GeekBrains

Задача классификации. Логистическая регрессия

Урок 3

План курса

1. Введение в машинное обучение. Основные определения
2. Задача регрессии
3. Задача классификации
4. Метод опорных векторов
5. Оценка точности модели
6. Качество данных
7. Подготовка данных
8. Деревья решений
9. Ансамблирование. Алгоритм случайного леса
10. Работа с признаками
11. Feature selection. Отбор признаков
12. Качество модели и ее улучшение

Что будет на уроке сегодня

📌 Задача классификации

📌 Линейный классификатор

📌 Логистическая регрессия

📌 Градиентный спуск

📌 Стохастический градиентный спуск

📌 Практика

Линейный классификатор

→

Модель линейной регрессии

Формула линейной регрессии:

$$a(x) = w_0 + w_1 x_1 + \dots + w_d x_l$$

Функция потерь, которую мы минимизируем:

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

Пример

Задача	Клиент отдаст кредит или нет
Объекты(x)	Клиенты
Целевая переменная(y)	1, -1

Линейная регрессия

Пусть w_0 будет признаком, который для всех элементов будет равен единице. Тогда наша формула приобретет вид:

$$a(x) = \langle wx \rangle$$

Бинарный классификатор

sign - знак скалярного произведения

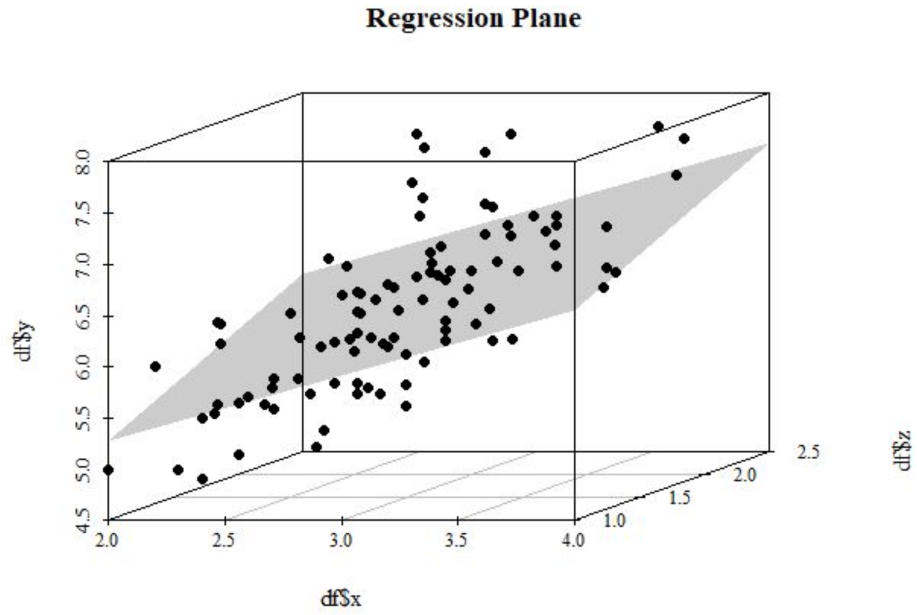
$$a(x) = \mathbf{sign}\langle wx \rangle$$

Бинарный классификатор

[] - индикатор события

$$a(x) = [(w,\ x) > 0]$$

Гиперплоскость



Решение задачи

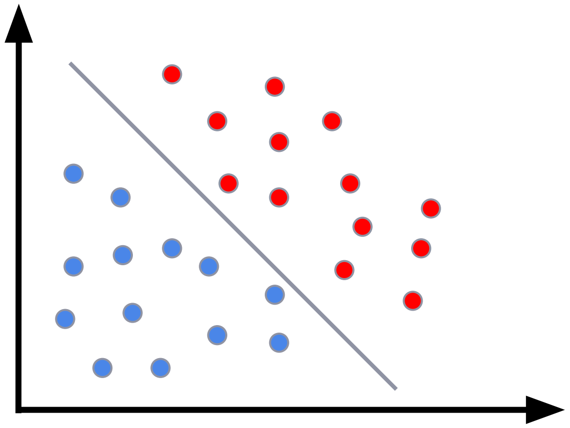
$$a(x) = \mathrm{sign}(w_0 + w_1x_1 + w_2x_2 + w_3x_3)$$

x_1 - зарплата клиента банка

x_2 - количество трат в месяц

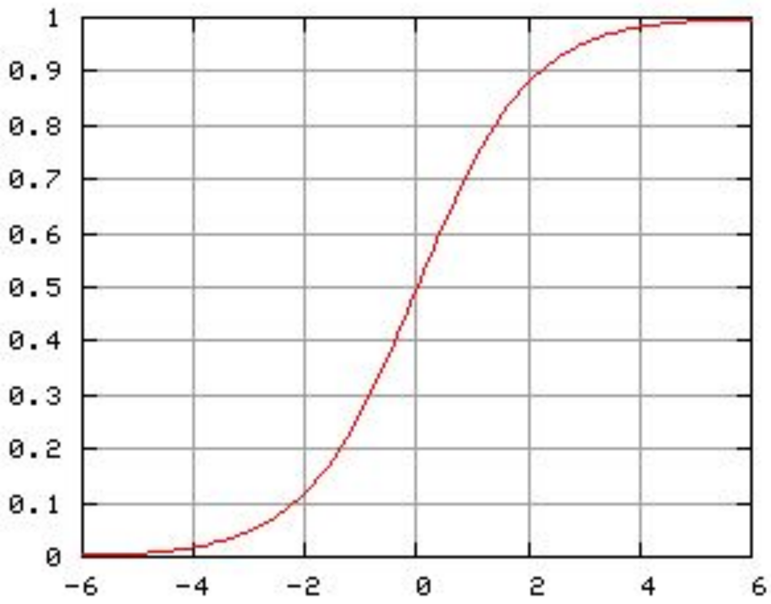
x_3 - задолженность по прошлым кредитам

Уверенность классификатора



Сигмоида

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

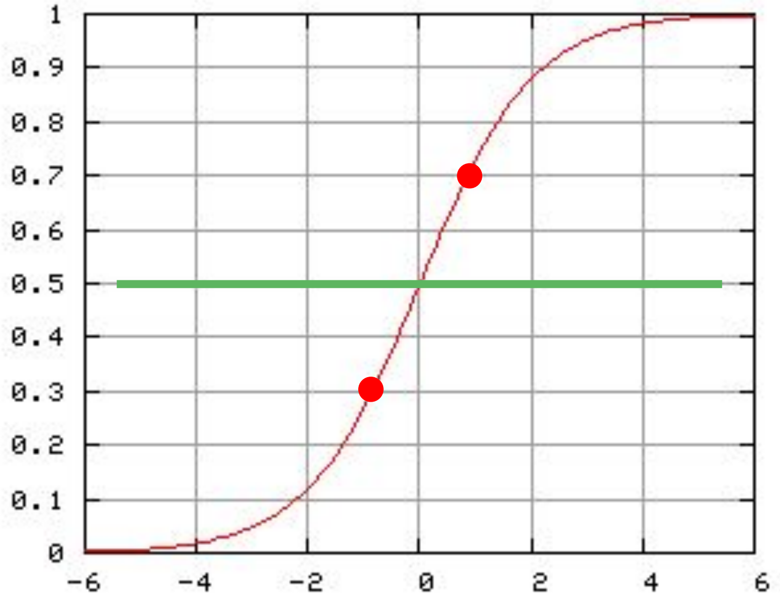


Формула логистической регрессии

$$a(x) = \frac{1}{1 + e^{-(w,x)}}$$

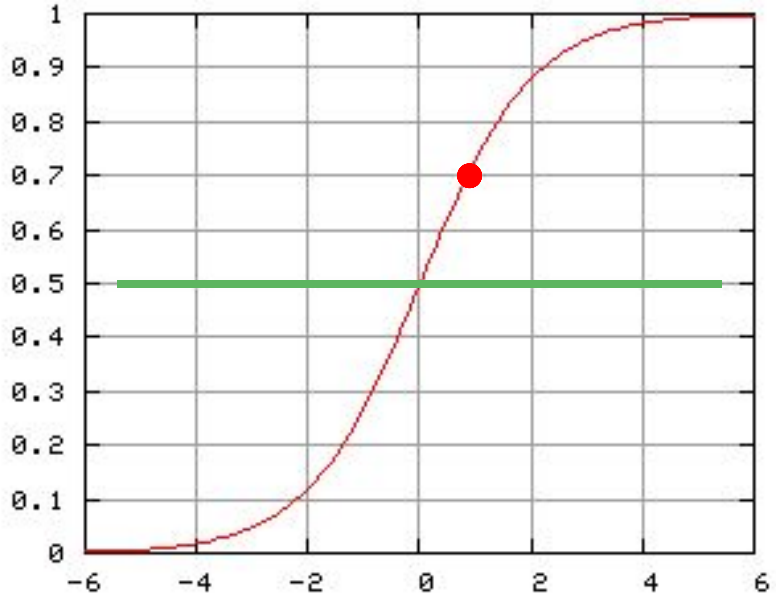
Сигмоида

Предсказание положительного класса



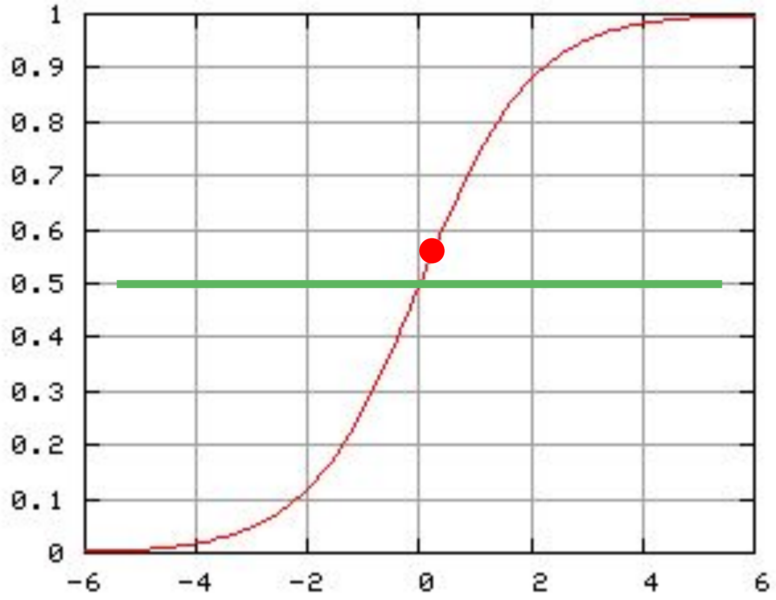
Предсказание отрицательного класса

Сигмоида



Предсказание положительного класса

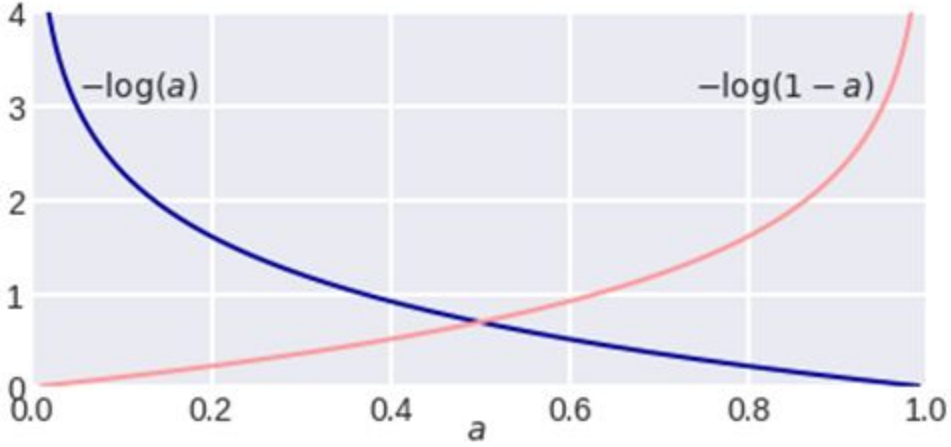
Сигмоида



Предсказание положительного класса с вероятностью чуть больше 0.5

Log-loss

$$-\begin{cases} \log a_i, & y_i = 1, \\ \log(1 - a_i), & y_i = 0. \end{cases}$$



Модель логистической регрессии

Формула логистической регрессии:

$$a(x) = \frac{1}{1 + e^{-(w,x)}}$$

Функция потерь, которую мы минимизируем:

Модель логистической регрессии

Формула логистической регрессии:

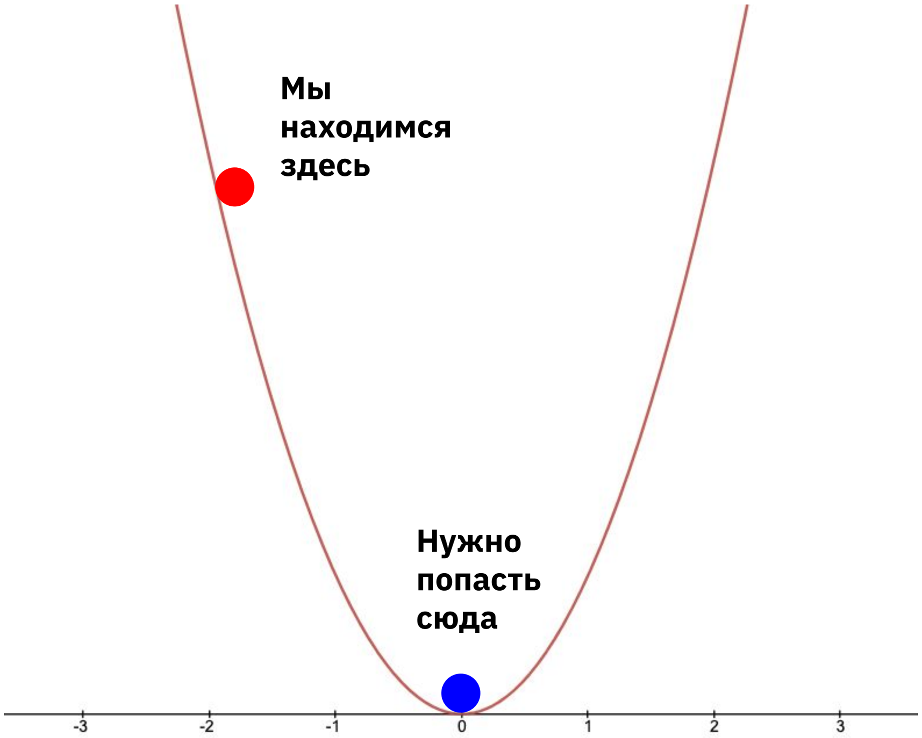
$$a(x) = \frac{1}{1 + e^{-(w,x)}}$$

Функция потерь, которую мы минимизируем:

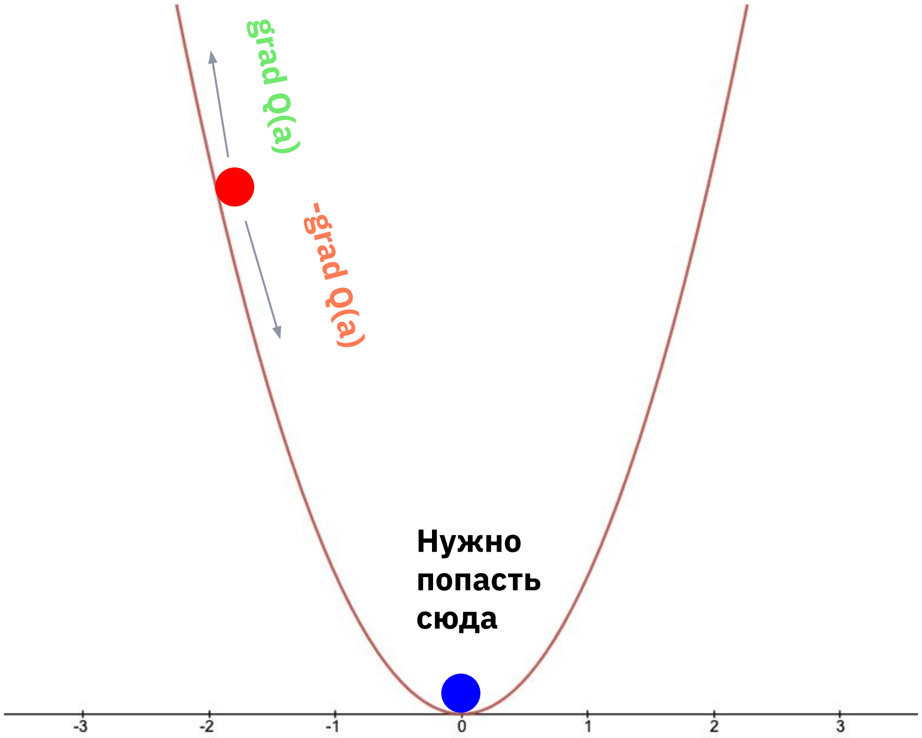
$$H(w) = -\frac{1}{N}\sum_{i=1}^{N} y_i \times \log(p(y_i)) + (1 - y_i) \times \log(1 - p(y_i))$$

Градиентный спуск

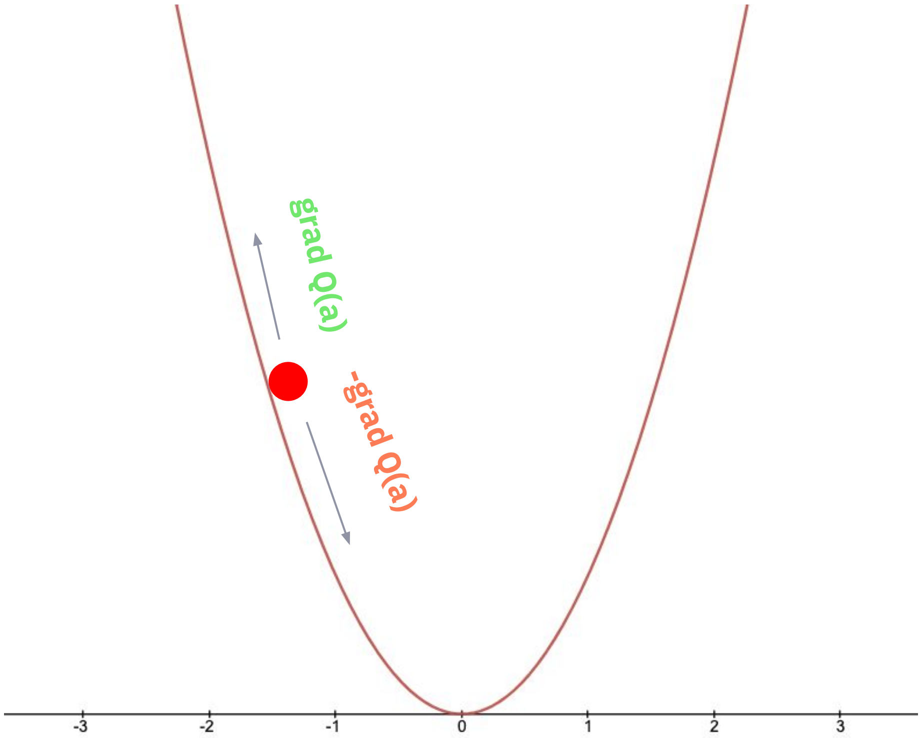
Как найти минимум функции?



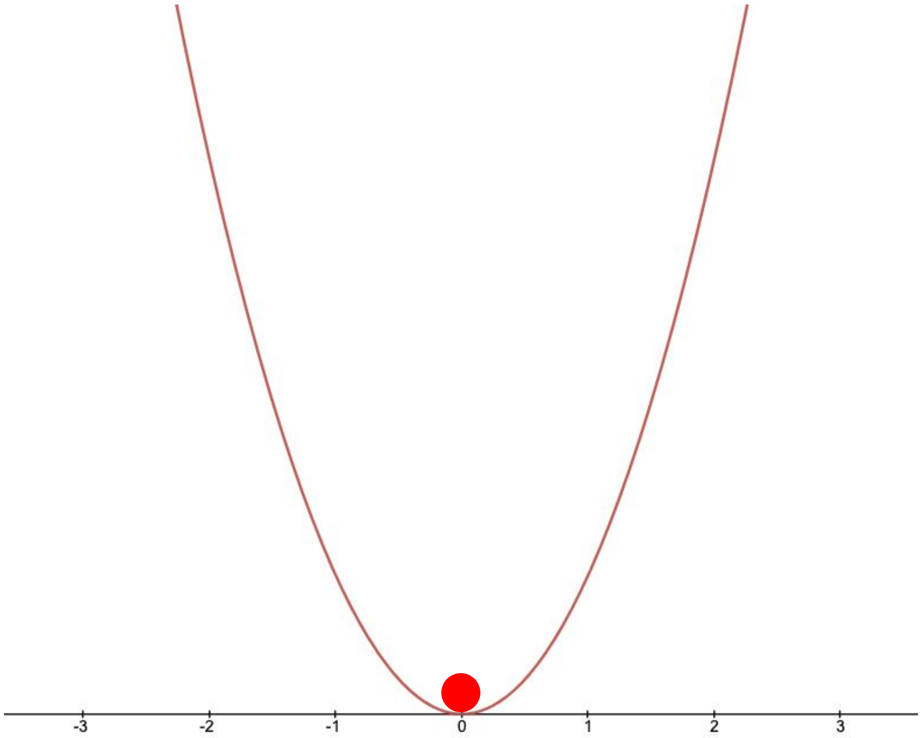
Как найти минимум функции?



Как найти минимум функции?



Как найти минимум функции?



Градиентный спуск

Вычисляем градиент Q(w) в точке и двигаемся в противоположную сторону

$$W_{new} = W_{old} - Q'(W_{old})$$

Градиентный спуск

Добавляем гиперпараметр — шаг градиентного спуска:

$$W_{new} = W_{old} - \alpha \times Q'(W_{old})$$

Градиентный спуск

W - вектор, тогда:

$$W_{new} = W_{old} - \alpha \times \nabla Q\left(W_{old}\right)$$

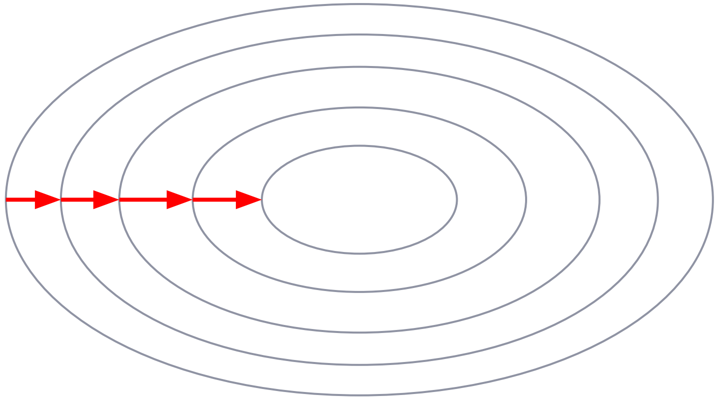
Стохастический градиентный спуск

Считаем градиент не по всем объектам, а выбираем случайный объект.
Формула останется такой же, но под второй частью выражения мы будем считать только один элемент.

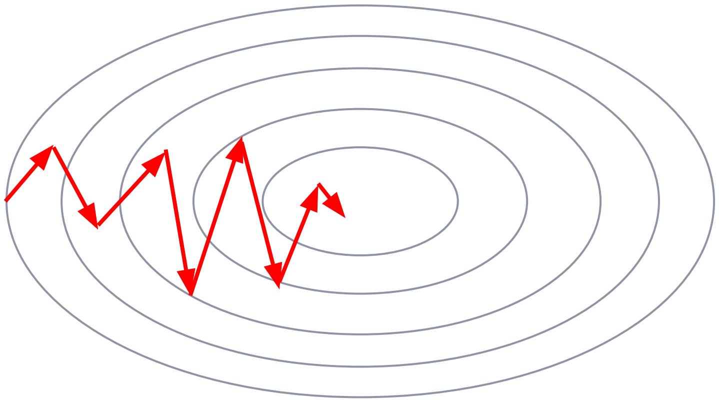
$$W_{new} = W_{old} - \alpha \times \nabla Q\left(W_{old}\right)$$

Сравнение видов градиентных спусков

Градиентный спуск



Стохастический градиентный спуск

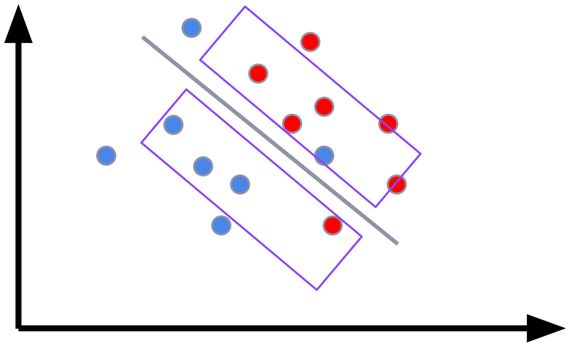


Логистическая регрессия

Логистическая регрессия

Зачастую нам недостаточно знать только определенный ответ, купит ли что-то клиент на сайте или нет. Нам важно понимать, с какой вероятностью он сделает действие, чтобы можно было удержать клиента.

Для этого нужно измерять расстояния от прямой до точек.

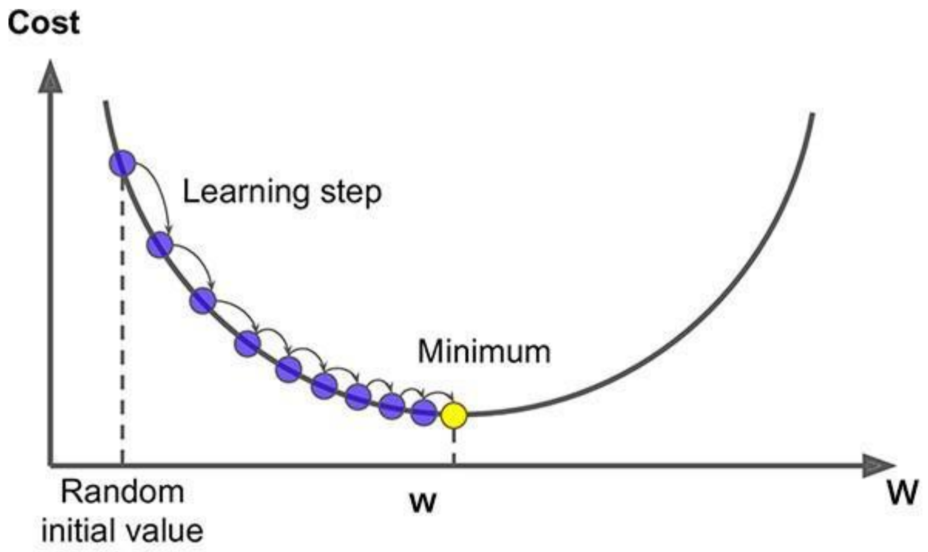


Градиентный спуск

→

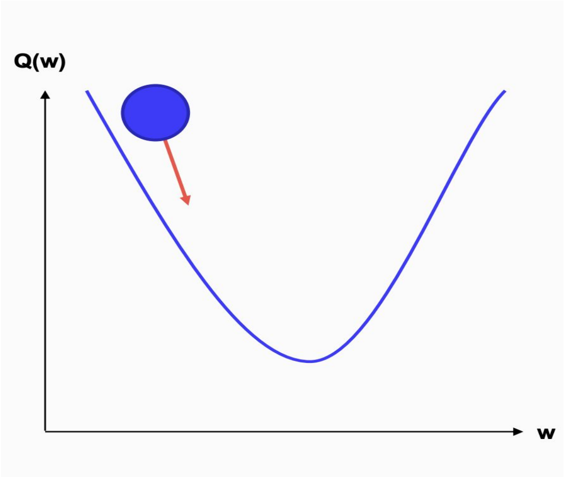
Градиентный спуск

Самая минимальная ошибка любых нашей моделей будет располагаться в самом низу функции, как нам искать ее?



Градиентный спуск

Градиент — вектор, в направлении которого функция быстрее всего растет.



Стохастический градиентный спуск

Модель машинного обучения

Модель машинного обучения — алгоритм, повторяющий мыслительный процесс человека.

Формальное определение: машинное обучение — поиск такой функции(зависимости), которая может искать взаимосвязи в данных и делать новые предположения для новых данных.