Machine Learning

myTRACKER

@ mail.ru group

МНК и ММП

Метод наименьших квадратов

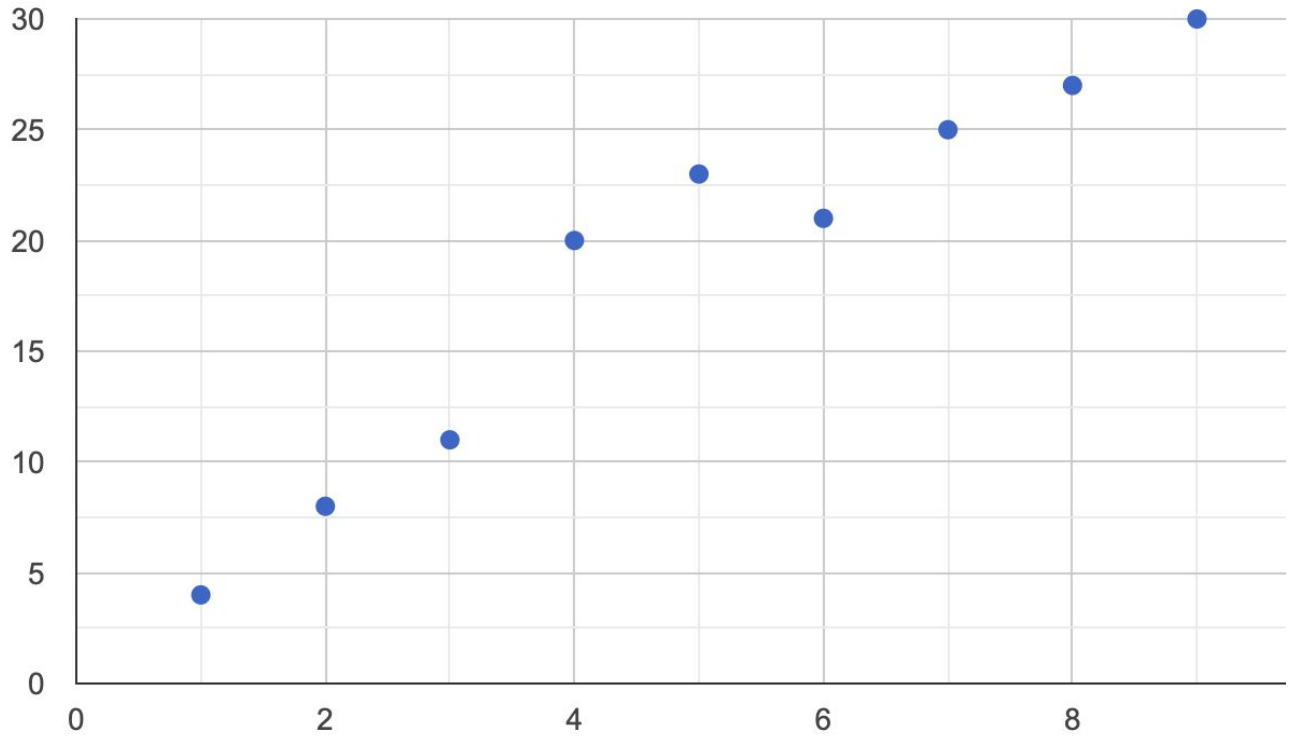
Пусть у нас есть некая зависимость одной переменной от другой

X	1	2	3	4	5	6	7	8	9
Y	4	8	15	20	23	24	25	27	30

Метод наименьших квадратов

Пусть у нас есть некая зависимость одной переменной от другой

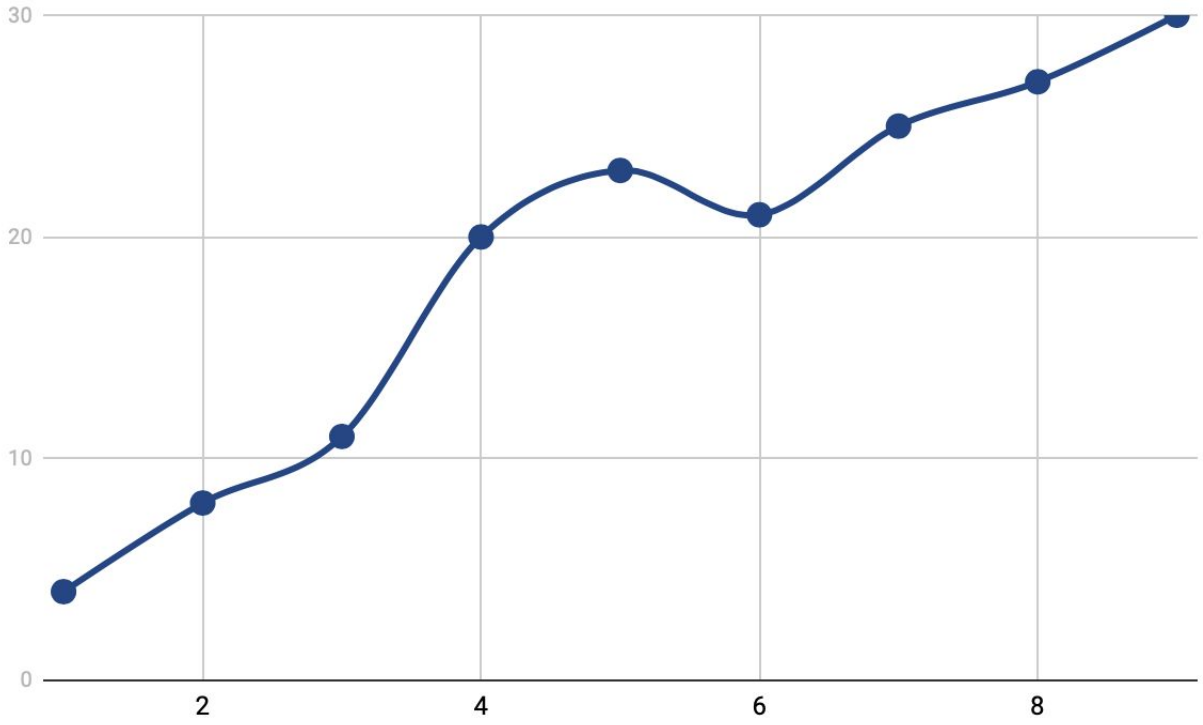
X	1	2	3	4	5	6	7	8	9
Y	4	8	15	20	23	24	25	27	30



Метод наименьших квадратов

Пусть у нас есть некая зависимость одной переменной от другой

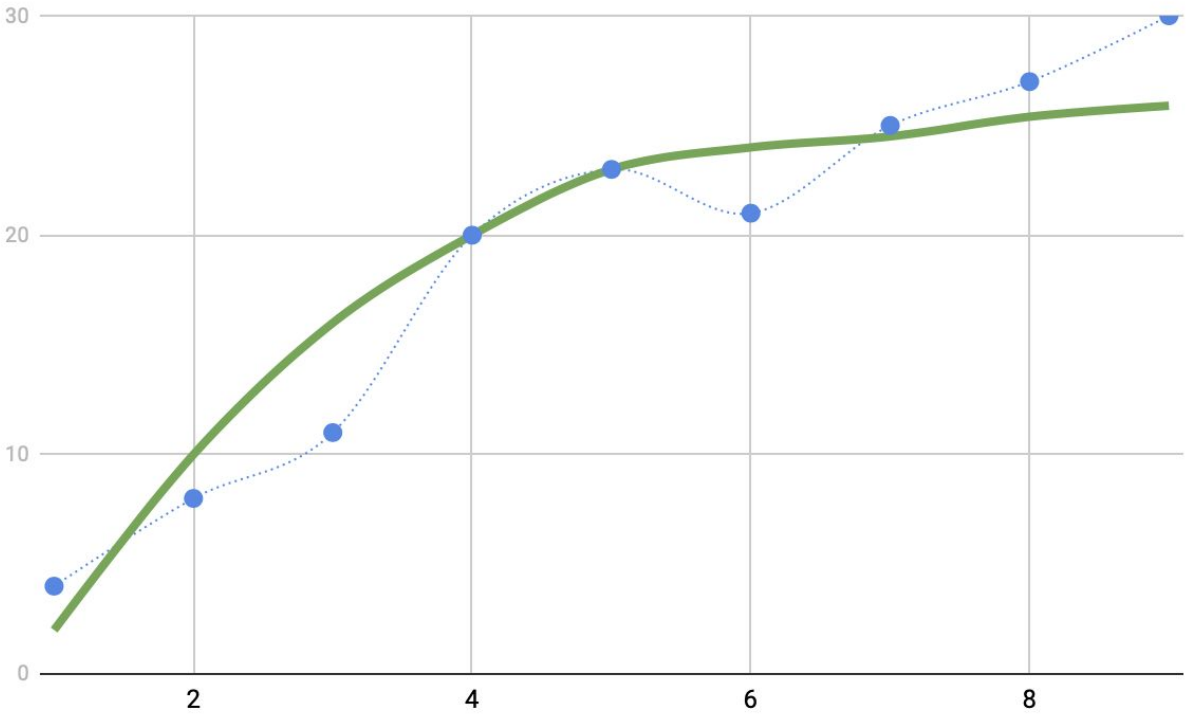
X	1	2	3	4	5	6	7	8	9
Y	4	8	11	20	23	21	25	27	30



Метод наименьших квадратов

Пусть у нас есть некая зависимость одной переменной от другой

X	1	2	3	4	5	6	7	8	9
Y	4	8	11	20	23	21	25	27	30



Метод наименьших квадратов

Определение. Пусть $f(x)$ - аппроксимирующая функция для набора точек $(x_i; y_i)$. Тогда ошибками будет называть $e_i = y_i - f(x_i)$.

Задача. Давайте оценивать аппроксимирующие функции с помощью ошибок.

Проблема. Как именно с помощью ошибок можно оценивать?

Ошибки - это набор точек, нужно придумать функцию, которая будет зависеть от ошибок и с помощью нее оценивать аппроксимирующие функции.

Метод наименьших квадратов

Определение. Пусть $f(x)$ - аппроксимирующая функция для набора точек $(x_i; y_i)$. Тогда ошибками будет называть $e_i = y_i - f(x_i)$.

Задача. Давайте оценивать аппроксимирующие функции с помощью ошибок.

Проблема. Как именно с помощью ошибок можно оценивать?

Варианты:

- **Простая сумма:** $e(x) = e_1 + \ldots + e_n$
- **Сумма модулей:** $e(x) = |e_1| + \ldots + |e_n|$
- **Сумма квадратов:** $e(x) = e_1^2 + \ldots + e_n^2$
- **Сумма больших степеней:** $e(x) = e_1^{10} + \ldots + e_n^{10}$

Метод наименьших квадратов

Определение. Пусть $f(x)$ - аппроксимирующая функция для набора точек $(x_i; y_i)$. Тогда ошибками будет называть $e_i = y_i - f(x_i)$.

Задача. Давайте оценивать аппроксимирующие функции с помощью ошибок.

Проблема. Как именно с помощью ошибок можно оценивать?

Варианты:

- **Простая сумма:** $e(x) = e_1 + \ldots + e_n$ Слагаемые могут сократиться между собой
- **Сумма модулей:** $e(x) = |e_1| + \ldots + |e_n|$
- **Сумма квадратов:** $e(x) = e_1^2 + \ldots + e_n^2$
- **Сумма больших степеней:** $e(x) = e_1^{10} + \ldots + e_n^{10}$ Сложно вычислять и слишком сильно “наказываем” за большие ошибки

Метод наименьших квадратов

Определение. Пусть $f(x)$ - аппроксимирующая функция для набора точек $(x_i; y_i)$. Тогда ошибками будет называть $e_i = y_i - f(x_i)$.

Задача. Давайте оценивать аппроксимирующие функции с помощью ошибок.

Проблема. Как именно с помощью ошибок можно оценивать?

Варианты:

- **Простая сумма:** $e(x) = e_1 + \ldots + e_n$ Слагаемые могут сократиться между собой
- **Сумма модулей:** $e(x) = |e_1| + \ldots + |e_n|$ Лучше подходит при нестандартном распределении ошибок
- **Сумма квадратов:** $e(x) = e_1^2 + \ldots + e_n^2$ Лучше подходит при нормальном и равномерном распределении ошибок
- **Сумма больших степеней:** $e(x) = e_1^{10} + \ldots + e_n^{10}$ Сложно вычислять и слишком сильно “наказываем” за большие ошибки

Метод наименьших квадратов

Определение. Пусть $f(x)$ - аппроксимирующая функция для набора точек $(x_i; y_i)$. Тогда ошибками будет называть $e_i = y_i - f(x_i)$.

Задача. Давайте оценивать аппроксимирующие функции с помощью ошибок.

Проблема. Как именно с помощью ошибок можно оценивать?

Варианты:

- **Простая сумма:** $e(x) = e_1 + \ldots + e_n$ Слагаемые могут сократиться между собой
- **Сумма модулей:** $e(x) = |e_1| + \ldots + |e_n|$ Лучше подходит при нестандартном распределении ошибок
- **Сумма квадратов:** $e(x) = e_1^2 + \ldots + e_n^2$ Лучше подходит при нормальном и равномерном распределении ошибок
- **Сумма больших степеней:** $e(x) = e_1^{10} + \ldots + e_n^{10}$ Сложно вычислять и слишком сильно “наказываем” за большие ошибки

В прикладных задачах чаще встречается нормальное распределение

Метод наименьших квадратов

Определение. Пусть задана такая зависимость: $y_t = f(x_t, b) + \varepsilon_t$, где ε_t - случайная ошибка модели и b - набор неизвестных параметров. Надо восстановить изначальную зависимость y от x. Для этого подберем параметры b наилучшим образом.

Метод наименьших квадратов

Определение. Пусть задана такая зависимость: $y_t = f(x_t, b) + \varepsilon_t$, где ε_t - случайная ошибка модели и b - набор неизвестных параметров. Надо восстановить изначальную зависимость y от x. Для этого подберем параметры b наилучшим образом.

Определение. Введем функцию “ошибки”, с помощью которой будем оценивать параметры b

$$RSS(b) = e^T e = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - f(x_t, b))^2$$

Метод наименьших квадратов

Определение. Пусть задана такая зависимость: $y_t = f(x_t, b) + \varepsilon_t$, где

ε_t - случайная ошибка модели и b - набор неизвестных параметров. Надо восстановить изначальную зависимость y от x. Для этого подберем параметры b наилучшим образом.

Определение. Введем функцию “ошибки”, с помощью которой будем оценивать параметры b

$$RSS(b) = e^T e = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - f(x_t, b))^2$$

Задача. Найти $\hat{b}_{OLS} = \arg\min_b RSS(b)$

Метод наименьших квадратов

Определение. Пусть задана такая зависимость: $y_t = f(x_t, b) + \varepsilon_t$, где ε_t - случайная ошибка модели и b - набор неизвестных параметров. Надо восстановить изначальную зависимость y от x. Для этого подберем параметры b наилучшим образом.

Определение. Введем функцию “ошибки”, с помощью которой будем оценивать параметры b

$$RSS(b) = e^T e = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - f(x_t, b))^2$$

Задача. Найти $\hat{b}_{OLS} = \arg\min_b RSS(b)$

Решение. Задачу можно решить с помощью методов оптимизации, а можно попытаться решить аналитически. Большинство задач можно решить аналитически, так что будем разбирать этот метод:

$$\sum_{t=1}^{n} (y_t - f(x_t, b)) \frac{\partial f(x_t, b)}{\partial b} = 0.$$

Метод наименьших квадратов (линейная модель)

Определение. Пусть задана линейная зависимость

$$y_t = \sum_{j=1}^{k} b_j x_{tj} + \varepsilon = x_t^T b + \varepsilon_t \quad <-> \quad y = Xb + \varepsilon.$$

Функциональное представление

Матричное представление

$$y_1 = b_1 x_{11} + b_2 x_{12}$$
$$y_2 = b_1 x_{21} + b_2 x_{22}$$

Feature 1	Feature 2	Target
x_{11}	x_{12}	y_1
x_{21}	x_{22}	y_2

Метод наименьших квадратов (линейная модель)

Определение. Пусть задана линейная зависимость

$$y_t = \sum_{j=1}^{k} b_j x_{tj} + \varepsilon = x_t^T b + \varepsilon_t \quad <-> \quad y = Xb + \varepsilon.$$

Функциональное представление

Матричное представление

Определение. Функция ошибки в матричном представлении имеет вид

$$RSS = e^T e = (y - Xb)^T (y - Xb)$$

Метод наименьших квадратов (линейная модель)

$$RSS = (Xb - y)^T (Xb - y) = (b^T X^T - y^T)(Xb - y)$$

$$RSS = b^T X^T Xb - b^T X^T y - y^T Xb + y^T y$$

Метод наименьших квадратов (линейная модель)

$$RSS = (Xb - y)^T (Xb - y) = (b^T X^T - y^T)(Xb - y)$$

$$RSS = b^T X^T Xb - b^T X^T y - y^T Xb + y^T y$$

Слагаемое	Формула производной	Слагаемое	Формула производной
$\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$	$\frac{d\,\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$	$\mathbf{y}^T\mathbf{X}\mathbf{w}$	$\frac{d\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{A}$
$\mathbf{w}^T\mathbf{X}^T\mathbf{y}$	$\frac{d\,\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \frac{d\,\mathbf{x}^\top\mathbf{a}}{d\mathbf{x}} = \mathbf{a}^\top$	$\mathbf{y}^T\mathbf{y}$	$\frac{da}{d\mathbf{x}} = \mathbf{0}^\top$ (row matrix)

$$\frac{dRSS}{db} = b^T(X^T X + X^T X) - (X^T y)^T - y^T X + 0$$

Метод наименьших квадратов (линейная модель)

$$RSS = (Xb - y)^T (Xb - y) = (b^T X^T - y^T)(Xb - y)$$

$$RSS = b^T X^T Xb - b^T X^T y - y^T Xb + y^T y$$

Слагаемое	Формула производной	Слагаемое	Формула производной
$\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$	$\frac{d\,\mathbf{x}^\top\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$	$\mathbf{y}^T\mathbf{X}\mathbf{w}$	$\frac{d\mathbf{A}\mathbf{x}}{d\mathbf{x}} = \mathbf{A}$
$\mathbf{w}^T\mathbf{X}^T\mathbf{y}$	$\frac{d\,\mathbf{a}^\top\mathbf{x}}{d\mathbf{x}} = \frac{d\,\mathbf{x}^\top\mathbf{a}}{d\mathbf{x}} = \mathbf{a}^\top$	$\mathbf{y}^T\mathbf{y}$	$\frac{da}{d\mathbf{x}} = \mathbf{0}^\top$ (row matrix)

$$\frac{dRSS}{db} = b^T(X^T X + X^T X) - (X^T y)^T - y^T X + 0$$

$$\frac{dRSS}{db} = 2b^T X^T X - 2y^T X = 0$$

$$2b^T X^T X - 2y^T X = 0 \qquad b = (X^T X)^{-1} X^T y$$

Метод наименьших квадратов (линейная модель)

Определение. Пусть задана линейная зависимость

$$y_t = \sum_{j=1}^{k} b_j x_{tj} + \varepsilon = x_t^T b + \varepsilon_t \quad <-> \quad y = Xb + \varepsilon.$$

Функциональное представление

Матричное представление

Определение. Функция ошибки в матричном представлении имеет вид

$$RSS = e^T e = (y - Xb)^T (y - Xb)$$

Если продифференцировать по вектору параметров b и приравняем производную к нулю, получаем

$$(X^T X)b = X^T y.$$

Подробнее про матричные производные: https://vk.cc/c3RNbp

Метод максимального правдоподобия

Будем пытаться оценивать неизвестные параметры с помощью математической статистики, но в общем задача остается такой же.

Метод максимального правдоподобия

Будем пытаться оценивать неизвестные параметры с помощью математической статистики, но в общем задача остается такой же.

Задача. Вы зашли в метро на незнакомой станции метро. И неожиданно для вас на этой станции метро сейчас мало людей и вы комфортно садитесь в вагон.

Гипотезы

- **Гипотеза 1.** Каждый день в это время на этой станции метро мало людей.
- **Гипотеза 2**. Раз в неделю есть день, когда на этой станции метро мало людей и вы попали именно в этот день.
- **Гипотеза 3**. Раз в месяц есть день, когда на этой станции метро мало людей и вы попали именно в этот день.
- **Гипотеза 4**. Раз в год есть день, когда на этой станции метро мало людей и вы попали именно в этот день.

Какую гипотезу выбрали бы вы?

Метод максимального правдоподобия

Будем пытаться оценивать неизвестные параметры с помощью математической статистики, но в общем задача остается такой же.

Задача. Вы зашли в метро на незнакомой станции метро. И неожиданно для вас на этой станции метро сейчас мало людей и вы комфортно садитесь в вагон.

Гипотезы

- **Гипотеза 1.** Каждый день в это время на этой станции метро мало людей.
- **Гипотеза 2**. Раз в неделю есть день, когда на этой станции метро мало людей и вы попали именно в этот день.
- **Гипотеза 3**. Раз в месяц есть день, когда на этой станции метро мало людей и вы попали именно в этот день.
- **Гипотеза 4**. Раз в год есть день, когда на этой станции метро мало людей и вы попали именно в этот день.

Метод максимального правдоподобия выберет гипотезу с максимальной вероятностью, а значит гипотезу 1.

Метод максимального правдоподобия

Определение. Пусть есть выборка $X_1, \ldots, X_n$ из распределения P_θ, где $\theta \in \Theta$ - неизвестные параметры.

Метод максимального правдоподобия

Определение. Пусть есть выборка $X_1, \ldots, X_n$ из распределения P_θ, где $\theta \in \Theta$ - неизвестные параметры.

Определение. Назовем $L(\mathbf{x} \mid \theta) \colon \Theta \to \mathbb{R}$ функцией правдоподобия, где $\mathbf{x} \in \mathbb{R}^n$

$L = \prod p(x_i|\theta)$ в случае дискретного распределения

$L = \prod f(x_i|\theta)$ в случае непрерывного распределения

Метод максимального правдоподобия

Определение. Пусть есть выборка $X_1, \ldots, X_n$ из распределения P_θ, где $\theta \in \Theta$ - неизвестные параметры.

Определение. Назовем $L(\mathbf{x} \mid \theta): \Theta \to \mathbb{R}$ функцией правдоподобия, где $\mathbf{x} \in \mathbb{R}^n$

$L = \prod p(x_i|\theta)$ в случае дискретного распределения

$L = \prod f(x_i|\theta)$ в случае непрерывного распределения

Будем искать точечную оценку для параметров

Определение. Точечную оценку $\hat{\theta}_{\text{МП}} = \hat{\theta}_{\text{МП}}(X_1, \ldots, X_n) = \underset{\theta \in \Theta}{\operatorname{argmax}} L(X_1, \ldots, X_n \mid \theta)$ будем называть оценкой максимального правдоподобия параметра θ.

То есть оценка ММП - это такая точечная оценка, при которой функция правдоподобия достигает своего максимума при заданных параметрах.

Свойства ММП

Метод максимального правдоподобия обладает несколькими очень полезными свойствами, которые выделяют его на фоне остальных

- Оценки ММП состоятельны, то есть $\hat{\theta}_{ML} \to \theta$ при $n \to \infty$
- Оценки ММП асимптотически несмещенные, то есть $M(\hat{\theta}_{ML}) \to \theta$ при $n \to \infty$
- Оценки ММП асимптотически эффективны, то есть дисперсия $D(\hat{\theta}_{ML})$ будет наименьшей среди асимптотически несмещенных оценок
- Оценки ММП асимптотически нормальны, то есть $\hat{\theta}_{ML} \sim N(\theta, I^{-1})$ при $n \to \infty$, где I - информация Фишера, $I = -ln(L''(\theta))$
- МНК является частным случаем ММП, если мы считаем что ошибка распределена по правилу: $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Проблема оптимальной остановки

Метод максимального правдоподобия иногда могут давать некорректные результаты. Все зависит от дизайна эксперимента.

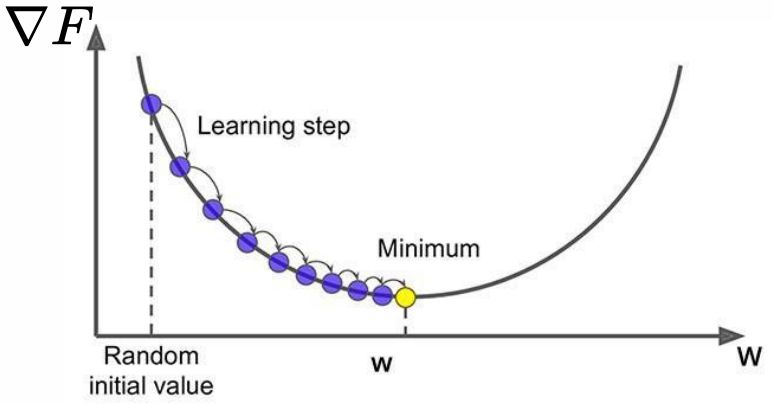
Пример. Предположим я сказал, что бросил монету 12 раз и получил 3 решки. Из этого вы сможете сделать некоторые выводы о вероятности выпадения решки у этой монеты. А теперь предположим, что я бросал монету пока решка не выпала 3 раза и бросал также 12 раз. Сделаете ли вы теперь другие выводы?

В обоих случаях функция правдоподобия будет одинакова и равна

$$p^3 (1-p)^9$$

Градиентный спуск

$$F = (w - 5)^4$$



Градиентный спуск описывается простой формулой:

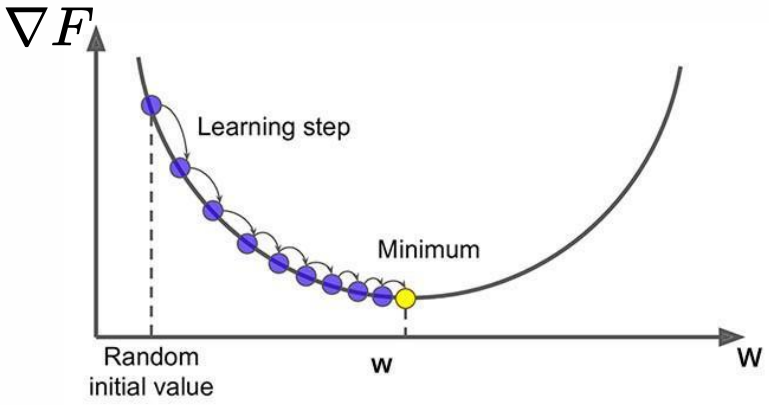
$$w_{n+1} = w_n - \gamma_n \nabla F(w_n)$$

Возьмем начальную точку: $w_0 = 3$

Пусть $\gamma_n = 0.01$

Градиентный спуск

$$F = (w-5)^4$$



Градиентный спуск описывается простой формулой:

$$w_{n+1} = w_n - \gamma_n \nabla F(w_n)$$

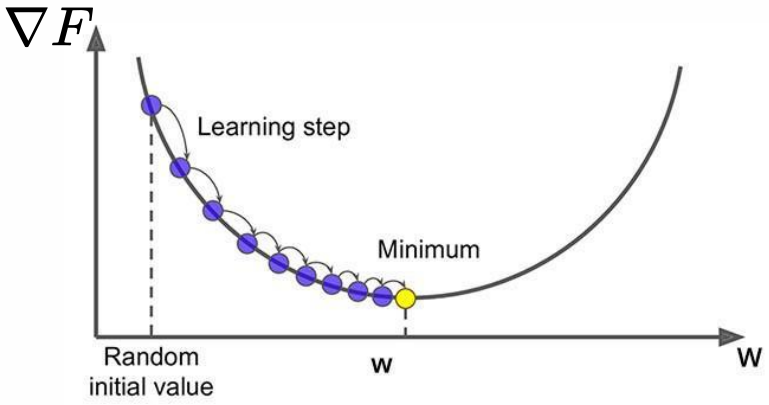
Возьмем начальную точку: $w_0 = 3$

Пусть $\gamma_n = 0.01$

$$\nabla F(w) = \frac{dF}{dw} = 4(w-5)^3$$

Градиентный спуск

$$F = (w-5)^4$$



Градиентный спуск описывается простой формулой:

$$w_{n+1} = w_n - \gamma_n \nabla F(w_n)$$

Возьмем начальную точку: $w_0 = 3$

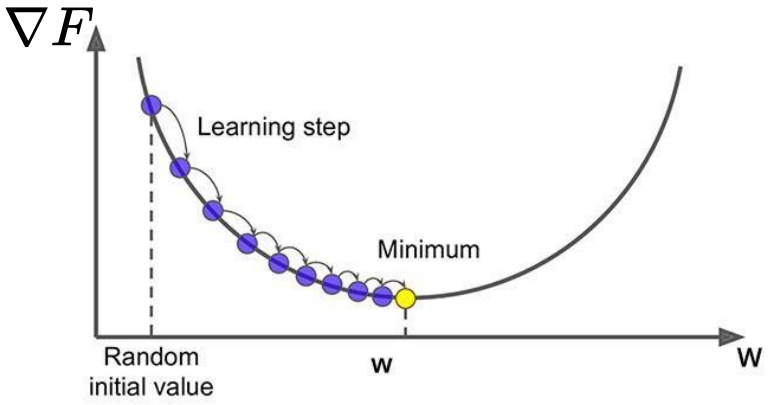
Пусть $\gamma_n = 0.01$

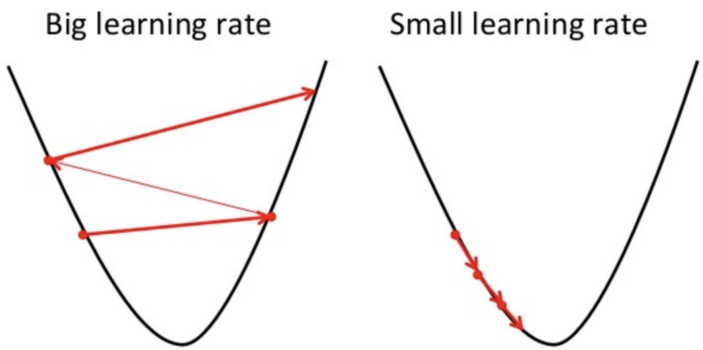
$$\nabla F(w) = \frac{dF}{dw} = 4(w-5)^3$$

$$\nabla F(w_0) = 4(w_0-5)^3 = -32$$

Градиентный спуск

$$F = (w-5)^4$$





Градиентный спуск описывается простой формулой:

$$w_{n+1} = w_n - \gamma_n \nabla F(w_n)$$

Возьмем начальную точку: $w_0 = 3$

Пусть $\gamma_n = 0.01$

$$\nabla F(w) = \frac{dF}{dw} = 4(w-5)^3$$

$$\nabla F(w_0) = 4(w_0-5)^3 = -32$$

$$w_1 = w_0 - \gamma_1 \nabla F(w_0) = 3.32$$