alpha = 0 &\Rightarrow \hat{y} = -0.367 + 0.416x, \\ \|\tilde{\mathbf{w}}\|^2 &= \left\| (-0.367, 0.416)^T \right\|^2 = 0.308, SSE = 6.34 \\ \alpha = 10 &\Rightarrow \hat{y} = -0.244 + 0.388x, \\ \|\tilde{\mathbf{w}}\|^2 &= \left\| (-0.244, 0.388)^T \right\|^2 = 0.210, SSE = 6.75 \\ \alpha = 100 &\Rightarrow \hat{y} = -0.021 + 0.328x, \\ \|\tilde{\mathbf{w}}\|^2 &= \left\| (-0.021, 0.328)^T \right\|^2 = 0.108, SSE = 9.97 \end{aligned}$$



По мере увеличения α больше внимания уделяется минимизации квадрата нормы $\tilde{\mathbf{w}}$. Поскольку с увеличением α роль слагаемого $\|\tilde{\mathbf{w}}\|^2$ в минимизации увеличивается, соответствие модели данным обучающего набора уменьшается, что видно по увеличению значений ошибки SSE.





Вместо обращения матрицы $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} + \alpha \mathbf{I}$, как это требуется в точном решении для гребневой регрессии, можно использовать алгоритм стохастического градиентного спуска. Градиент функции $J(\tilde{\mathbf{w}})$ по $\tilde{\mathbf{w}}$, умноженный для удобства на $\frac{1}{2}$, равен

$$\nabla_{\tilde{\mathbf{w}}} = \frac{\partial}{\partial \tilde{\mathbf{w}}} J(\tilde{\mathbf{w}}) = -\tilde{\mathbf{D}}^T \mathbf{Y} + (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}) \tilde{\mathbf{w}} + \alpha \tilde{\mathbf{w}}$$

Используя (пакетный) градиентный спуск, можно итеративно вычислить $\tilde{\mathbf{w}}$ следующим образом

$$\tilde{\mathbf{w}}^{(t+1)} = \tilde{\mathbf{w}}^{(t)} - \eta \nabla_{\tilde{\mathbf{w}}} = (1 - \eta \alpha) \tilde{\mathbf{w}}^{(t)} + \eta \tilde{\mathbf{D}}^T (\mathbf{Y} - \tilde{\mathbf{D}} \tilde{\mathbf{w}}^{(t)})$$

В методе SGD вектор весов обновляется по одной (случайной) точке на каждой итерации:

$$\tilde{\mathbf{w}}^{(t+1)} = \tilde{\mathbf{w}}^{(t)} - \eta \nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) = \left(1 - \frac{\eta \alpha}{n}\right) \tilde{\mathbf{w}}^{(t)} + \eta \left(y_k - \tilde{\mathbf{x}}_k \tilde{\mathbf{w}}^{(t)}\right) \tilde{\mathbf{x}}_k$$

Здесь константа регуляризации α масштабируется делением на n , так как исходное значение предназначалось для всех n точек набора данных \mathbf{D} .



Входными данными для алгоритма множественной гребневой регрессии при помощи SGD являются матрица входных данных \mathbf{D} , вектор откликов \mathbf{Y} для точек набора \mathbf{D} , шаг обучения $\eta > 0$, требуемая точность $\varepsilon > 0$.

Ridge Regression: SGD ($\mathbf{D}, \mathbf{Y}, \eta, \varepsilon$):

```

1  $\tilde{\mathbf{D}} \leftarrow (\mathbf{1} \quad \mathbf{D})$  // дополненные входные данные
2  $t \leftarrow 0$  // инициализация счетчика шагов/итераций
3  $\tilde{\mathbf{w}}^{(0)} \leftarrow$  случайный вектор в  $\mathbb{R}^{d+1}$  // начальный вектор весов
4 repeat
5     foreach  $k = 1, 2, \dots, n$  (в случайном порядке) do
6          $\nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k) \leftarrow -(y_k - \tilde{\mathbf{x}}_k^T \tilde{\mathbf{w}}^{(t)}) \cdot \tilde{\mathbf{x}}_k + \frac{\alpha}{n} \tilde{\mathbf{w}}$  // градиент в точке  $\tilde{\mathbf{x}}_k$ 
7          $\tilde{\mathbf{w}}^{(t+1)} \leftarrow \tilde{\mathbf{w}}^{(t)} - \eta \cdot \nabla_{\tilde{\mathbf{w}}}(\tilde{\mathbf{x}}_k)$  // обновить оценку для весов
8      $t \leftarrow t + 1$ 
9 until  $\|\tilde{\mathbf{w}}^{(t)} - \tilde{\mathbf{w}}^{(t-1)}\| \leq \varepsilon$ 
    
```



Применим гребневую регрессию к набору данных Ирисы ($n = 150$), используя длину лепестка X (petal length) в качестве независимого признака и ширину лепестка Y (petal width) в качестве переменной отклика.

Используя SGD (с параметрами $\eta = 0.001$ и $\varepsilon = 0.0001$), получим уравнения гребневой регрессии для разных значений константы регуляризации α :

$$\alpha = 0 \Rightarrow \hat{y} = -0.366 + 0.416 x, SSE_{SGD} = 6.37, SSE_{Ridge} = 6.34$$

$$\alpha = 10 \Rightarrow \hat{y} = -0.244 + 0.387 x, SSE_{SGD} = 6.76, SSE_{Ridge} = 6.38$$

$$\alpha = 100 \Rightarrow \hat{y} = -0.022 + 0.327 x, SSE_{SGD} = 10.04, SSE_{Ridge} = 8.87$$

Полученные при помощи SGD уравнения гребневой регрессии, в целом, соответствуют уравнениям гребневой регрессии, полученным ранее точным способом.



Лассо (least absolute selection and shrinkage operator, lasso) – это метод регуляризации, направленный на обнуление части весов регрессии.

Сделаем допущение, что признаки X_1, X_2, \dots, X_d и отклик Y центрированы (будем использовать обозначения $\bar{\mathbf{D}}$ и $\bar{\mathbf{Y}}$). Центрирование освобождает нас от необходимости явного использования в регрессии коэффициента смещения $b = w_0$.

Регрессия лассо (в отличие от гребневой регрессии) использует для регуляризации норму L_1 :

$$\min_{\mathbf{w}} J(\mathbf{w}), \quad J(\mathbf{w}) = \frac{1}{2} \|\bar{\mathbf{Y}} - \bar{\mathbf{D}}\mathbf{w}\|^2 + \alpha \|\mathbf{w}\|_1,$$

где коэффициент $\alpha \geq 0$ – константа регуляризации и для вектора весов $\mathbf{w} = (w_1, w_2, \dots, w_d)$

$$\|\mathbf{w}\|_1 = \sum_{i=1}^d |w_i|$$



Использование нормы L_1 приводит к разреженности вектора весов \mathbf{w} .

Гребневая регрессия L_2 уменьшает значения коэффициентов регрессии w_i , но они могут оставаться небольшими, но все же отличными от нуля.

Регрессия L_1 способна обнулять коэффициенты регрессии, что приводит к более интерпретируемой модели, особенно когда в наборе данных много признаков.

Целевая функция в регрессии лассо состоит из двух частей: функции квадрата ошибки $\|\bar{\mathbf{Y}} - \bar{\mathbf{D}}\mathbf{w}\|^2$, являющейся выпуклой и дифференцируемой, и функции штрафа L_1

$$\alpha \|\mathbf{w}\|_1 = \alpha \sum_{i=1}^d |w_i|,$$

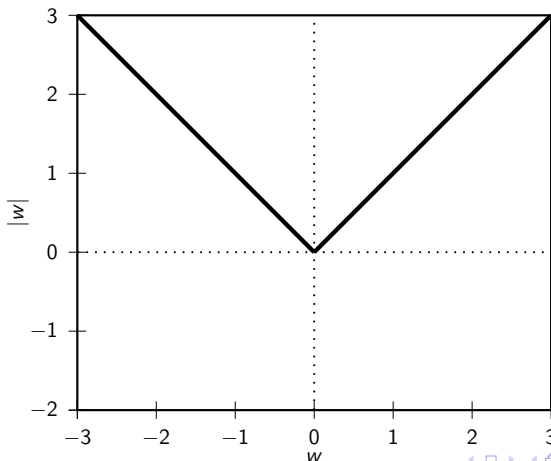
которая является выпуклой, но недифференцируемой при $w_i = 0$. Поэтому мы не можем просто вычислить градиент и приравнять его к нулю, как это делается в случае гребневой регрессии. Задачу минимизации в регрессии лассо можно решить с помощью обобщенного подхода субградиентов.



Рассмотрим функцию абсолютного значения $f(w) = |w|$.

Когда $w > 0$, имеем $f'(w) = +1$, а когда $w < 0$, имеем $f'(w) = -1$.

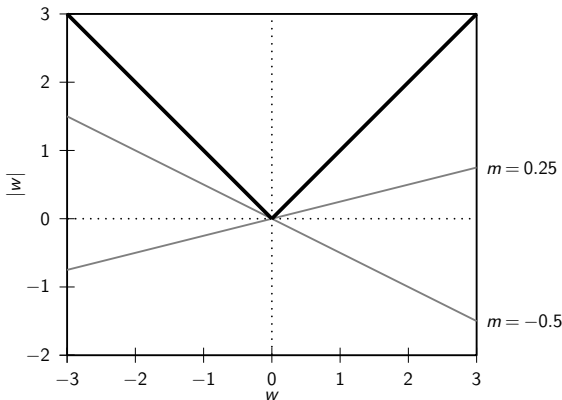
В точке $w = 0$ производная не существует.





Субградиенты обобщают понятие производной.

Для функции $f(w) = |w|$ наклон m любой прямой, проходящей через точку $w = 0$ и остающейся ниже или касающейся графика функции f , называется **субградиентом функции f** в точке $w = 0$.





Множество всех субградиентов функции $|w|$ называется **субдифференциалом** и обозначается как $\partial |w|$.

Субдифференциал функции $f(w) = |w|$ при $w = 0$ определяется формулой $\partial |w| = [-1, 1]$.

Рассматривая любые значения w , получим следующую формулу для субдифференциала функции $f(w) = |w|$:

$$\partial |w| = \begin{cases} +1, & w > 0 \\ -1, & w < 0 \\ [-1, 1], & w = 0 \end{cases}$$

Когда производная (градиент) функции существует, субдифференциал принимает единственное значение и равен значению производной (или градиенту). Когда производной не существует, субдифференциал соответствует набору субградиентов.



Рассмотрим парную регрессию лассо с единственным независимым признаком \bar{X} и откликом \bar{Y} (оба признака центрированы). Тогда модель парной регрессии задается в виде

$$\hat{y}_i = w x_i.$$

Целевая функция регрессии лассо записывается в виде

$$J(w) = \frac{1}{2} \sum_{i=1}^n (y_i - w x_i)^2 + \alpha |w|.$$

Субдифференциал функции $J(w)$ вычисляется следующим образом:

$$\begin{aligned} \partial J(w) &= \frac{1}{2} \sum_{i=1}^n 2(y_i - w x_i)(-x_i) + \alpha \partial |w| = \\ &= - \sum_{i=1}^n x_i y_i + w \sum_{i=1}^n x_i^2 + \alpha \partial |w| = -\bar{\mathbf{X}}^T \bar{\mathbf{Y}} + w \|\bar{\mathbf{X}}\|^2 + \alpha \partial |w| \end{aligned}$$



Приравняем субдифференциал $J(w)$ к нулю и получим

$$\partial J(w) = 0 \Rightarrow w \|\bar{\mathbf{X}}\|^2 + \alpha \partial |w| = \bar{\mathbf{X}}^T \bar{\mathbf{Y}} \Rightarrow w + \eta \alpha \partial |w| = \eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}}, \eta = \frac{1}{\|\bar{\mathbf{X}}\|^2}$$

В соответствии с тремя случаями для субдифференциала функции абсолютного значения $|w|$, нужно рассмотреть три случая:

❶ $w > 0, \partial |w| = +1:$

$$\begin{aligned} w &= \eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} - \eta \alpha \\ w > 0 &\Rightarrow \eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} > \eta \alpha \Rightarrow |\eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}}| > \eta \alpha \end{aligned}$$

❷ $w < 0, \partial |w| = -1:$

$$\begin{aligned} w &= \eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} + \eta \alpha \\ w < 0 &\Rightarrow \eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} < -\eta \alpha \Rightarrow |\eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}}| > \eta \alpha \end{aligned}$$

❸ $w = 0, \partial |w| \in [-1, +1]:$

$$w \in [\eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} - \eta \alpha, \eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} + \eta \alpha]$$

$$w = 0 \Rightarrow |\eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}}| \leq \eta \alpha$$



Пусть $\tau \geq 0$ – некоторое фиксированное значение. Определим **функцию мягкого порога** (soft-threshold function) $S_\tau : \mathbb{R} \rightarrow \mathbb{R}$ следующим образом:

$$S_\tau(z) = \text{sign}(z) \max\{0, |z| - \tau\}$$

Тогда указанные выше три случая можно компактно записать как:

$$w = S_{\eta\alpha}(\eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}}),$$

где $\tau = \eta\alpha$. Таким образом, полученная формула задает оптимальное решение (вектор весов) задачи парной регрессии лассо.



L_1 -Regression $(\mathbf{D}, Y, \alpha, \eta, \varepsilon)$:

```

1   $\boldsymbol{\mu} \leftarrow \text{mean}(\mathbf{D}), \mu_Y \leftarrow \text{mean}(\mathbf{Y})$  // вычисляем средние значения
2   $\bar{\mathbf{D}} \leftarrow \mathbf{D} - \mathbf{1} \cdot \boldsymbol{\mu}^T$  // центрируем входные данные
3   $\bar{Y} \leftarrow Y - \mu_Y \cdot \mathbf{1}$  // центрируем отклик
4   $t \leftarrow 0$  // счетчик шагов/итераций
5   $\mathbf{w}^{(0)} \leftarrow$  случайный вектор в  $\mathbb{R}^d$  // начальный вектор весов
6  repeat
7      foreach  $k = 1, 2, \dots, d$  do
8           $\nabla(w_k^{(t)}) \leftarrow -\bar{X}_k^T (\bar{\mathbf{Y}} - \bar{\mathbf{D}} \mathbf{w}^{(t)})$  // вычисляем градиент
9           $w_k^{(t+1)} \leftarrow w_k^{(t)} - \eta \cdot \nabla(w_k^{(t)})$  // обновить оценку весов
10          $w_k^{(t+1)} \leftarrow \mathcal{S}_{\eta \cdot \alpha}(w_k^{(t+1)})$  // функция мягкого порога
11      $t \leftarrow t + 1$ 
12 until  $\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\| \leq \varepsilon$ 
13  $b \leftarrow \mu_Y - (\mathbf{w}^{(t)})^T \boldsymbol{\mu}$  // вычислить смещение
    
```



Применим регрессию лассо к полному набору данных Ирисы с $n = 150$ точками и четырьмя независимыми признаками, а именно шириной чашелистика X_1 (sepal-width), длиной чашелистика X_2 (sepal-length), шириной лепестка X_3 (petal-width) и длиной лепестка X_4 (petal-length).

Признак типа ириса содержит переменную отклика Y . Существует три типа ирисов, а именно Iris-setosa, Iris-versicolor и Iris-virginica, которые имеют коды 0, 1 и 2 соответственно.

Результаты регрессии L_1 для различных α и $\eta = 0.0001$ показаны ниже:

$$\alpha = 0 : \hat{y} = +0.19 - 0.11 x_1 - 0.05 x_2 + 0.23 x_3 + 0.61 x_4, SSE = 6.96, \|\mathbf{w}\|_1 = 0.44$$

$$\alpha = 1 : \hat{y} = -0.08 - 0.08 x_1 - 0.02 x_2 + 0.25 x_3 + 0.52 x_4, SSE = 7.09, \|\mathbf{w}\|_1 = 0.34$$

$$\alpha = 5 : \hat{y} = -0.55 + 0.00 x_1 + 0.00 x_2 + 0.36 x_3 + 0.17 x_4, SSE = 8.82, \|\mathbf{w}\|_1 = 0.16$$

$$\alpha = 10 : \hat{y} = -0.58 + 0.00 x_1 + 0.00 x_2 + 0.42 x_3 + 0.00 x_4, SSE = 10.15, \|\mathbf{w}\|_1 = 0.18$$

Обратите внимание на эффект обнуления некоторых весов для значений $\alpha = 5$ и $\alpha = 10$.



Построим и сравним коэффициенты гребневой регрессии (L_2) и регрессии лассо (L_1) с одинаковым уровнем квадратичной ошибки.

При $\alpha = 5$ модель регрессии лассо имеет ошибку $SSE = 8.82$.

Установим значение параметра $\alpha = 35$ в гребневой регрессии, что приведет к аналогичной ошибке SSE . Две модели имеют следующее представление:

$$L_1 : \hat{y} = -0.553 + 0.00 x_1 + 0.00 x_2 + 0.359 x_3 + 0.17 x_4, \|\mathbf{w}\|_1 = 0.156$$

$$L_2 : \hat{y} = -0.394 + 0.019 x_1 - 0.051 x_2 + 0.316 x_3 + 0.212 x_4, \|\mathbf{w}\|_1 = 0.598$$

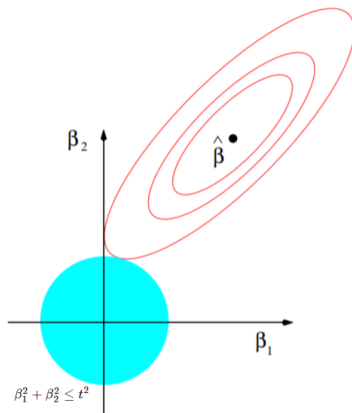
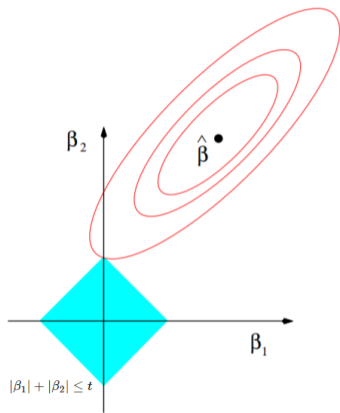
В модели гребневой регрессии коэффициенты при x_1 и x_2 малы и, следовательно, менее важны, но они не равны нулю.

В модели регрессии лассо коэффициенты для x_1 и x_2 в точности равны нулю, остаются только признаки x_3 и x_4 .

Таким образом, регрессия лассо может осуществлять **отбор значимых признаков**.



Основное различие регрессии лассо (L_1) и гребневой регрессии (L_2) заключается в том, что регрессия лассо может приводить к обнулению весов некоторых независимых переменных, тогда как гребневая регрессия уменьшает их до значений, близких к нулю.

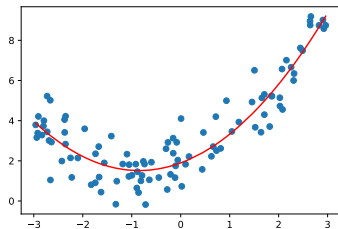
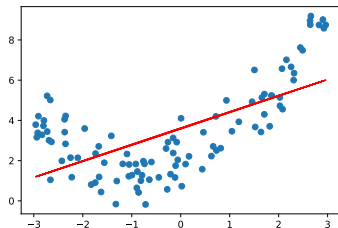




Линейной регрессии вида

$$Y = \beta + \boldsymbol{\omega}^T \mathbf{X} + \varepsilon = \beta + \sum_{i=1}^d \omega_i X_i + \varepsilon$$

может быть недостаточно для выявления взаимосвязи между признаками X_1, X_2, \dots, X_d и откликом Y в случае, когда эта взаимосвязь является нелинейной. В этом случае можно воспользоваться т.н. полиномиальной регрессией, в которой используются степени независимых переменных, которые рассматриваются как отдельные независимые переменные в модели множественной регрессии.





Полиномиальная регрессия – это форма регрессионного анализа, в которой взаимосвязь между независимой переменной X и зависимой переменной Y моделируется как полином m -й степени от X

$$Y = f(X, \mathbf{w}) + \varepsilon = w_0 + w_1X + w_2X^2 + \dots + w_mX^m + \varepsilon$$

Полиномиальная регрессия соответствует нелинейной зависимости между значением X и соответствующим условным средним значением Y , обозначаемым $\mathbb{E}[Y | X]$. Хотя полиномиальная регрессия подгоняет нелинейную модель к нелинейным данным, как задача статистической оценки она является линейной в том смысле, что функция регрессии $\mathbb{E}[Y | X]$ является линейной по неизвестным параметрам $\mathbf{w} = (w_0, w_1, \dots, w_m)^T$, которые оцениваются на основе имеющихся данных. По этой причине полиномиальная регрессия считается частным случаем множественной линейной регрессии.



Пусть входные данные имеют вид $\mathbf{D} = \{\mathbf{X} = (x_1, \dots, x_n)^T, \mathbf{Y} = (y_1, \dots, y_n)^T\}$, тогда модель полиномиальной регрессии принимает вид системы линейных уравнений относительно весов $\mathbf{w} = (w_0, w_1, \dots, w_m)^T$:

$$\mathbf{Y} = \mathbf{V} \mathbf{w} + \boldsymbol{\varepsilon},$$

где \mathbf{V} представляет собой $n \times (m + 1)$ -матрицу Вандермонда вида

$$\mathbf{V} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Можно доказать, что матрица Вандермонда \mathbf{V} при $m < n$ имеет максимальный ранг $m + 1$ тогда и только тогда, когда все точки x_i различны. Тогда $(m + 1) \times (m + 1)$ -матрица $\mathbf{V}^T \mathbf{V}$ будет невырожденной, т.е. будет существовать обратная матрица $(\mathbf{V}^T \mathbf{V})^{-1}$.



Матрица \mathbf{V} нелинейным образом зависит от значений признака \mathbf{X} , но функция полиномиальной регрессии является линейной относительно коэффициентов регрессии (весов) $\mathbf{w} = (w_0, w_1, \dots, w_m)^T$:

$$\hat{\mathbf{Y}} = f(\mathbf{X}, \mathbf{w}) = \mathbf{V}(\mathbf{X}) \mathbf{w},$$

поэтому в случае максимального ранга матрицы \mathbf{V} коэффициенты регрессии \mathbf{w} могут быть определены методом наименьших квадратов:

$$\min_{\mathbf{w}} SSE = \min_{\mathbf{w}} \|\boldsymbol{\epsilon}\|^2 = \min_{\mathbf{w}} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \min_{\mathbf{w}} \|\mathbf{Y} - \mathbf{V}(\mathbf{X}) \mathbf{w}\|^2$$

Приравнявая градиент SSE по \mathbf{w} к нулю и решая полученные уравнения, получим формулу для коэффициентов полиномиальной регрессии

$$\mathbf{w} = (\mathbf{V}^T \mathbf{V})^{-1} \mathbf{V}^T \mathbf{Y}$$

Таким образом, задачи полиномиальной регрессии решаются с помощью методов множественной регрессии, если переменные X, X^2, X^3, \dots трактуются как отдельные независимые переменные в модели множественной регрессии.



Полиномиальную регрессию можно обобщить на случай линейной регрессии общего вида, когда восстанавливается зависимость переменной \mathbf{Y} от другой или нескольких других переменных (признаков) $\mathbf{X} \in \mathbb{R}^d$ с линейной зависимостью от неизвестных коэффициентов $\mathbf{w} = (w_1, \dots, w_m) \in \mathbb{R}^m$ вида:

$$\mathbf{Y} = f(\mathbf{X}, \mathbf{w}) = \sum_{k=1}^m w_k f_k(\mathbf{X}),$$

где $f_1(\mathbf{X}), \dots, f_m(\mathbf{X})$ – некоторые базисные функции. В качестве базисных функций могут рассматриваться различные полиномы, сплайны, радиальные базисные функции, вейвлеты и т.п.

Пусть даны значения независимых переменных $\mathbf{x}_i \in \mathbb{R}^d, i = \overline{1, n}$ и соответствующие значения зависимой переменной $y_i \in \mathbb{R}, i = \overline{1, n}$. Введем матричные обозначения $\mathbf{F}(\mathbf{X}) = \begin{pmatrix} f_1(\mathbf{x}_1) & \dots & f_m(\mathbf{x}_1) \\ \dots & \dots & \dots \\ f_1(\mathbf{x}_n) & \dots & f_m(\mathbf{x}_n) \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}.$



$n \times m$ -матрица $\mathbf{F}(\mathbf{X})$ имеет нелинейную зависимость общего вида от значений признаков \mathbf{X} , но функция регрессии $f(\mathbf{X}, \mathbf{w})$ является линейной относительно коэффициентов регрессии (весов) \mathbf{w} :

$$\hat{\mathbf{Y}} = f(\mathbf{X}, \mathbf{w}) = \mathbf{F}(\mathbf{X}) \mathbf{w},$$

поэтому в случае максимальности ранга матрицы \mathbf{F} аналогично случаю полиномиальной регрессии коэффициенты регрессии \mathbf{w} могут быть определены методом наименьших квадратов:

$$\min_{\mathbf{w}} SSE = \min_{\mathbf{w}} \|\boldsymbol{\varepsilon}\|^2 = \min_{\mathbf{w}} \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|^2 = \min_{\mathbf{w}} \left\| \mathbf{Y} - \mathbf{F}(\mathbf{X}) \mathbf{w} \right\|^2$$

Приравнивая градиент SSE по \mathbf{w} к нулю и решая полученные уравнения, получим формулу для коэффициентов линейной регрессии общего вида

$$\mathbf{w} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{Y}$$

Таким образом, задачи линейной регрессии общего вида также решаются с помощью методов множественной регрессии, если переменные $f_1(\mathbf{X})$, ..., $f_m(\mathbf{X})$ трактуются как отдельные независимые переменные в модели множественной регрессии.



Матрица \mathbf{F} имеет размеры $n \times m$, где, вообще говоря, $n > m$, матрица \mathbf{F}^T имеет размеры $m \times n$, а матрицы $\mathbf{F}^T \mathbf{F}$ и $(\mathbf{F}^T \mathbf{F})^{-1}$ – размеры $m \times m$.

Матрица $\mathbf{F}^+ = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T$ в формуле весов линейной регрессии общего вида называется **псевдообратной матрицей** (или матрицей Мура–Пенроуза):

$$\mathbf{F}^+ \mathbf{F} = (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{F} = \mathbf{1}_m,$$

где $\mathbf{1}_m$ – единичная $m \times m$ -матрица. Матрица $\mathbf{F} \mathbf{F}^+ = \mathbf{F} (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T$ (размерами $n \times n$) называется **проекционной матрицей**.

Значение функции SSE для построенного вектора весов \mathbf{w} будет равно

$$SSE = \|\mathbf{Y} - \mathbf{F} \mathbf{F}^+ \mathbf{Y}\|^2$$

Если столбцы матрицы \mathbf{F} линейно-зависимы (случай мультиколлинеарности), то матрица $\mathbf{F}^T \mathbf{F}$ будет вырожденной и обратная матрица $(\mathbf{F}^T \mathbf{F})^{-1}$ не существует.

Если же столбцы матрицы \mathbf{F} будут близки к линейной зависимости, то обращение матрицы $\mathbf{F}^T \mathbf{F}$ будет сложной вычислительной задачей.



Если определитель матрицы $\mathbf{F}^T \mathbf{F}$ близок к нулю (имеются собственные числа близкие к нулю), то

- решение задачи линейной регрессии общего вида \mathbf{w} становится неустойчивым и неинтерпретируемым и может содержать слишком большие компоненты w_j различных знаков
- на обучающих данных функция $SSE = \|\mathbf{Y} - \mathbf{F}\mathbf{w}\|^2$ может иметь малые значения, а на контрольных (или новых) данных функция $SSE = \|\mathbf{Y}' - \mathbf{F}'\mathbf{w}\|^2$ может принимать значительно большие значения (т.е. будет иметь место переобучение)

Для устранения мультиколлинеарности (и переобучения) можно провести:

- **отбор признаков**, то есть отбрасывание тех признаков, которые могут оказаться линейно-зависимыми с другими признаками
- **регуляризацию** (накладываем дополнительные ограничения на вектор коэффициентов вида $\|\mathbf{w}\| \leq \alpha$)
- **преобразование признаков** (отобразить входные данные на новое пространство признаков)



Пусть $\mathbf{Y} = (y_1, y_2, \dots, y_n)^T$ – вектор значений откликов, а $\hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)^T$ – вектор прогнозируемых значений откликов. Тогда для оценки качества регрессии можно использовать показатели:

- Среднее квадратичное отклонение (Mean Squared Error, MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Корень среднеквадратичного отклонения (Root Mean Squared Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- Среднее квадратичное процентное отклонение (Mean Squared Percentage Error, MSPE)

$$MSPE = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2$$



- Среднее абсолютное отклонение (Mean Absolute Error, MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Метрика MAE более устойчива к выбросам, чем MSE. Например, что если для одной точки отклонение очень большое (точка представляет собой выброс), а для остальных точек — маленькое, то значение MAE увеличится от этой одной точки меньше, чем MSE, т.к. в MSE отклонения возводятся в квадрат.

- Среднее абсолютное процентное отклонение (Mean Average Percentage Error, MAPE)

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Метрика MAPE усредняет значения отклонений, деленных на значение целевой переменной.



- Среднее квадратичное логарифмическое отклонение (Mean Squared Logarithmic Error, MSLE)

$$MSLE = \frac{1}{n} \sum_{i=1}^n (\ln(1 + y_i) - \ln(1 + \hat{y}_i))^2$$

- Медианное абсолютное отклонение (Median Absolute Error, MedAE)

$$MedAE = \text{median}(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|)$$

- Максимальная ошибка (maximum residual error, MaxErr)

$$MaxErr = \max_{i=\overline{1,n}} |y_i - \hat{y}_i|$$



Коэффициент детерминации (coefficient of determination) R^2 представляет собой долю дисперсии \mathbf{Y} , которая была объяснена независимой переменной модели.

Коэффициент R^2 вычисляется по формуле

$$R^2 = 1 - \frac{MSE}{\mathbb{V}[\mathbf{Y}]} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

где $\mathbb{V}[\mathbf{Y}] = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$ – дисперсия \mathbf{Y} , $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ – выборочное среднее зависимой переменной \mathbf{Y} . Чем больше значение коэффициента детерминации R^2 , тем лучше данные описываются заданной моделью. В случае, когда $R^2 = 1$, все точки набора данных в точности лежат на линии регрессии.



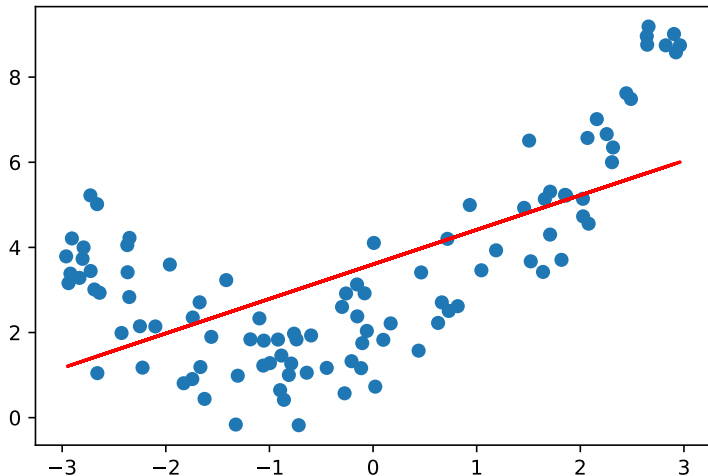
Одной из основных проблем машинного обучения является т.н. переобучение.

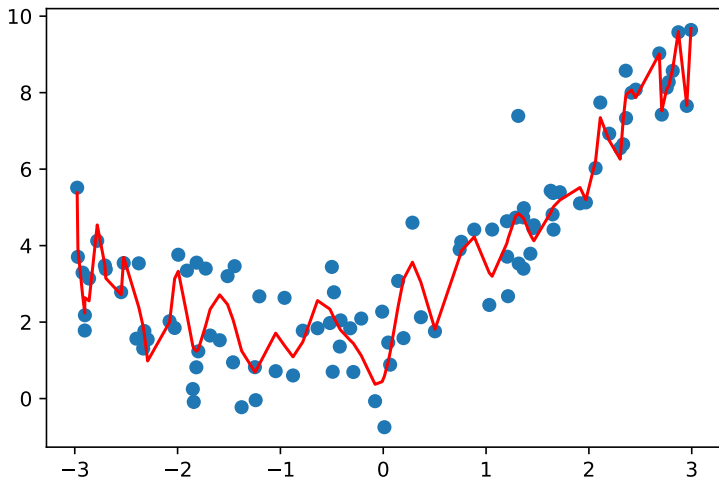
Переобучение (overfitting, overtraining) — это ситуация, когда обучаемая модель для данных обучающего набора прогнозирует отклики, которые близко или даже точно соответствуют откликам в этом наборе, однако для данных, не участвовавших в процессе обучения, модель вырабатывает предсказания низкого качества. Переобучение может возникать

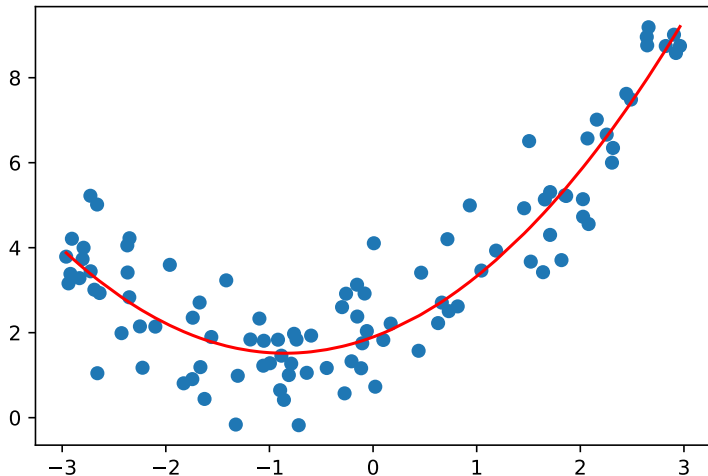
- при использовании слишком сложных моделей
- при слишком долгом процессе обучения
- при неудачной обучающей выборке

Недообучение (underfitting) — это ситуация, когда обучаемая модель не обеспечивает приемлемого качества (достаточно малой величины средней ошибки) даже на обучающем наборе. Недообучение может возникать

- при использовании слишком простых моделей
- при прекращении процесса обучения до достижения состояния с достаточно малой ошибкой
- при неудачной обучающей выборке





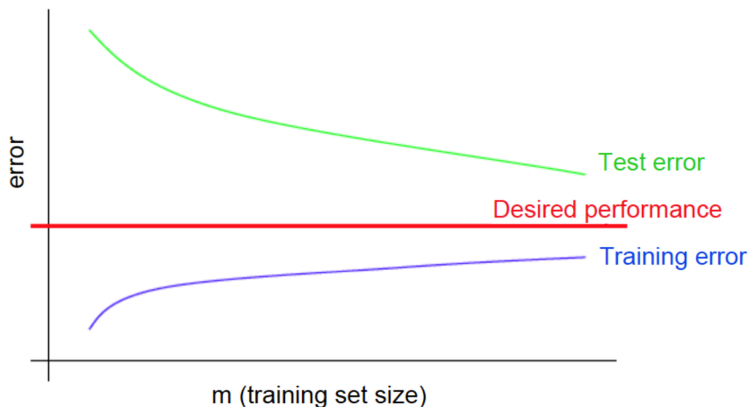




Кривая обучения — графическое представление зависимости меры (показателя) качества обучения (по вертикальной оси) от определенного показателя модели обучения (по горизонтальной оси). Например, в примерах ниже представлена зависимость средней ошибки от объема датасета.



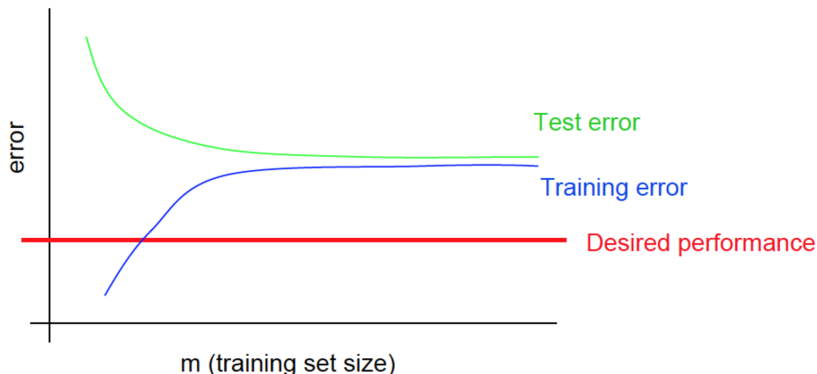
При переобучении небольшая средняя ошибка на обучающей выборке не обеспечивает такую же малую ошибку на тестовой выборке.



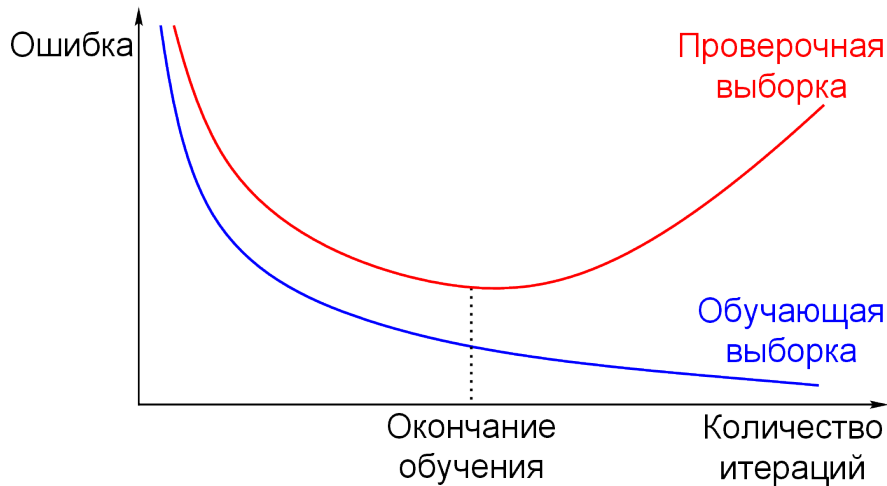
Переобучение связано с высокой дисперсией (variance) алгоритма обучения.



При недообучении независимо от объема обучающего датасета как на обучающей выборке, так и на тестовой выборке небольшая средняя ошибка не достигается.



Недообучение связано с высоким смещением (bias) алгоритма обучения.





Возможные решения при переобучении:

- Увеличение количества данных в наборе;
- Уменьшение количества параметров модели;
- Добавление регуляризации (увеличение коэффициента регуляризации).

Возможные решения при недообучении:

- Добавление новых параметров в модель;
- Использование для описания модели более сложных функций (с более высокой степенью);
- Уменьшение коэффициента регуляризации.



- линейная регрессия (LinearRegression)
- полиномиальная регрессия (PolynomialFeatures+LinearRegression)
- стохастический градиентный спуск (SGDRegressor)
- гребневая регрессия (Ridge)
- регрессия лассо (Lasso)
- регрессия эластичная сеть (ElasticNet)
- регрессия на основе метода ближайших соседей (KNeighborsRegressor)
- регрессия на основе деревьев решений (DecisionTreeRegressor)
- регрессия на основе метода опорных векторов (SVR)