ЗАНЯТИЕ 2. РЕГРЕССИЯ

Градиентный спуск

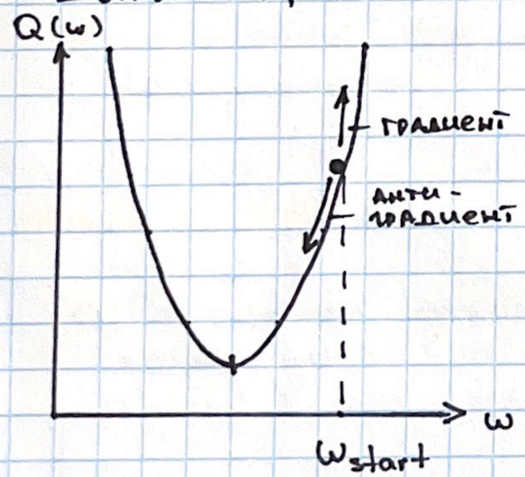
При обучении модели у нас есть функция потерь, которую мы должны минимизировать. ($Q(\overline{w}) \Rightarrow \min\limits_{w}$).
Есть 2 подхода:

1. аналитический: $\left.\begin{aligned} Q'(w_0) &= 0 \\ Q'(w_1) &= 0 \\ &\dots \\ Q'(w_n) &= 0 \end{aligned}\right\}$

Проблема: не каждое уровнение можно решить; не всегда можем найти точное решение

2. численный: градиентный спуск

Допустим, мы хотим минимизировать MSE:



Наша цель - научить компьютер проходить численно в точку мин-ма

Шаг 1: встаём в произвольную точку на функции (w start) и в этой точке посчитаем градиент

Градиент - это вектор, который указывает направление наискорейшего роста функции.

В нашем случае нужен антиградиент (ищем минимум). Получаем формулу:

$$\boxed{w_{new} := w_{old} - \underbrace{\nabla Q(w_{old})}_{\text{градиент}}}$$

Пояснение: мы встаём в точку, считаем градиент и сдвигаемся в сторону антиградиента, т.е. из веса вычитаем значение антиградиента и попадаем в новую точку

Другая запись:

$$\boxed{w^{(k+1)} = w^{(k)} - \nabla Q(w^{(k)})}$$

Проблема: какой шаг делать?

→ нужно добавить параметр η, который будет задавать шаг градиентного спуска (гиперпараметр метода, то число, которое мы задаём сами)

$$\boxed{w^{(k+1)} = w^{(k)} - \eta \nabla Q(w^{(k)})}$$

На практике η [0,001, 0,01] – нормальный градиентный шаг (learning rate)

Стахастический градиентный спуск, SCD

При большом объёме данных обычный градиентный будет слишком долго работать.

Предлагается на 1 итерации брать 1 случайный объект:

$$w^{(k+1)} := w^{(k)} - \eta \nabla_{i_k} Q(w^{(k)})$$

Нормализация / масштабирование / стандартизация признаков

Допустим у нас есть признаки:

X_1 - возраст ~ десятки
X_2 - з/п ~ тысячи
X_3 - женат / не женат ~ 0/1

Так как признаки разных масштабов, желательно приводить их к одному масштабу. Есть разные способы масштабирования, но самый простой:

$$\boxed{\frac{x - \min}{\max - \min}} \in [0; 1]$$

Методы борьбы с переобучением: регуляризация

Если в выборке есть линейно-зависимые признаки, то задача оптимизации $Q(w) \to \min$ имеет бесконечное число решений.

Большие значения параметров (весов) модели w - признак переобучения.

Решение проблемы - регуляризация.
Будем минимизировать регуляризованный функционал ошибки:

$$\boxed{Q_{alpha}(w) = Q(w) + \alpha \cdot R(w) \to \min_w}$$, где

$R(w)$ - регуляризатор

Регуляризация штрафует слишком большие веса.

Наиболее используемые регуляризаторы:

- L_2 - регуляризатор : $R(w) = ||w||_2 = \sum_{i=1}^{d} w_i^2$
- L_1 - регуляризатор: $R(w) = ||w||_1 = \sum_{i=1}^{d} |w_i|$

Пример регуляризованного функционала:

$$Q(a(w), X) = \frac{1}{l} \sum_{i=1}^{l} ((w, x_i) - y_i)^2 + \alpha \sum_{i=1}^{d} w_i^2,$$

где α - коэффициент регуляризации