

# Методы машинного обучения

Шорохов С.Г.

кафедра математического моделирования и искусственного интеллекта

## Лекция 1. Линейная регрессия



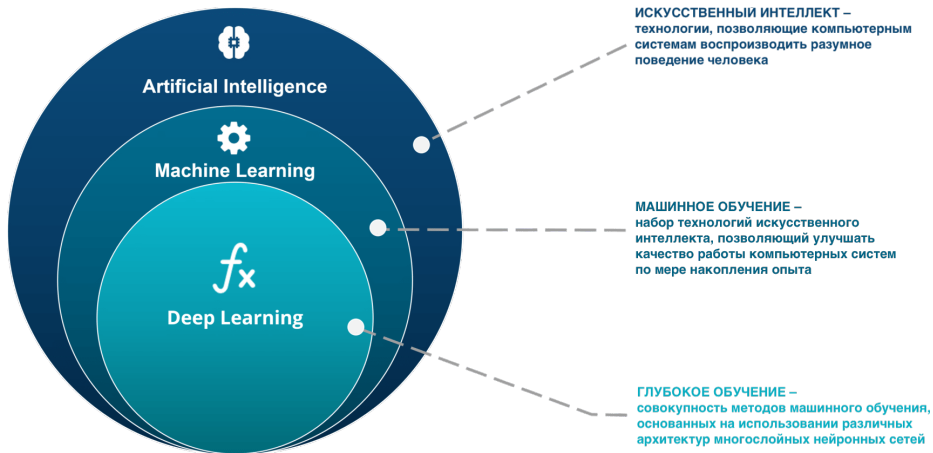


Термин «машинное обучение» (machine learning) был впервые введен А. Самюэлом из компании IBM (A.L. Samuel, Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229).

Основоположниками современной теории машинного обучения принято считать советских ученых В. Вапника и А. Червоненкиса:

- В.Н. Вапник, А.Я. Червоненкис. Теория распознавания образов. Статистические проблемы обучения. – М.: Наука, 1974. – 416 с.
- В.Н. Вапник. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979. – 448 с.

В. Вапник и А. Червоненкис использовали термин «обучение машин распознаванию образов». В дальнейшем В. Вапник использовал термин «Statistical Learning Theory».





*When you're fundraising, it's AI*

*When you're hiring, it's ML*

*When you're implementing, it's linear regression*

*When you're debugging, it's printf()*

*— Baron Schwartz (@xaprb) November 15, 2017*



## Лекции (теория):

- M.Zaki, W.Meira, Data Mining and Machine Learning. Fundamental Concepts and Algorithms 2e (2020)

## Лабораторные работы:

- F.Chollet, Deep Learning with Python 2e (2021)



- ❶ 15.04 – линейная регрессия
- ❷ 22.04 – нелинейная регрессия
- ❸ 29.04 – выходной
- ❹ 06.05 – искусственные нейронные сети и их обучение
- ❺ 13.05 – нейронные сети MLP
- ❻ 20.05 – нейронные сети RNN
- ❼ 27.05 – нейронные сети CNN
- ❽ 03.06 – автокодировщики



Дано пространство объектов (признаков)  $\mathbf{X}$  с плотностью распределения вероятности  $\rho(x)$  (вообще говоря, неизвестной).

Также дано пространство ответов (откликов)  $\mathbf{Y}$  и определена функция  $f : \mathbf{X} \rightarrow \mathbf{Y}$ , которая неизвестна, но известны значения  $y_i = f(x_i)$ , которые функция  $f$  принимает на множестве  $\{x_i\} \subset \mathbf{X}$  (обучающая выборка).

Требуется построить функцию (алгоритм)  $f' : \mathbf{X} \rightarrow \mathbf{Y}$ , приближающую функцию  $f$  так, чтобы на функции  $f'$  достигался минимум функционала (критерия качества)

$$J[u(x)] = \int_{\mathbf{X}} F(x, u(x)) \rho(x) dx,$$

причем функция (алгоритм)  $f'$  принадлежит некоторому классу функций (алгоритмов)  $\mathcal{F}$ :

$$f'(x) = \arg \min_{u(x) \in \mathcal{F}} \int_{\mathbf{X}} F(x, u(x)) \rho(x) dx.$$



**Дискриминативные модели** (discriminative models) изучают условное вероятностное распределение  $p(\mathbf{y} \mid \mathbf{x})$  с входными данными  $\mathbf{x}$  и целевой меткой  $\mathbf{y}$ . Дискриминативные модели предсказывают метку  $\mathbf{y}$  по входным данным  $\mathbf{x}$ . Дискриминативные модели в основном используются в таких задачах машинного обучения, как классификация и регрессия.

**Генеративные (порождающие) модели** (generative models) предназначены для изучения совместного вероятностного распределения данных (и меток)  $p(\mathbf{x}, \mathbf{y})$ . Генеративные модели обычно используются для получения (генерации) данных путем изучения распределения наблюдаемых данных. Из совместного распределения можно получить условное распределение

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})},$$

но совместное распределение  $p(\mathbf{x}, \mathbf{y})$  даёт больше информации и его можно использовать, например, для генерации новых изображений объектов, которые выглядят как настоящие.







- **Задача кластеризации** — это задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.
- В задаче классификации имеется множество объектов, разделённых некоторым образом на классы, и задано конечное подмножество объектов (выборка), для которых известно, к каким классам они относятся (классовая принадлежность остальных объектов неизвестна). **Задача классификации** – это задача построения алгоритма, способного классифицировать (определять метку или наименование класса) произвольного объекта из исходного множества.
- В задаче регрессии на множестве объектов определена некоторая (вообще говоря) неизвестная функция и задано конечное подмножество объектов (выборка), для которых известны значения функции. **Задача регрессии** – это задача построения алгоритма, способного находить значение функции для произвольного объекта из исходного множества.



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  $\varepsilon$  – 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  $\varepsilon$  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  $f$  предполагается линейной по  $\mathbf{X}$ , 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  $\beta$  – истинное (неизвестное) смещение (bias),  $\omega_i$  – истинный (неизвестный) коэффициент регрессии или вес для признака  $X_i$ ,  $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_d)^T$  – истинный  $d$ -мерный вектор весов.

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  $\beta$  задает точки пересечения гиперплоскости с осями координат пространства признаков. Функция регрессии  $f$  полностью определяется  $d + 1$  параметрами  $\beta$  и  $\omega_i$  для  $i = 1, \dots, d$ .

Наиболее распространенным подходом к прогнозированию параметров линейной регрессии (коэффициентов смещения  $b$  и регрессии  $\mathbf{w}$ ) является использование **метода 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  $y$ . Суть **метода наименьших квадратов** 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 соответствующие 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)

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

Основные подходы к минимизации SSE:

- продифференцировать SSE по неизвестным параметрам  $b, \mathbf{w}$ , приравнять производные к нулю и решить полученную систему уравнений
- использовать численные методы минимизации функций



В **парной** (bivariate) **регрессии** обучающие данные  $\mathbf{D}$  содержат один признак  $X = (x_1, x_2, \dots, x_n)^T$  вместе с откликом  $Y = (y_1, y_2, \dots, y_n)^T$ , а линейная функция регрессии зависит от двух 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 условия:

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

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

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



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  $X$  и  $Y$  и дисперсию  $X$ :

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



Итак, в парной регрессии

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

оценка параметров регрессии  $b$  и  $w$  по обучающим данным

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

производится по формулам

$$b = \mu_Y - \frac{\sigma_{XY}}{\sigma_X^2} \mu_X, w = \frac{\sigma_{XY}}{\sigma_X^2},$$

где

$$\mu_X = \frac{1}{n} \sum_{i=1}^n x_i, \mu_Y = \frac{1}{n} \sum_{i=1}^n y_i,$$

$$\sigma_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y),$$

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





Набор данных Ирисы (Iris) изучался Р. Фишером в 1936 г. Набор состоит из 150 записей и содержит следующие пять признаков:

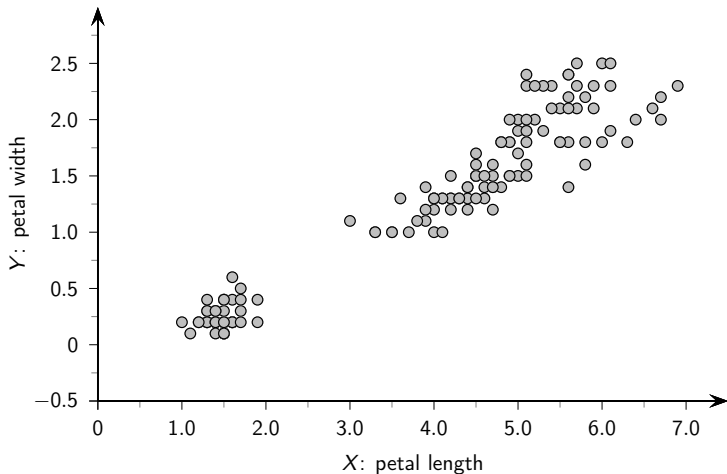
- ❶ длина чашелистника (sepal length) в см
- ❷ ширина чашелистника (sepal width) в см
- ❸ длина лепестка (petal length) в см
- ❹ ширина лепестка (petal width) в см
- ❺ класс (class) ириса, принимающий значения:
  - Iris-setosa
  - Iris-versicolour
  - Iris-virginica

Набор данных Ирисы размещен в репозитории данных машинного обучения UCI по адресу <https://archive.ics.uci.edu/ml/datasets/Iris>, количество обращений к набору с 2007 года составляет более 4 миллионов.

# Пример парной регрессии в наборе Ирисы



Даны два признака: длина лепестка (petal length)  $X$  (независимая переменная) и ширина лепестка (petal width)  $Y$  (отклик) в наборе данных Ирисы. 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

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

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

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



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

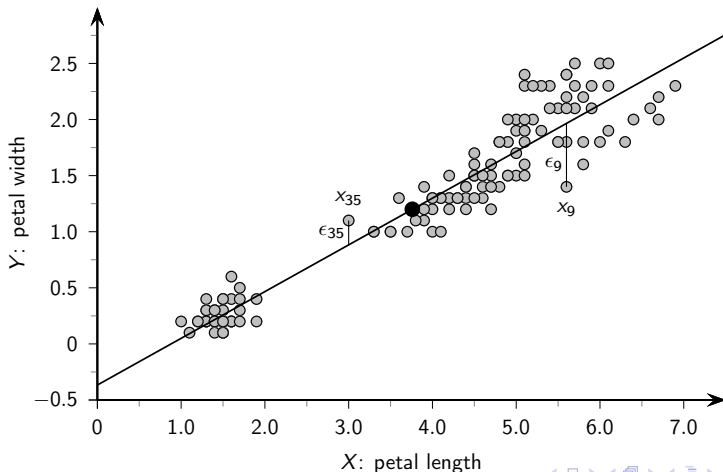
$$\hat{y} = -0.3665 + 0.4164 x$$

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  $\mathbf{Y}$ . 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:

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

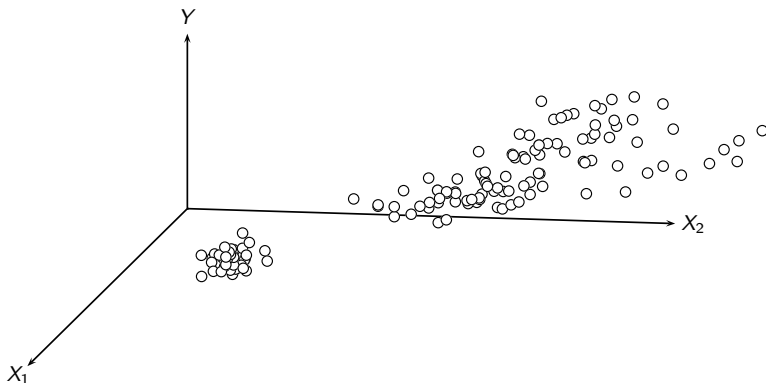
Дифференцируя SSE по  $\tilde{\mathbf{w}}$  и приравнивая результат к нулю, получим, что оптимальный вектор весов  $\tilde{\mathbf{w}}$  множественной регрессии задается формулой

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

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

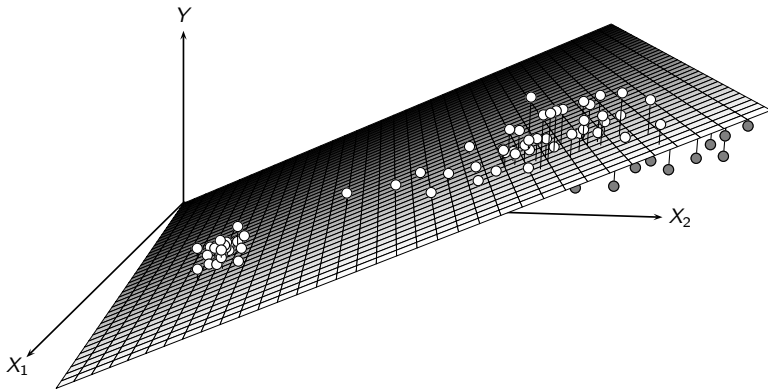
$$\tilde{\mathbf{w}} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T \mathbf{Y} = \begin{pmatrix} -0.014 \\ -0.082 \\ 0.45 \end{pmatrix}$$

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 дополненной матрицы данных  $\tilde{\mathbf{D}}$  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  $\Delta$  и  $\Delta^{-1}$  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 последнего 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

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

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  $\tilde{\mathbf{x}}_k$  набора  $\tilde{\mathbf{D}}$  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 стохастического градиентного спуска 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 и неопределены, 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 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.



Для построения точного решения дифференцируем функцию  $J(\tilde{\mathbf{w}})$  по  $\tilde{\mathbf{w}}$  и приравниваем 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).

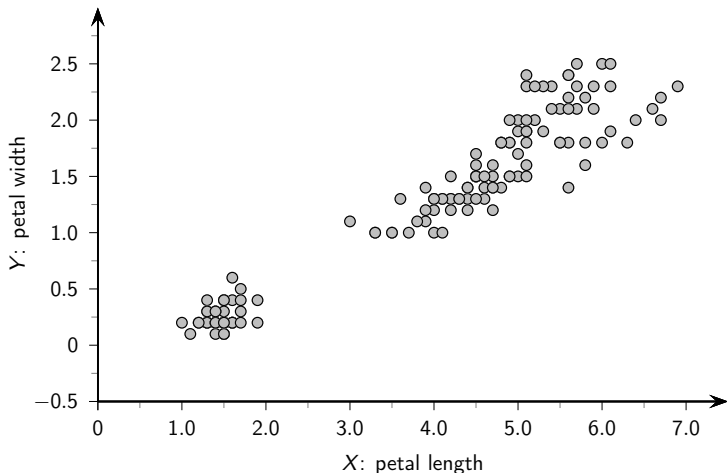
Если число  $\lambda_i$  является собственным значением матрицы  $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$ , то число  $\lambda_i + \alpha$  является собственным значением матрицы  $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}} + \alpha \mathbf{I}$ . Поскольку матрица  $\tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$  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

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

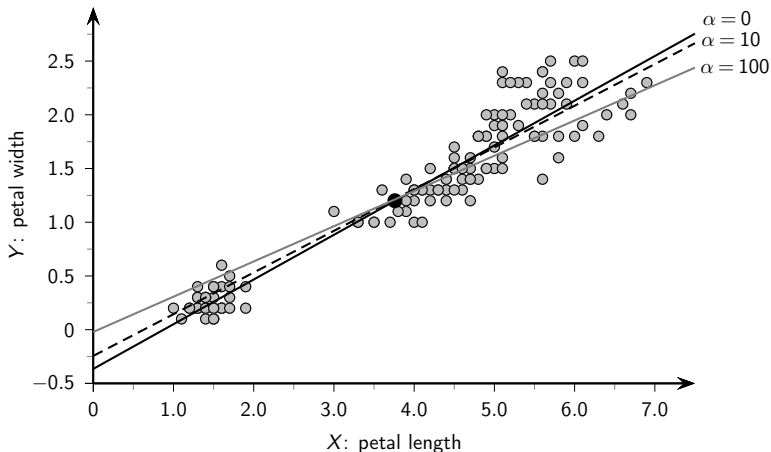
Получим различные линии наилучшего соответствия для различных значений параметра регуляризации  $\alpha$ :

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

# Линии гребневой регрессии для набора Ирисы



По мере увеличения  $\alpha$  больше внимания уделяется минимизации квадрата нормы  $\tilde{\mathbf{w}}$ . Поскольку с увеличением  $\alpha$  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  $\alpha$  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  $\alpha$ :

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

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

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

Полученные уравнения 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 использует норму  $L_1$  для регуляризации:

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

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  $w_i = 0$ .

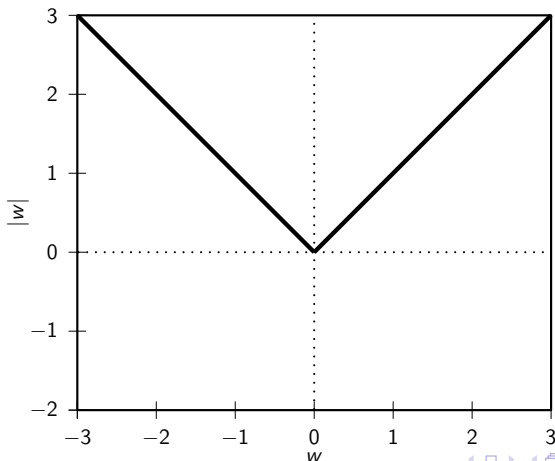
Поэтому мы не можем просто вычислить градиент и приравнять его к нулю, как это делается в случае гребневой регрессии. Задачу минимизации можно решить с помощью обобщенного подхода субградиентов.



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

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

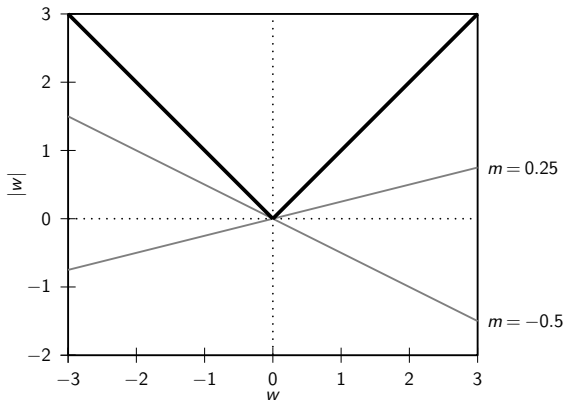
В точке  $w = 0$  производная не существует.





Понятие субградиента обобщает 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  $L_1$  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  $w = 0, \partial |w| \in [-1, +1]:$

$$\begin{aligned} w &\in [\eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} - \eta \alpha, \eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}} + \eta \alpha] \\ w = 0 &\Rightarrow |\eta \bar{\mathbf{X}}^T \bar{\mathbf{Y}}| \leq \eta \alpha \end{aligned}$$



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 двумерной регрессии  $L_1$  (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      $\nabla(w_k^{(t)}) \leftarrow -\bar{X}_k^T (\bar{Y} - \bar{\mathbf{D}} \mathbf{w}^{(t)})$  // 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
```



Применяем регрессию  $L_1$  к полному набору данных Ирисы с  $n = 150$  точками и четырьмя независимыми признаками, а именно шириной чашелистика  $X_1$  (sepal-width), длиной чашелистника  $X_2$  (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  $\alpha$  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 регрессии  $L_2$  (гребневая) и  $L_1$  (лассо) с одинаковым уровнем квадратичной ошибки.

При  $\alpha = 5$  модель регрессии  $L_1$  имеет ошибку  $SSE = 8.82$ .

Установим значение параметра  $\alpha = 35$  в регрессии  $L_2$ , 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 регрессии  $L_2$  коэффициенты при  $x_1$  и  $x_2$  малы и, следовательно, менее важны, но они не равны нулю.

В модели регрессии  $L_1$  коэффициенты для  $x_1$  и  $x_2$  в точности равны нулю, остаются только признаки  $x_3$  и  $x_4$ .

Таким образом, регрессия  $L_1$  (лассо) может осуществлять **отбор значимых признаков**.



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

