Линейная регрессия. Градиентный спуск

2021

Мария Корлякова

ПЛАН

1.Линейная регрессия. MSE

2.Метод наименьших квадратов

3.Градиентный спуск.

Машинное обучение

Объект:

Описание объекта:

- пол: female male
- окрас: sable sable
- рост: 45 см. 50 см

Ответ: Колли / Шарпей

Типы признаков

Бинарные: (самец/самка)

Вещественные : рост, вес

Порядковые: число ног

Категорийные: цвет

Множественные: подмножество из множества

Обучение

Без учителя (выделение классов)

С учителем (отнесение к классу)

Обучение с подкреплением

Supervised learning

Unsupervised learning

Reinforcement learning

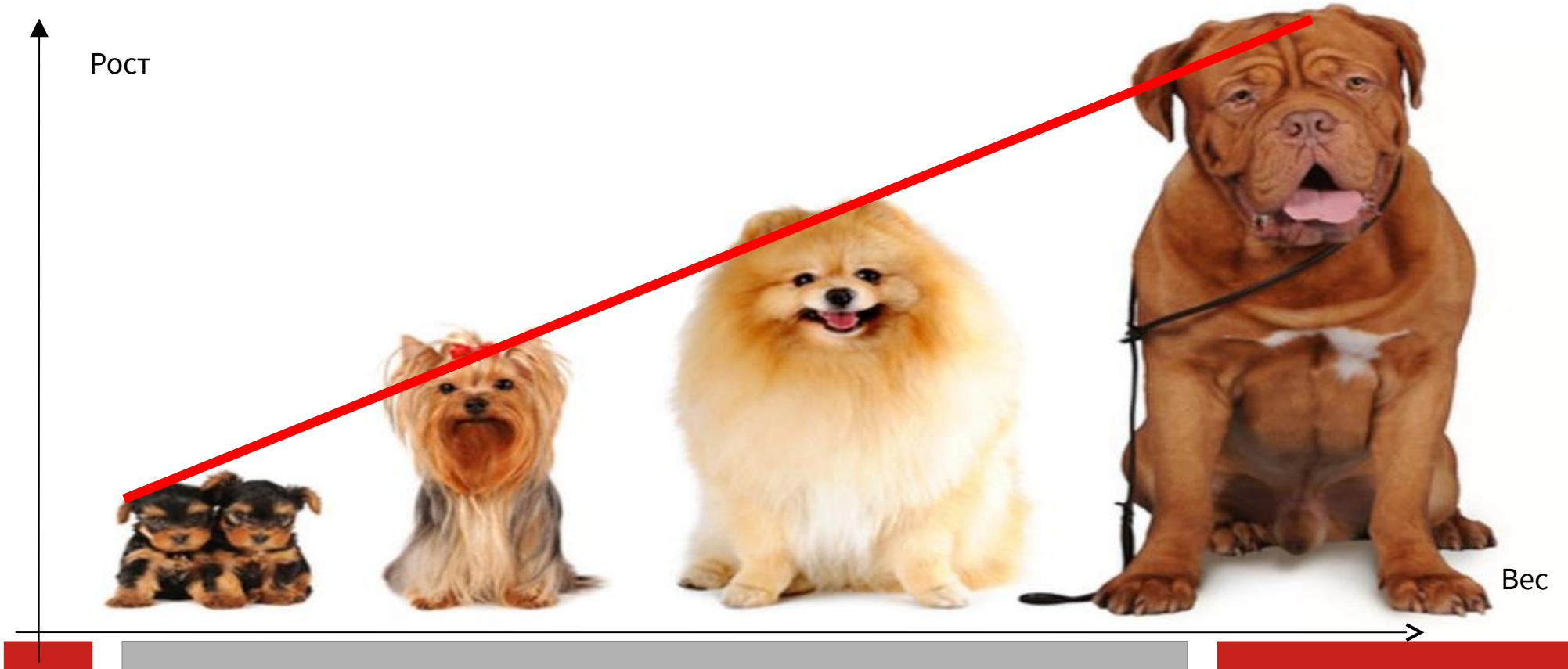
Задача классификации

Разделить объекты на группы и сказать, к какой из них относиться новый объект:

Dog

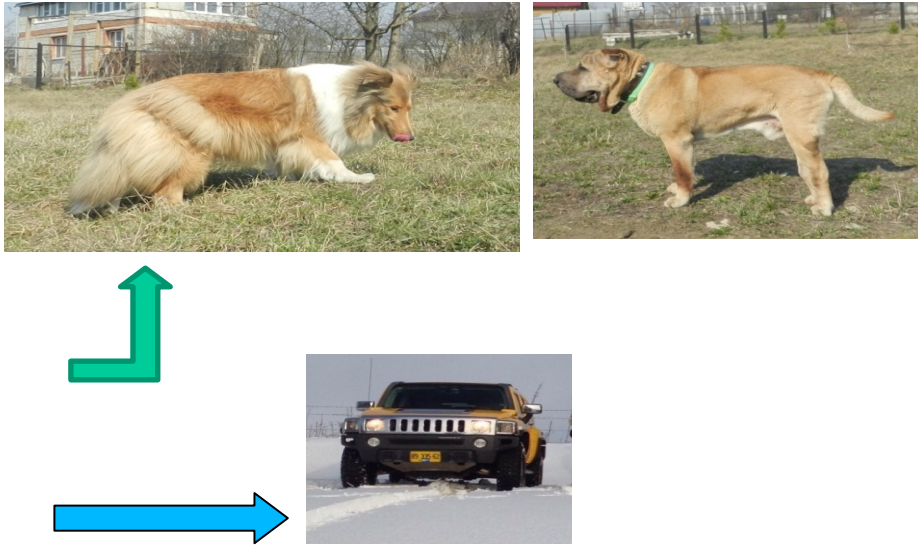
Car

Задача Регрессии



Задача Кластеризации

Необходимо определить группы, которые сформированы на основании метрики близости.



Постановка задачи

Задача обучения с учителем

$$X = (x^i, y_i)_{i=1}^l$$

$$a(x) = w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_d x^d$$

$$a(x) = sign(w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_d x^d)$$

$$a(x) :$$

$$Q(a, X) \to min$$

Линейная регрессия

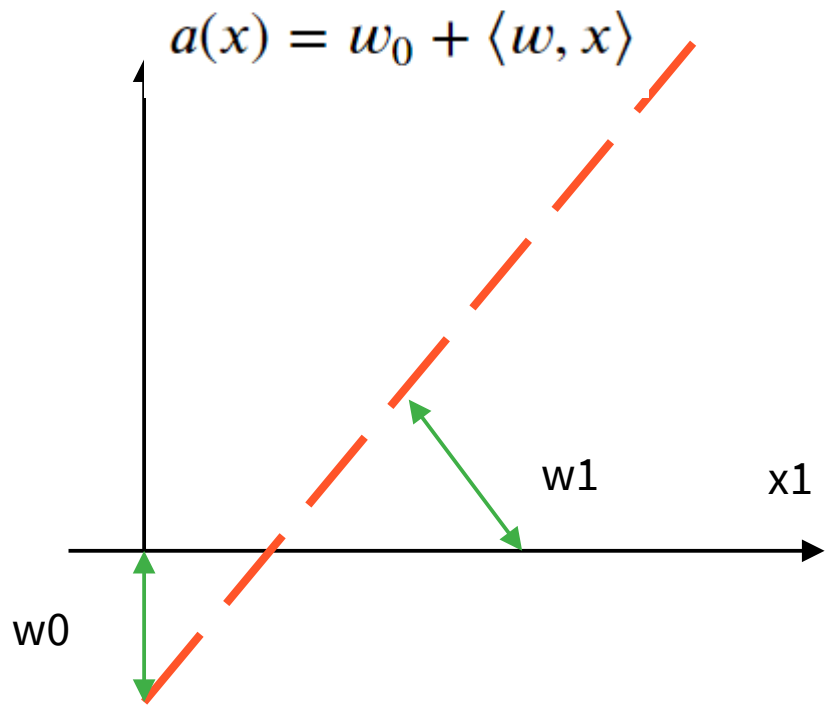
Уравнение линейной регрессии

$$a(x) = w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_d x^d$$

$$a(x) = w_0 + \sum_{i=1}^{d} w_i x^i.$$

$$a(x) = w_0 + \langle w, x \rangle$$

$$a(x) = \langle w, x \rangle \qquad x0 = 1$$



Качество решения задачи линейной регрессии

Средняя Абсолютная Ошибка

Mean absolut error (MAE)

$$Q(a, x) = \frac{1}{l} \sum_{i=1}^{l} |a(x_i) - y_i|$$

Средняя Квадратичнная Ошибка

Mean Square Error (MSE)

$$Q(a, x) = \frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2$$

Метод наименьших квадратов

Критерий

$$Q(w, x) = \frac{1}{l}\sum_{i=1}^{l}(\langle w, x_i\rangle - y_i)^2 \to \min_w.$$

$$X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{l1} & \dots & x_{ld} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix}$$

$$Q(w, X) = \frac{1}{l}||Xw - y||^2 \to \min_w$$

$$Xw = y$$

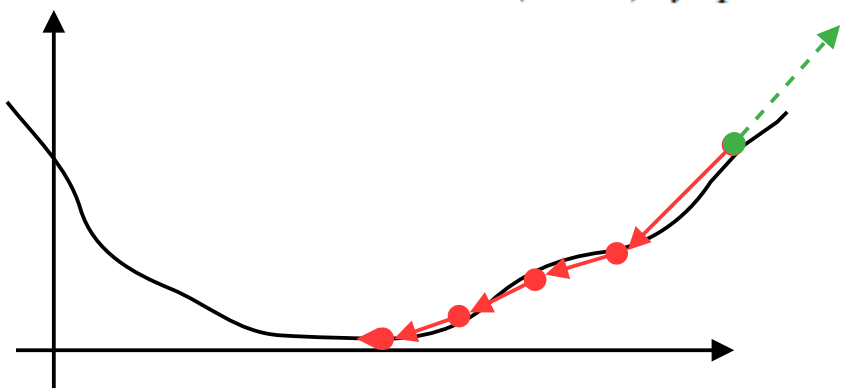
$$w = (X^T X)^{-1} X^T y.$$

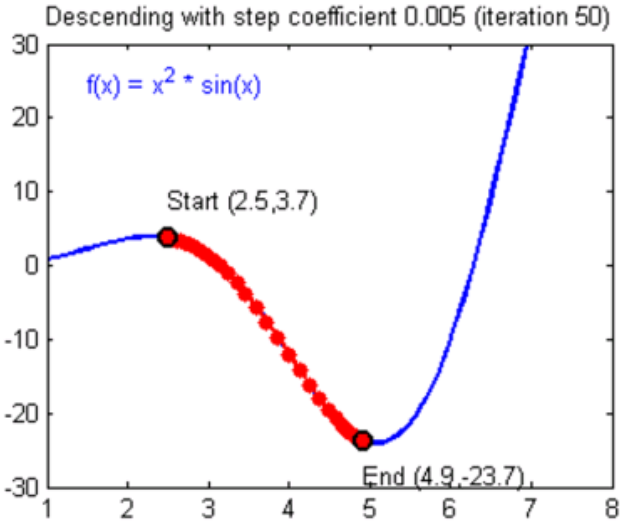
https://habr.com/ru/company/ods/blog/323890/#metod-naimenshih-kvadratov

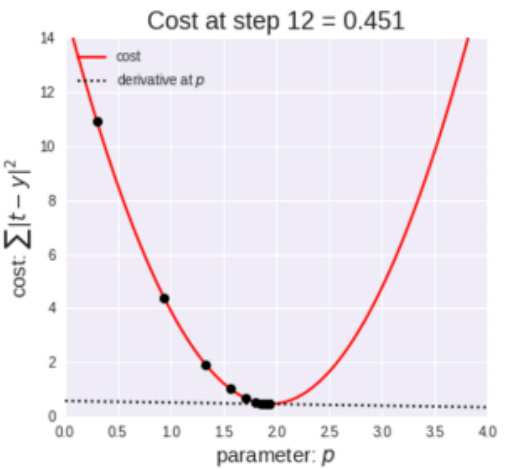
Градиентный спуск

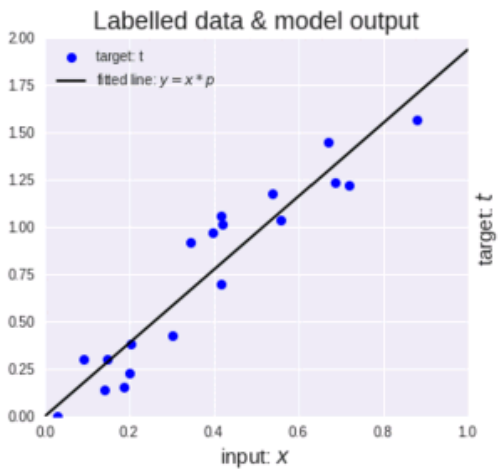
Минимум функции

$$\nabla f(x_1, \ldots, x_d) = \left(\frac{\partial f}{\partial x_i} \right)_{i=1}^{d}$$









Градиентный спуск

Критерий

$$Q(w, X) = \frac{1}{l}||Xw - y||^2 \to \min_w$$

$$X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ x_{l1} & \dots & x_{ld} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix}$$

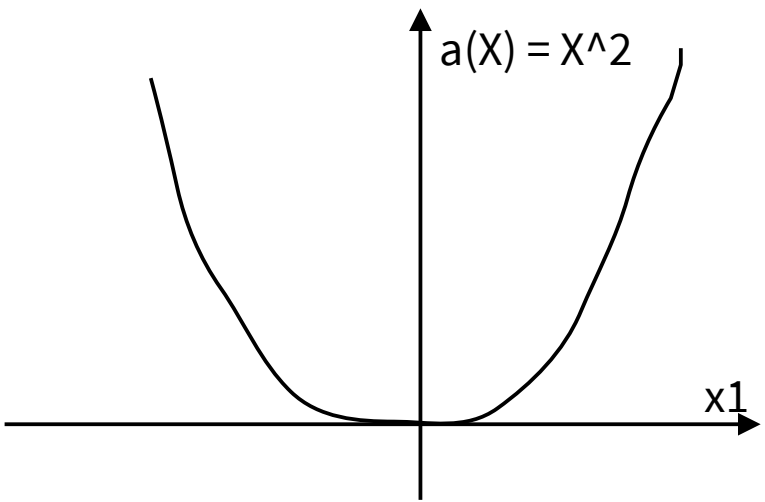
$$w^k = w^{k-1} - \eta_k \nabla Q(w^{k-1}, X).$$

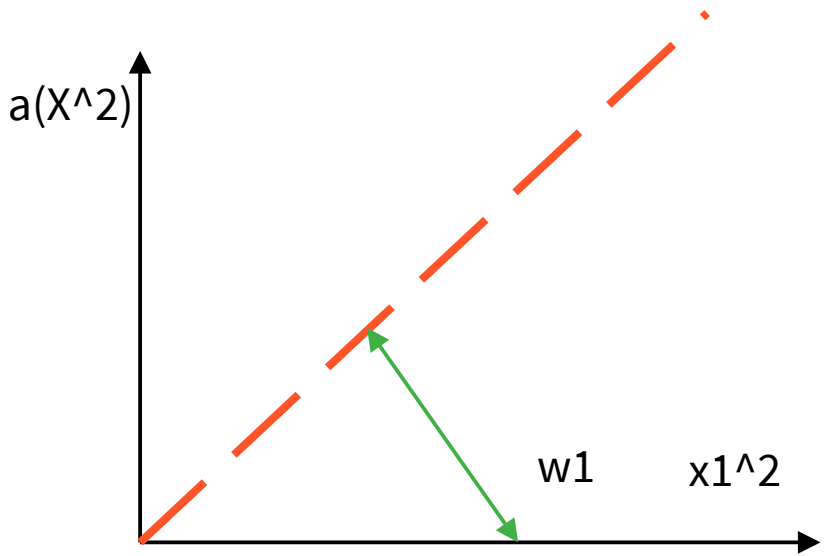
$$\nabla_w Q(w, X) = \frac{2}{l} X^T (Xw - y).$$

Линейная регрессия

Уравнение линейной регрессии

$$a(x) = \langle w, x \rangle$$





Домашнее задание:

Задача: предсказание баллов ЕГЭ ученика в зависимости от кол-ва лет стажа его репетитора (Lesson_1_HW)

1. Подберите скорость обучения (alpha) и количество итераций

*2. В коде п.2 есть ошибка, исправьте ее

*3. Вместо того, чтобы задавать количество итераций, задайте условие остановки алгоритма - когда ошибка за итерацию начинает изменяться ниже определенного порога (упрощенный аналог параметра tol в линейной регрессии в sklearn).

4. Сделайте выводы по результатам работы с GD : что повышает качество результата, что понижает.