Машинное обучение: часть 3

Линейная регрессия

Пусть x — рост котика, а y — его вес.

Что мы знаем?

- чем крупнее котик, тем больший вес он имеет;
- котики одинакового роста могут иметь разный вес.

Выводы:

- для фиксированного роста котика x

  его вес $y = f(x)$ является случайной величиной;
- в среднем вес $f(x)$ возрастает при увеличении роста котика x.

Простая зависимость:

$$y = \theta_0 + \theta_1 x + \varepsilon,$$

x — рост котика,

y — вес котика,

θ_0, θ_1 — неизвестные параметры,

ε — случайная составляющая
с нулевым средним.

Более сложная зависимость:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 + \varepsilon,$$

x_1 — рост котика,

x_2 — обхват туловища котика,

y — вес котика,

$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры,

ε — случайная составляющая
с нулевым средним.

Простая зависимость:

$$y = \theta_0 + \theta_1 x + \varepsilon,$$

x — рост котика,

y — вес котика,

θ_0, θ_1 — неизвестные параметры,

ε — случайная составляющая с нулевым средним.

Зависимость

- **линейна по параметрам**,
- линейна по аргументу.

Более сложная зависимость:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 + \varepsilon,$$

x_1 — рост котика,

x_2 — обхват туловища котика,

y — вес котика,

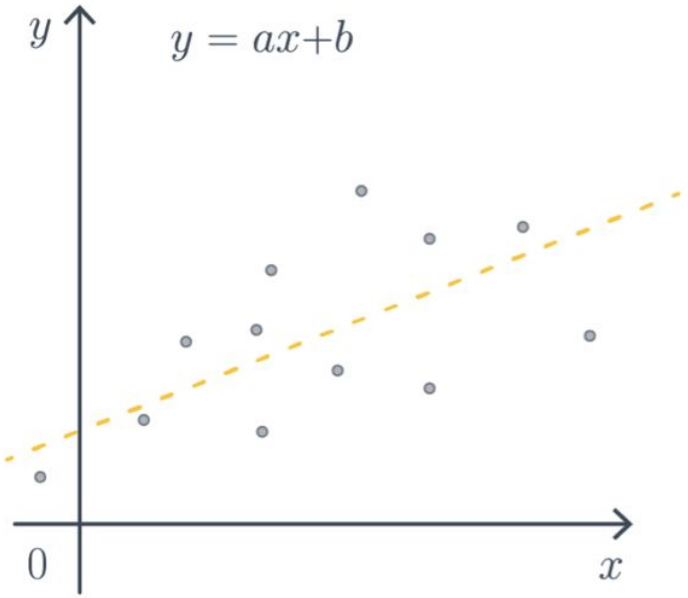
$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры,

ε — случайная составляющая с нулевым средним.

Зависимость

- **линейна по параметрам**,
- квадратична по аргументам.

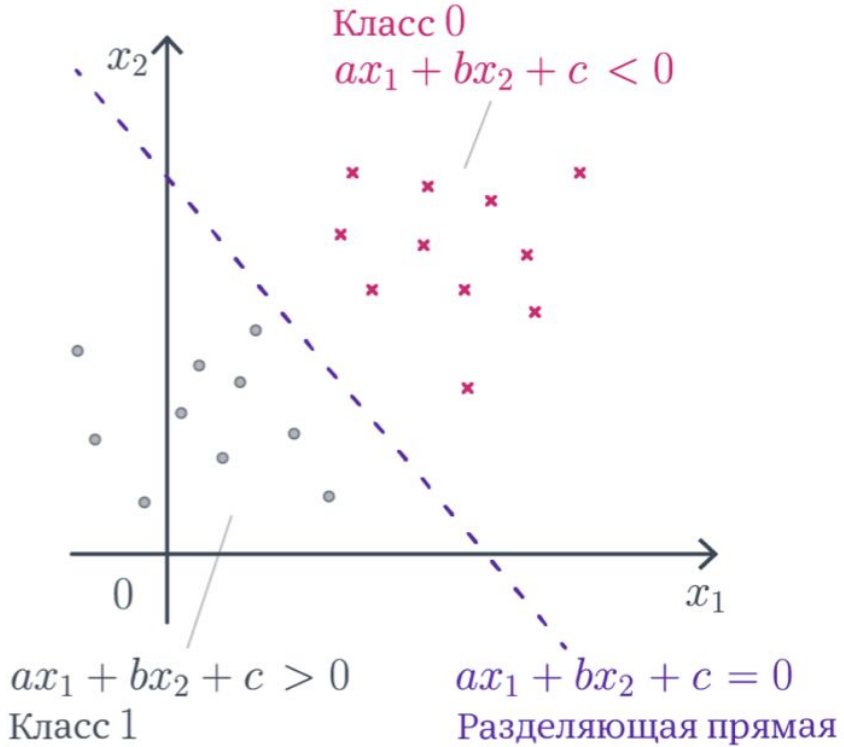
Регрессия



x — (единственный) признак

y — таргет

Классификация



x_1, x_2 — признаки

Обучение линейной регрессии

	Признак 1	Признак 2	Признак 3	Признак 4	Признак 5		
	Площадь	Удаленность от центра	Год постройки	Лет после ремонта	Тип постройки	Стоимость квартиры > 100k $	Ответ модели линейной регрессии
1	25	3	2005	1	1	65	78
2	55	10	1987	5	2	120	115
3	50	12	1990	6	5	125	105
	k_1	k_2	k_3	k_4	k_5	y	$\hat{y}$

$$\hat{y} = k_0 + \sum x_i k_i = k_0 + k_1 x_1 + k_2 x_2 + k_3 x_3 + k_4 x_4 + k_5 x_5$$

Возьмем средний квадрат ошибки (квадрат разницы между истинным ответов и предсказанием) и будем минимизировать его

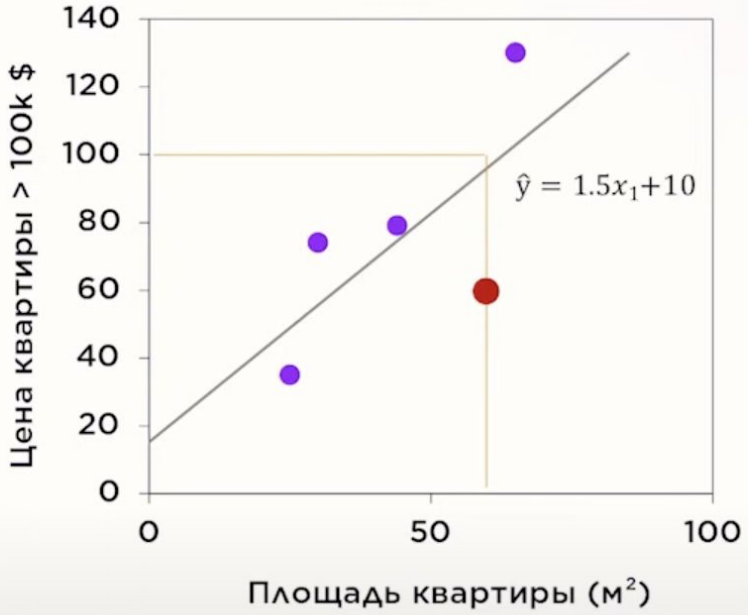
Такая метрика называется MSE (Mean Squared Error)

А мы будем искать значения весов, которые минимизируют MSE

Геометрический смысл MSE

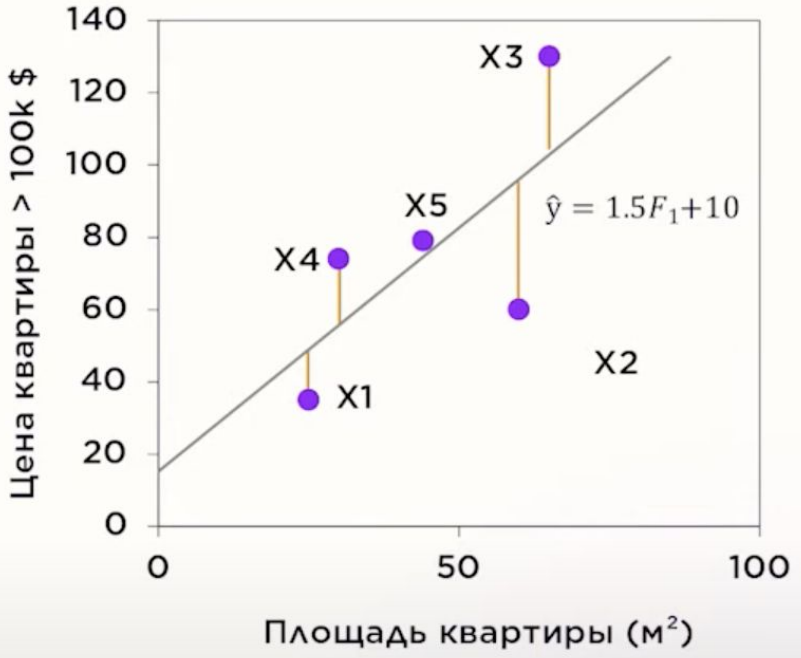
Пусть у нас только один признак – площадь

	Признак 1		
	Площадь	Стоимость квартиры > 100k $	Ответ линейной регрессии
1	25	35	47.5
2	60	60	100
3	65	130	107.5
4	30	74	55
5	44	79	76

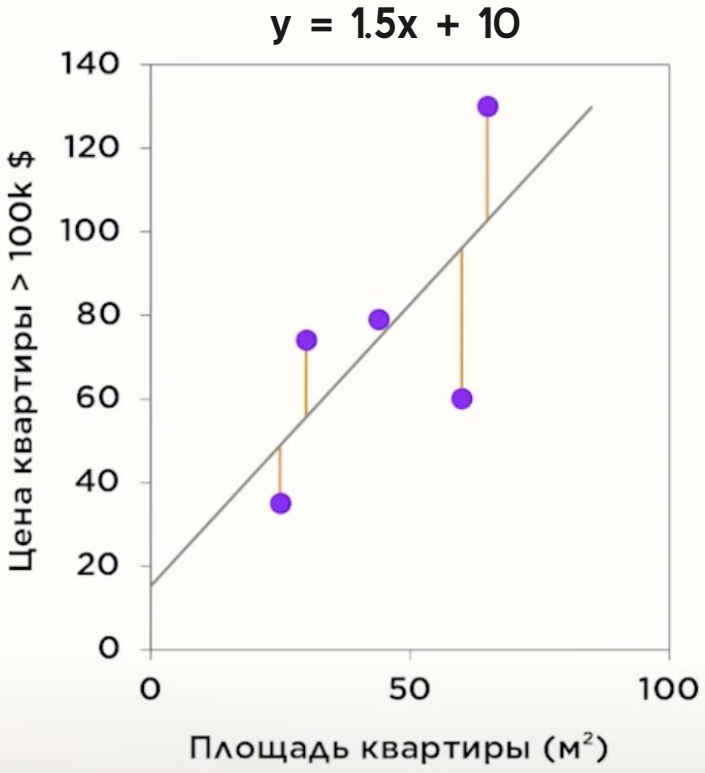


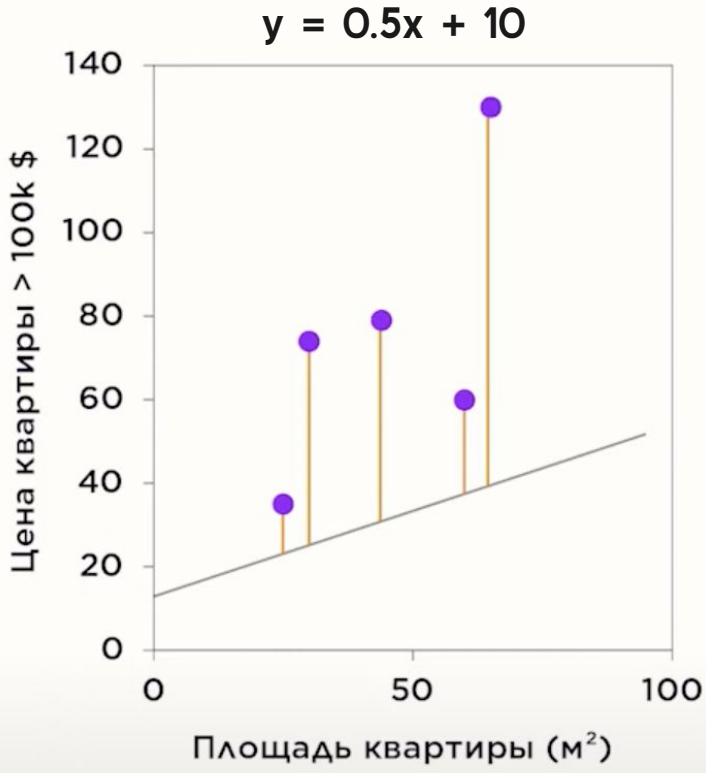
Это расстояние между

	Признак 1		
	Площадь	Стоимость квартиры > 100k $	Ответ линейной регрессии
1	25	35	47.5
2	60	60	100
3	65	130	107.5
4	30	74	55
5	44	79	76



Пусть теперь у нас есть другая модель





Регрессия плохо работает, когда зависимость между признаками и ответами – нелинейная

	Признак 1	Ответ
1	25	35
2	40	15
3	50	30
4	60	60
5	55	80
6	35	100
7	25	75