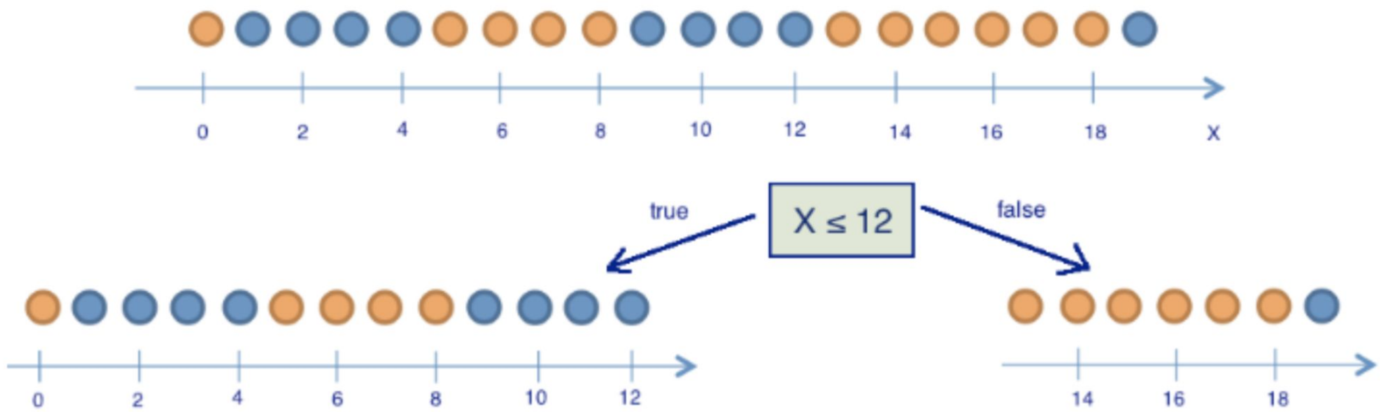
Курсы по машинному обучению

Тема 6. Линейные модели

Деревья решений (регрессия)



Критерий. Выборочное среднее

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$$

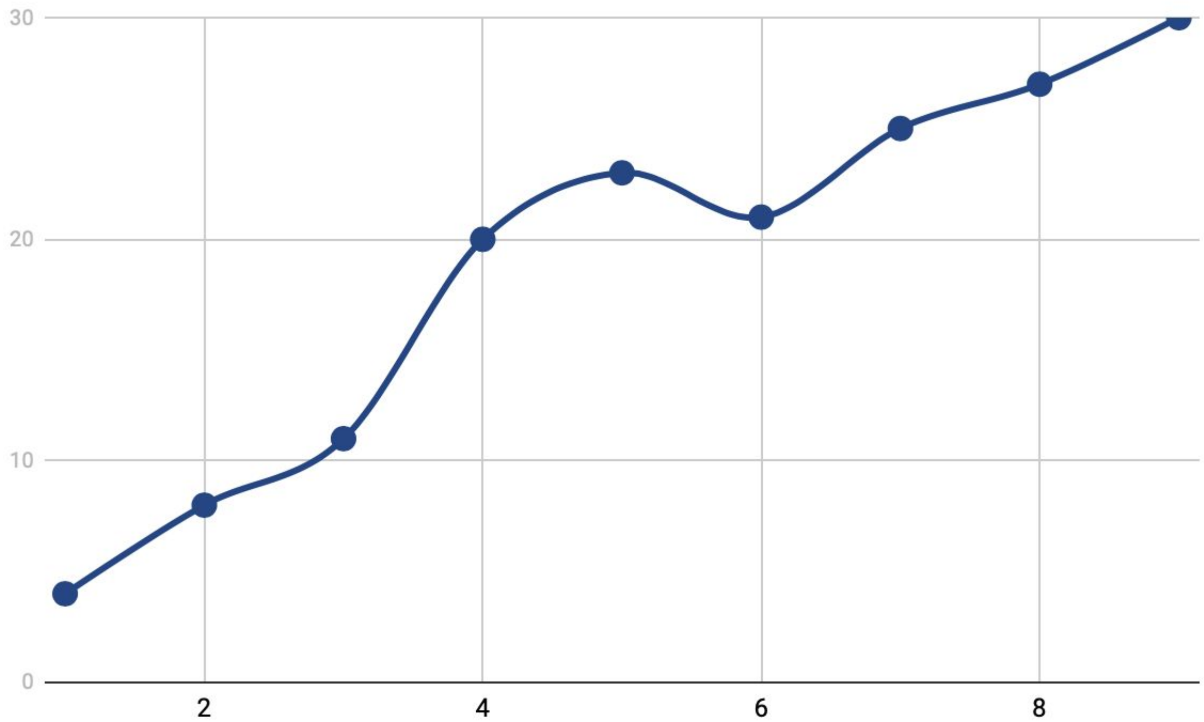
Ответ. Среднее значений в листе

$$X_{leaf} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Метод наименьших квадратов

Пусть у нас есть некая зависимость одной переменной от другой

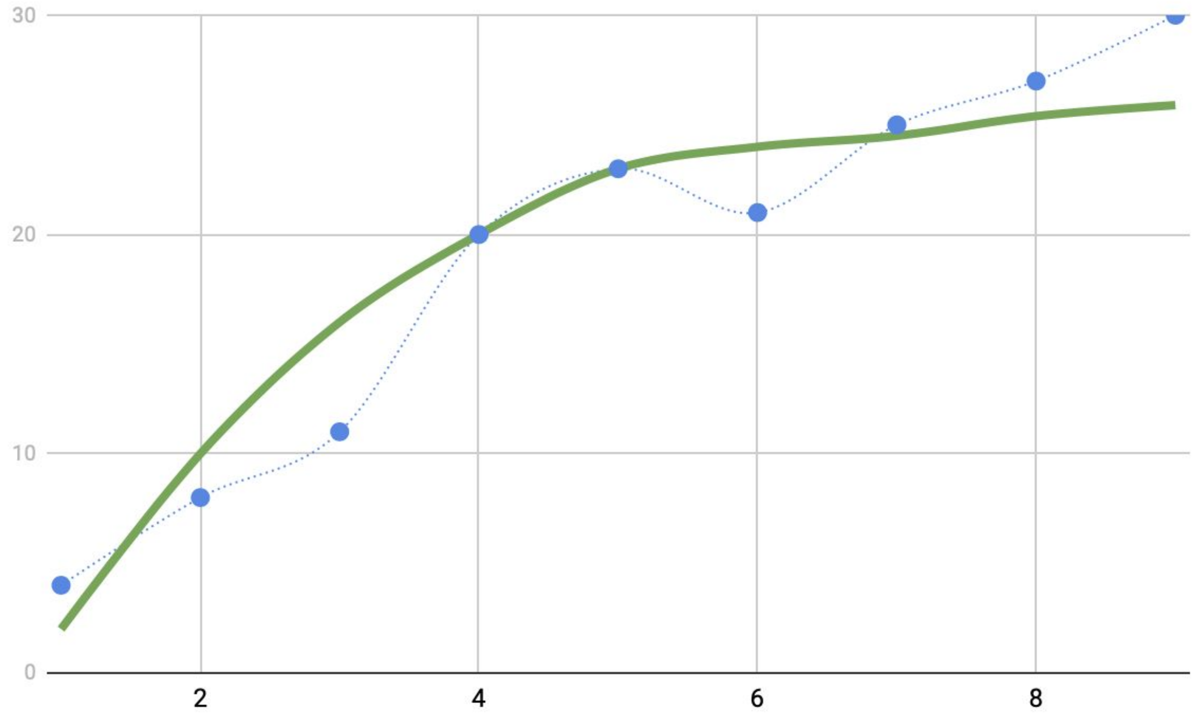
X	1	2	3	4	5	6	7	8	9
Y	4	8	11	20	23	21	25	27	30



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 (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

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)

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

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

Регуляризация

В итоге решение линейной регрессии можно найти по формуле:

$$b = (X^T X)^{-1} X^T y$$

Регуляризация

В итоге решение линейной регрессии можно найти по формуле:

$$b = (X^T X)^{-1} X^T y$$

Проблема. Матрица $X^T X$ становится необратимой.

Решение. Задачу надо изменить и сделать матрицу обратимой или регулярной.

Регуляризация

В итоге решение линейной регрессии можно найти по формуле:

$$b = (X^T X)^{-1} X^T y$$

Проблема. Матрица $X^T X$ становится необратимой.

Решение. Задачу надо изменить и сделать матрицу обратимой или регулярной.

Проблема. Матрица с мультиколлинарными столбцами дает нестабильную оценку параметров.

Решение. Добавить ограничение на параметры.

$$Error = RSS + \lambda b^T b$$

Регуляризация

В итоге решение линейной регрессии можно найти по формуле:

$$b = (X^T X)^{-1} X^T y$$

Проблема. Матрица $X^T X$ становится необратимой.

Решение. Задачу надо изменить и сделать матрицу обратимой или регулярной.

Проблема. Матрица с мультиколлинарными столбцами дает нестабильную оценку параметров.

Решение. Добавить ограничение на параметры.

$$Error = RSS + \lambda b^T b$$

В итоге получаем гребневую регрессию (Ridge regression):

$$b = (X^T X + \lambda E)^{-1} X^T y$$

Линейные регрессии с регуляризациями

Ridge regression (гребневая регрессия, регрессия с $L2$ регуляризацией)

$$Error = RSS + \lambda \sum_i b_i^2$$

LASSO regression (лассо регрессия, регрессия с $L1$ регуляризацией)

$$Error = RSS + \lambda \sum_i |b_i|$$

Можно комбинировать регуляризации (Elastic Net regression)

$$Error = RSS + \lambda_1 \sum_i |b_i| + \lambda_2 \sum_i b_i^2$$

Линейные регрессии с регуляризациями

Ridge regression (гребневая регрессия, регрессия с $L2$ регуляризацией)

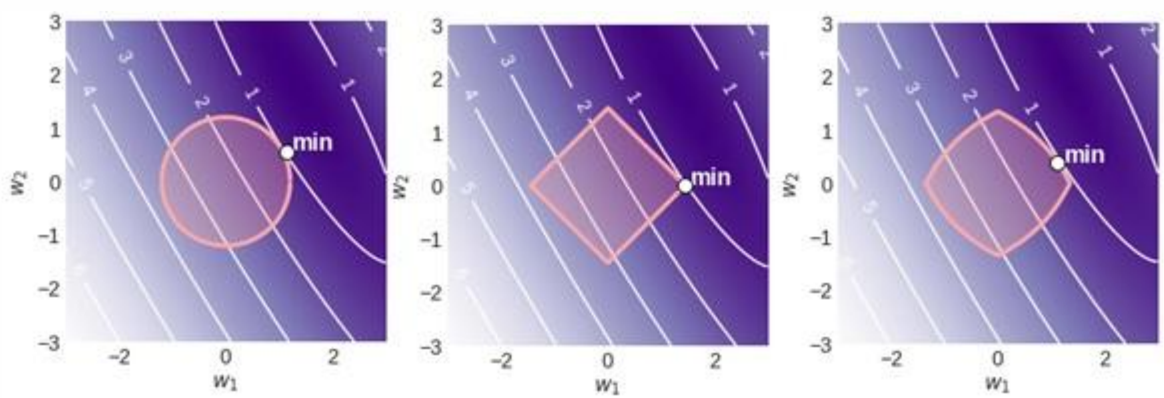
$$Error = RSS + \lambda \sum_i b_i^2$$

LASSO regression (лассо регрессия, регрессия с $L1$ регуляризацией)

$$Error = RSS + \lambda \sum_i |b_i|$$

Можно комбинировать регуляризации (Elastic Net regression)

$$Error = RSS + \lambda_1 \sum_i |b_i| + \lambda_2 \sum_i b_i^2$$



L_2 L_1 $L_1 + L_2$

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 $L(\mathbf{x} \mid \theta)\colon \Theta \to \mathbb{R}$ 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

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: $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Логистическая регрессия

Определение. Логит-функцией (сигмоидой) назовем функцию вида:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Определение. Вероятностью отнесения объекта к положительному классу будем рассчитывать по формуле:

$$p_+(x) = P(y = 1|x, b) = \sigma(b^T x)$$

Логистическая регрессия

Определение. Логит-функцией (сигмоидой) назовем функцию вида:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Определение. Вероятностью отнесения объекта к положительному классу будем рассчитывать по формуле:

$$p_+(x) = P(y=1|x,b) = \sigma(b^T x)$$

Определение. Вероятностью отнесения объекта к отрицательному классу будем рассчитывать по формуле:

$$p_-(x) = P(y=-1|x,b) = 1 - \sigma(b^T x) = \sigma(-b^T x)$$

Определение. Вероятностью отнесения объекта к своему классу будем рассчитывать по формуле:

$$p(x) = \sigma(yb^T x)$$

Логистическая регрессия

Вопрос. Как обучать такую модель?

Ответ. Воспользуемся методом максимального правдоподобия.

Логистическая регрессия

Вопрос. Как обучать такую модель?

Ответ. Воспользуемся методом максимального правдоподобия.

Решение.

1. Построим функцию максимального правдоподобия

$$L = \prod p(x_i|\theta) = \prod_i \sigma(y_i b^T x_i)$$

Логистическая регрессия

Вопрос. Как обучать такую модель?

Ответ. Воспользуемся методом максимального правдоподобия.

Решение.

1. Построим функцию максимального правдоподобия

$$L = \prod p(x_i|\theta) = \prod_i \sigma(y_i b^T x_i)$$

2. Применим логарифм к функции правдоподобия

$$\log L = \log \prod_i \sigma(y_i b^T x_i) = \sum_i \log \sigma(y_i b^T x_i) = -\sum_i \log(1 + e^{y_i b^T x_i})$$

Логистическая регрессия

Вопрос. Как обучать такую модель?

Ответ. Воспользуемся методом максимального правдоподобия.

Решение.

1. Построим функцию максимального правдоподобия

$$L = \prod p(x_i|\theta) = \prod_i \sigma(y_i b^T x_i)$$

2. Применим логарифм к функции правдоподобия

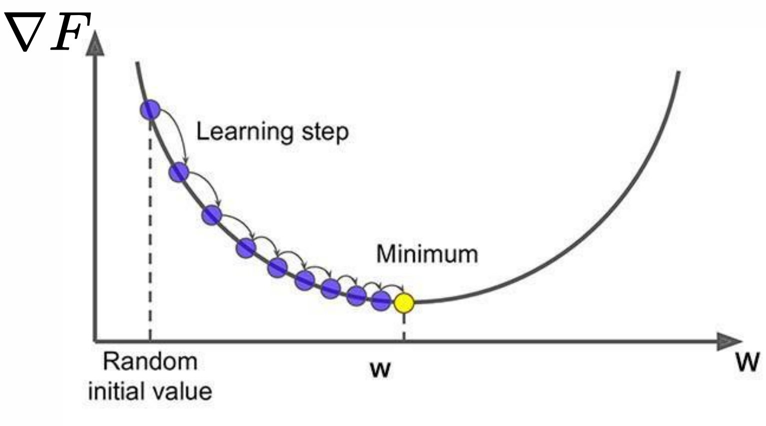
$$\log L = \log \prod_i \sigma(y_i b^T x_i) = \sum_i \log \sigma(y_i b^T x_i) = -\sum_i \log(1 + e^{y_i b^T x_i})$$

3. Получаем следующий функционал, который надо минимизировать:

$$Error = \sum_i \log(1 + e^{y_i b^T x_i})$$

Градиентный спуск

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



Градиентный спуск описывается простой формулой:

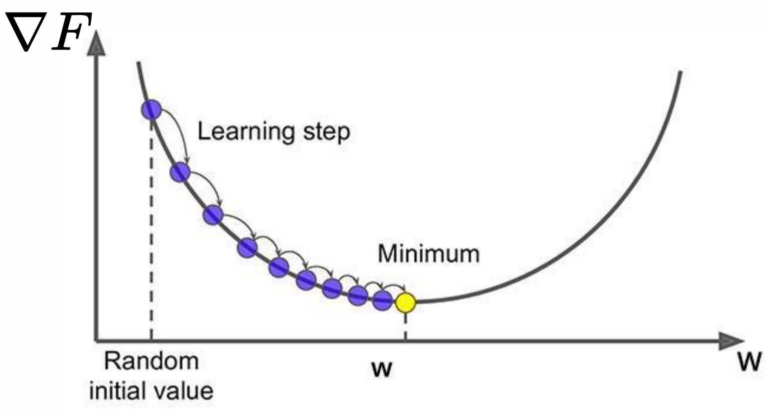
$$w_{n+1} = w_n - \gamma_n \nabla F(w_n)$$

Возьмем начальную точку: $w_0 = 3$

Пусть $\gamma_n = 0.01$

Градиентный спуск

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



Градиентный спуск описывается простой формулой:

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

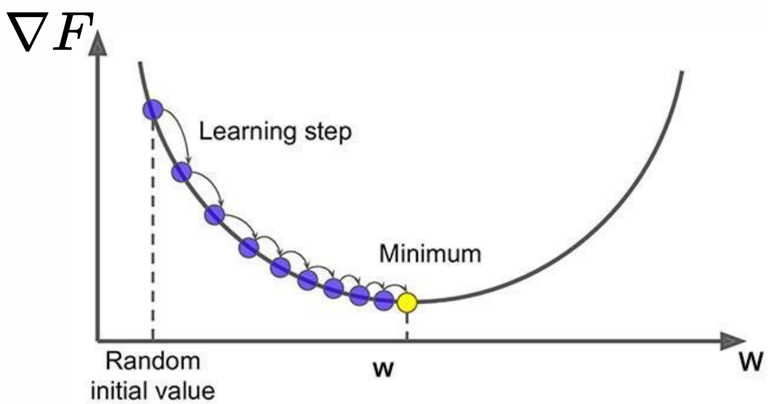
Возьмем начальную точку: $w_0 = 3$

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: $w_0 = 3$

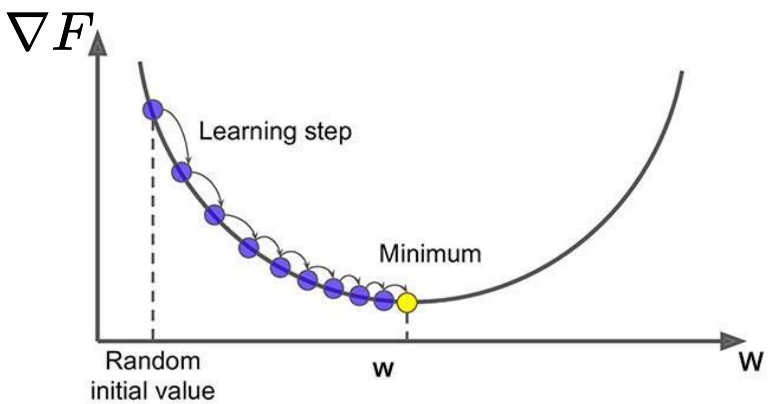
Пусть $\gamma_n = 0.01$

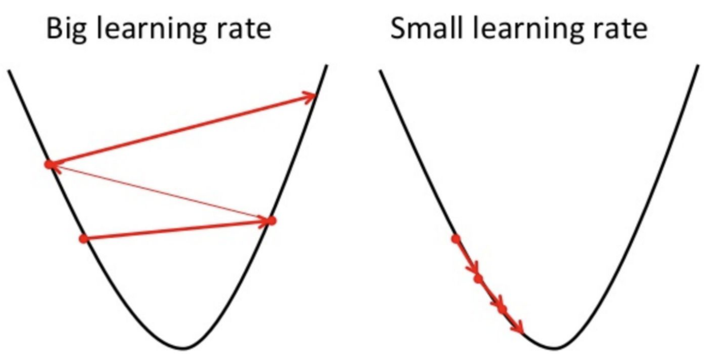
$$\nabla F(w) = \frac{dF}{dw} = 4(w - 5)^3$$

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

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 $\gamma_n = 0.01$

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

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

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