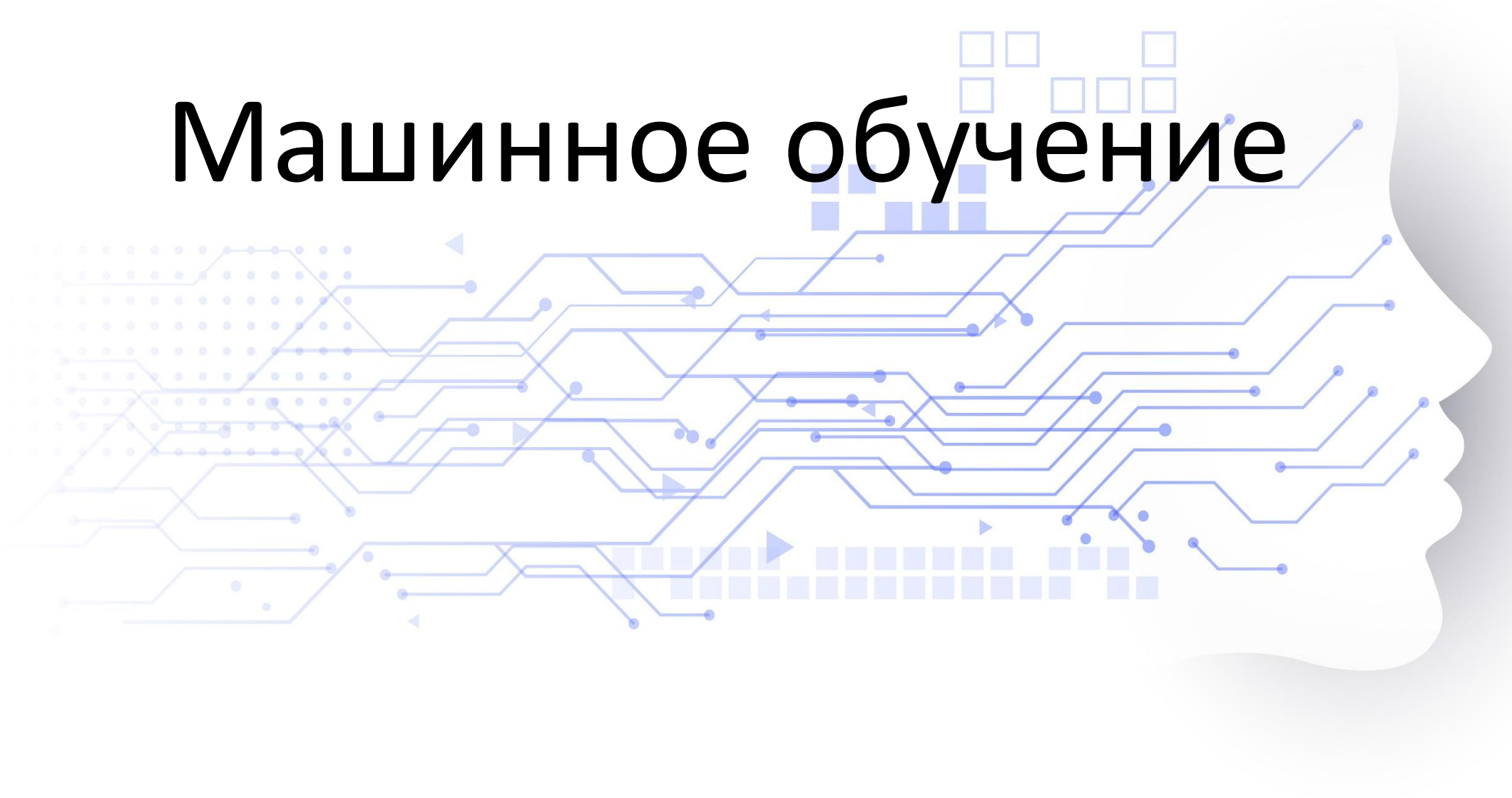
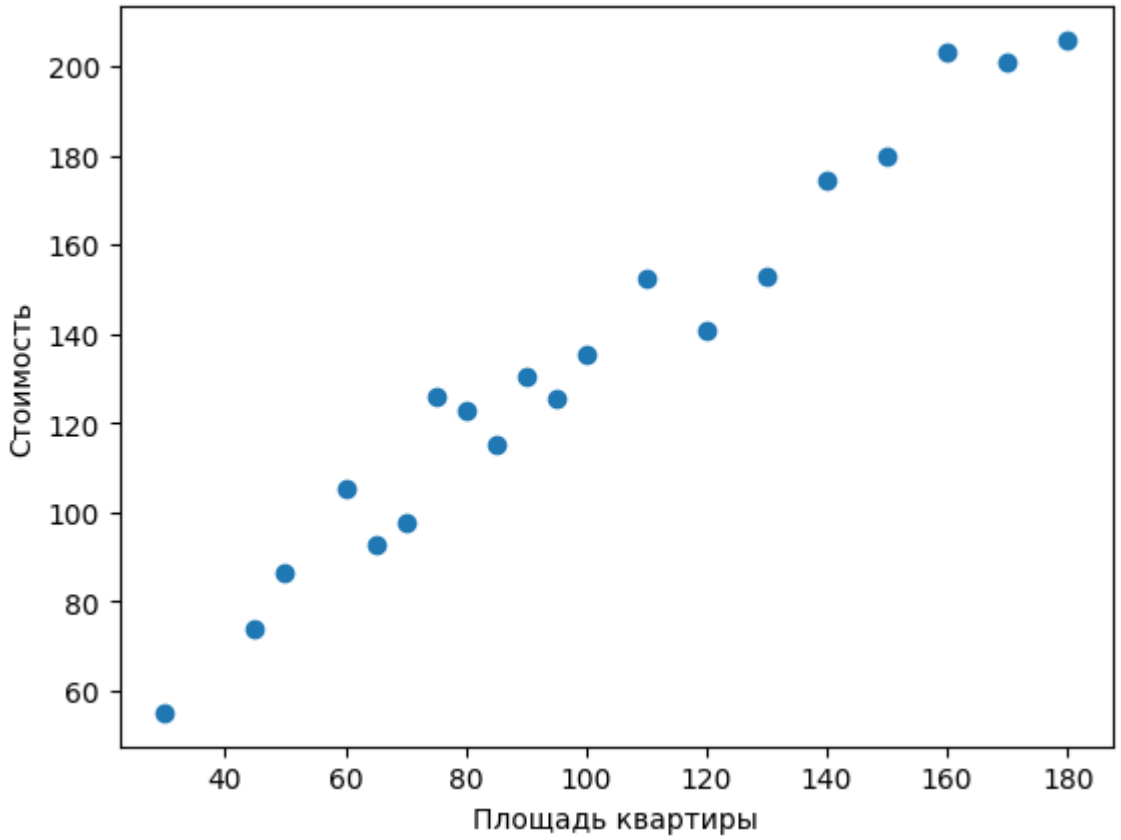
Машинное обучение

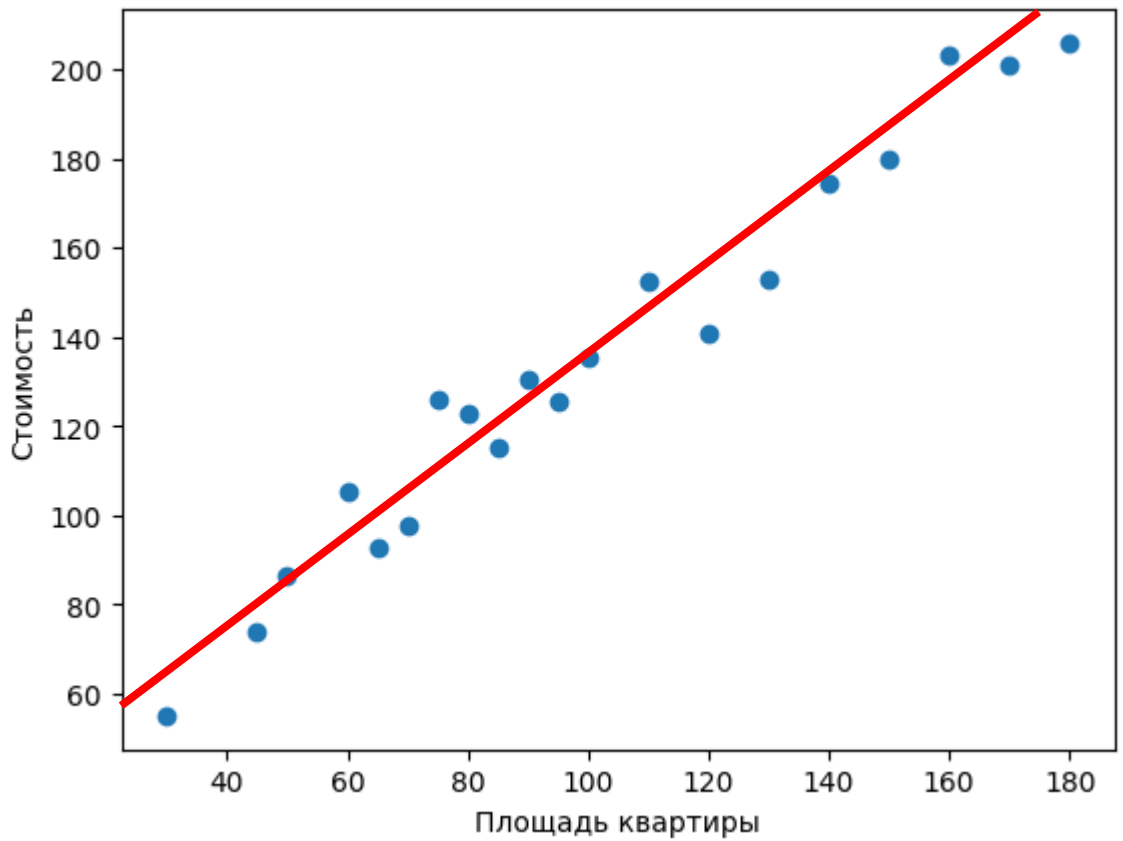


Линейная регрессия

Модель линейной регрессии

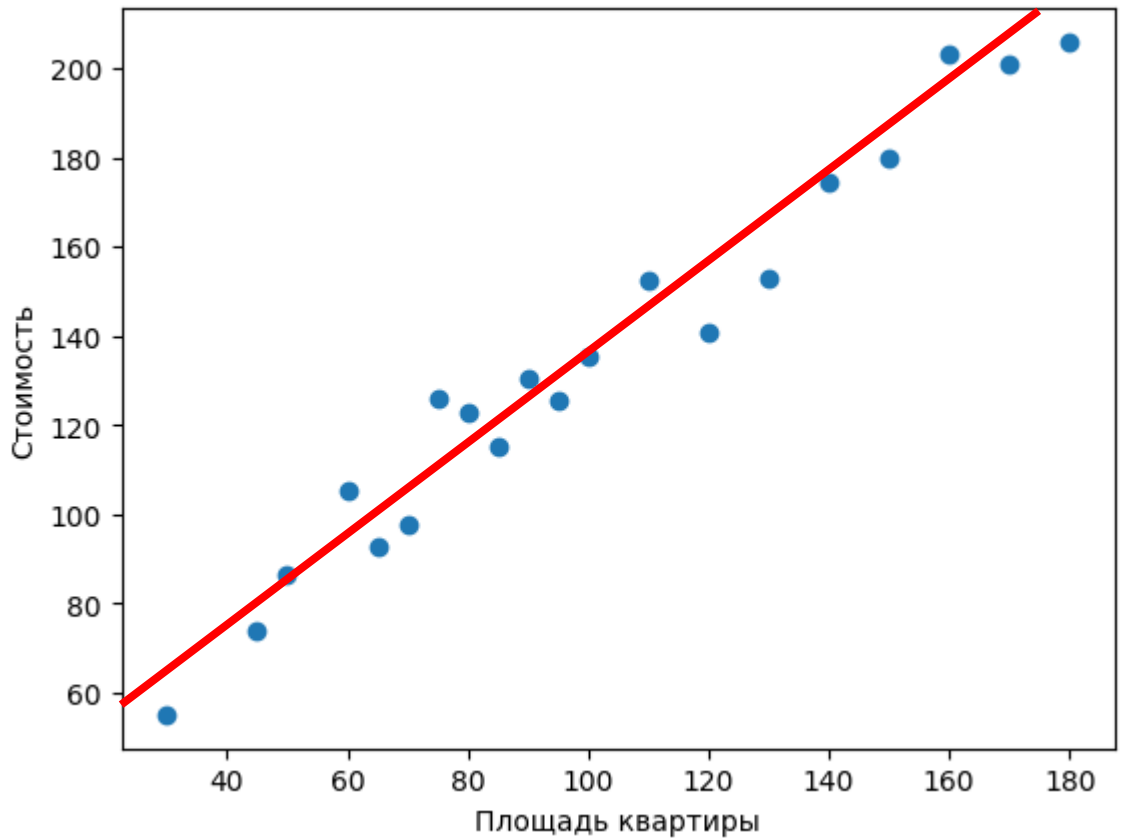


Модель линейной регрессии



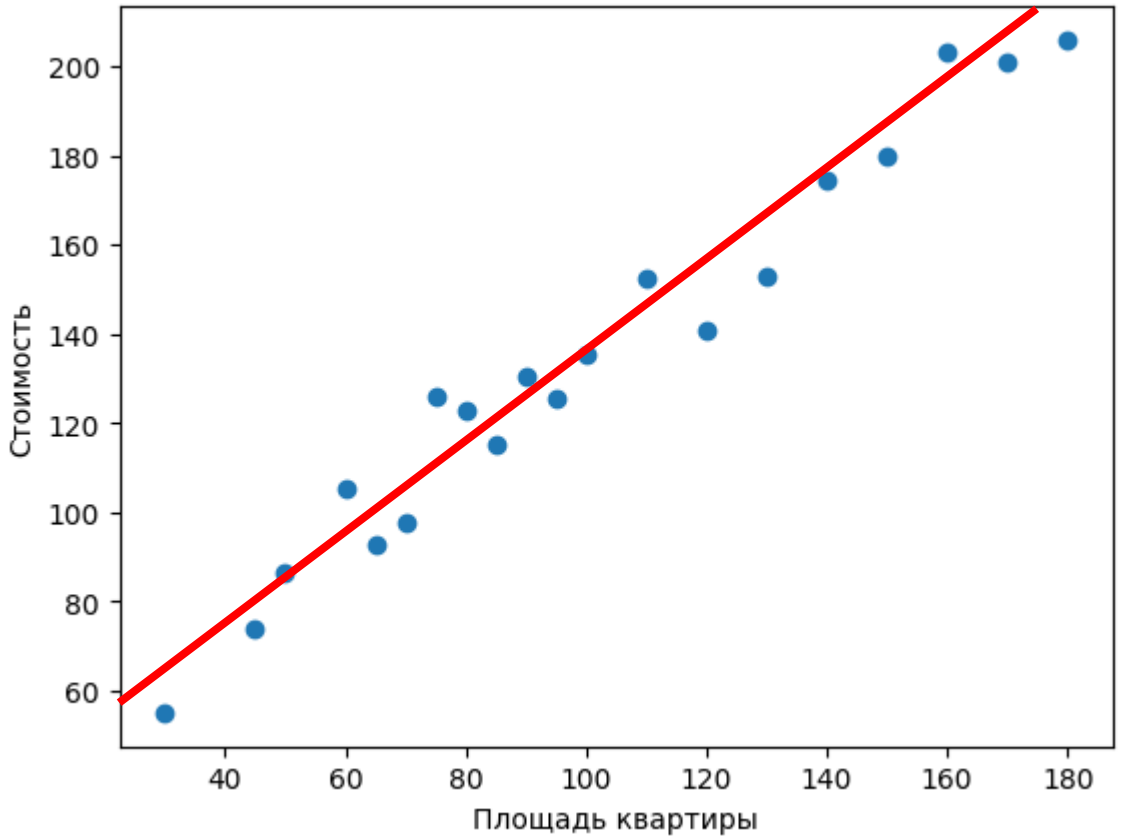
Модель линейной регрессии

$$y = ax + b$$



Модель линейной регрессии

$$y = ax + b$$



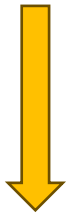
a — тангенс угла наклона

b — где прямая пересекает ось ординат

Парная регрессия

Регрессия с одной переменной

$$y = ax + b$$



Модель: $$a(x) = \theta_1 x_1 + \theta_0$$

$$a(x) = \theta_1 x_{\text{площадь квартиры}} + \theta_0$$

Количество параметров: Два

θ_1 – тангенс угла наклона

θ_0 – где прямая пересекает ось ординат

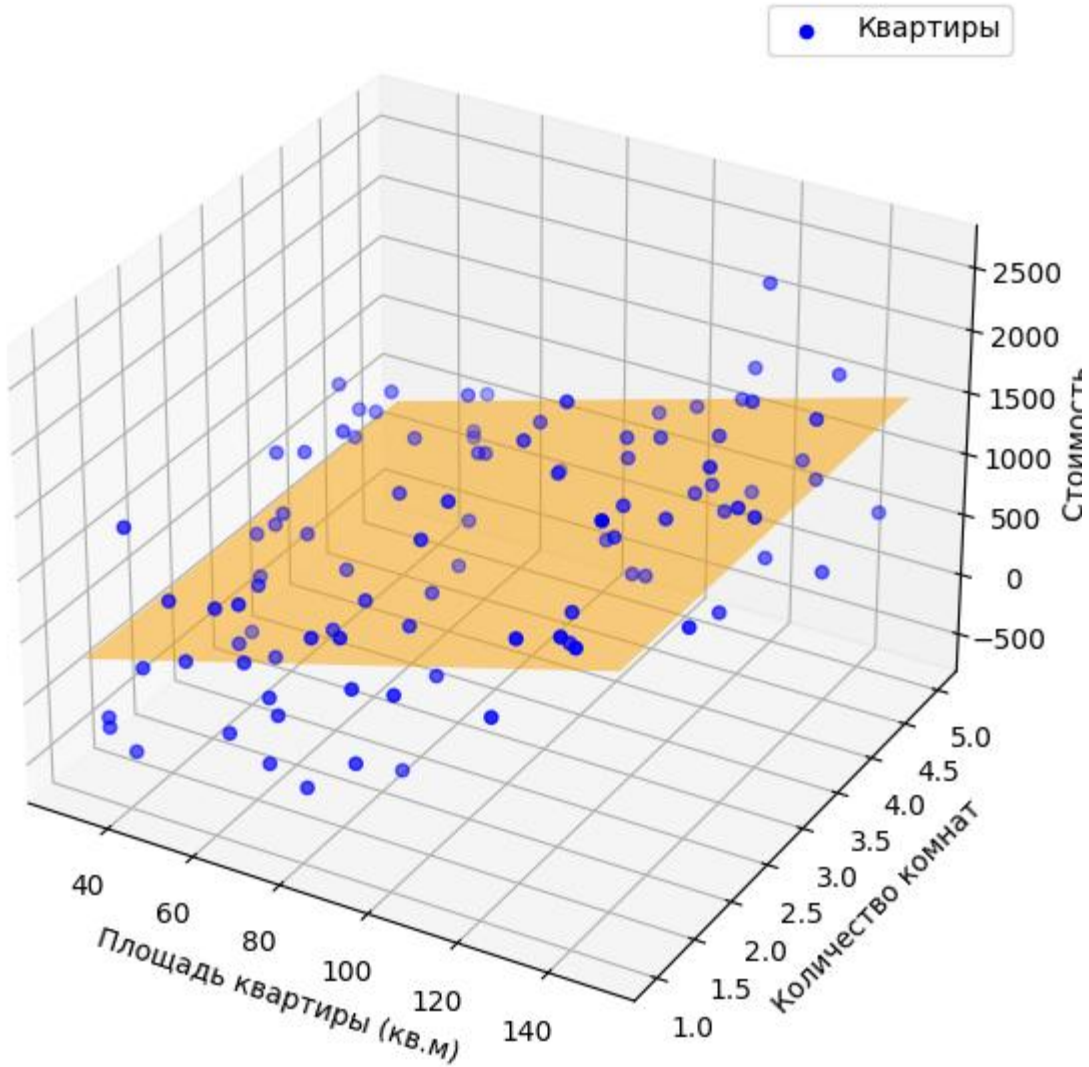
Линейная регрессия с двумя признаками

Модель: $$\boldsymbol{a(x) = \theta_1 x_1 + \theta_2 x_2 + \theta_0}$$

$$\boldsymbol{a(x) = \theta_1 x_{\text{площадь квартиры}} + \theta_2 x_{\text{количество комнат}} + \theta_0}$$

Количество параметров: Три

Задачи регрессии



Линейная регрессия с несколькими переменными

Площадь квартиры (x_1)	Этаж квартиры (x_2)	Площадь кухни (x_3)	Количество комнат (x_4)	Стоимость квартиры ($\mathbb{Y}$)
460	2	15	6	195
230	7	9	4	130
315	1	20	3	140
178	3	25	4	80
…	…	…	…	….

$$\boldsymbol{a(x)} = \boldsymbol{\theta_0} + \boldsymbol{\theta_1 x}_{\text{площадь квартиры}} + \boldsymbol{\theta_2 x}_{\text{этаж квартиры}} + \boldsymbol{\theta_3 x}_{\text{площадь кухни}} + \boldsymbol{\theta_4 x}_{\text{количество комнат}}$$

Регрессия с несколькими переменными

Количество признаков: d признаков

Модель: $$\boldsymbol{a(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d}$$

Количество параметров: $d + 1$

Регрессия с несколькими переменными

Количество признаков: d признаков

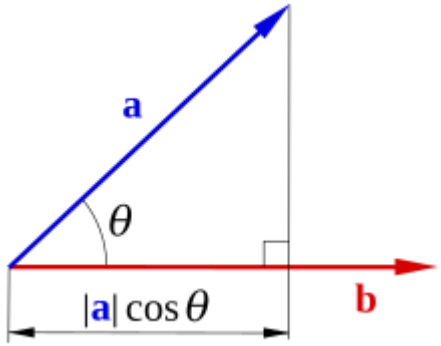
Модель: $\boldsymbol{a(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d}$

Свободный коэффициент / bias

Веса/коэффициенты

Количество параметров: $d + 1$

Скалярное произведение



$$\langle a.b\rangle = \sum_{i=1}^{n} a_i b_i$$

a- первый вектор
b- второй вектор
n- размерность векторного пространства
a_i- компонент вектора a
b_i- компонент вектора b

Линейная регрессия в векторном виде

Площадь квартиры (x_1)	Этаж квартиры (x_2)	… (x_d)	Стоимость квартиры $(\mathbb{Y})$
460	2	…	195
230	7	…	130
315	1	…	140
178	3	…	80
…	…	…	….

$$\boldsymbol{a(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d}$$

Линейная регрессия в векторном виде

$$a(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d$$

$$a(x) = \theta_0 + \sum_{i=1}^{d} \theta_i x_i$$

$$a(x) = \theta_0 + \langle \theta, x \rangle$$

Линейная регрессия в векторном виде

$$a(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d$$

$$a(x) = \theta_0 + \sum_{i=1}^{d} \theta_i x_i$$

$$a(x) = \theta_0 + \langle \theta, x \rangle$$

Линейная регрессия в векторном виде

Bias (x_0)	Площадь квартиры (x_1)	Этаж квартиры (x_2)	… (x_d)	Стоимость квартиры $(\mathbb{Y})$
1	460	2	…	195
1	230	7	…	130
1	315	1	…	140
1	178	3	…	80
…	…	…	…	….

Линейная регрессия в векторном виде

$$a(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d = \theta_0 + \langle \theta, x \rangle$$

Есть признак, всегда равный единице $x_0 = 1$

$$a(x) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_d x_d$$

$$a(x) = \theta_0 * 1 + \theta_1 x_1 + \ldots + \theta_d x_d$$

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

Модель линейной регрессии

$$a(x) = \theta_1 x_1 + ... + \theta_d x_d$$

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

Нет гарантий, что целевая переменная именно так зависит от признаков

Надо формировать признаки так, чтобы модель подходила

Предсказание стоимости квартиры

Признаки: площадь, район, расстояние до метро

Целевая переменная: рыночная стоимость квартиры

Площадь (x_1)	Район(x_2)	Расстояние до метро (x_3)	Стоимость квартиры $(\mathbb{Y})$
460	ЦАО	0.2	195
230	ЮАО	2	130
315	ЦАО	1.2	140
178	САО	5	80
87	ЮАО	0.8	98

Линейная модель:

$$\boldsymbol{a(x)} = \boldsymbol{\theta_0} + \boldsymbol{\theta_1}\, \boldsymbol{x}_{(\text{площадь})} + \boldsymbol{\theta_2}\, \boldsymbol{x}_{(\text{район})} + \boldsymbol{\theta_3 x}_{(\text{расстояние до метро})}$$

Предсказание стоимости квартиры

$$a(x) = \theta_0 + \theta_1 x_{(\text{площадь})} + \theta_2 x_{(\text{район})} + \theta_3 x_{(\text{расстояние до метро})}$$

Площадь (x_1)	Район(x_2)	Расстояние до метро (x_3)	Стоимость квартиры $(\mathbb{Y})$
460	ЦАО	0.2	195
230	ЮАО	2	130
315	ЦАО	1.2	140
178	САО	5	80
87	ЮАО	0.8	98

За каждый квадратный метр добавляем θ_1 к прогнозу

Предсказание стоимости квартиры

$$a(x) = \theta_0 + \theta_1 x_{(\text{площадь})} + \theta_2 x_{(\text{район})} + \theta_3 x_{(\text{расстояние до метро})}$$

Площадь (x_1)	Район(x_2)	Расстояние до метро (x_3)	Стоимость квартиры $(\mathbb{Y})$
460	ЦАО	0.2	195
230	ЮАО	2	130
315	ЦАО	1.2	140
178	САО	5	80
87	ЮАО	0.8	98

Кодирование категориальных признаков

Значения признака «район»: $U = \{u_1, \dots, u_m\}$

Новые признаки вместо x_j: $[x_j = u_1], \dots, [x_j = u_m]$

One-hot кодирование

Кодирование категориальных признаков

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

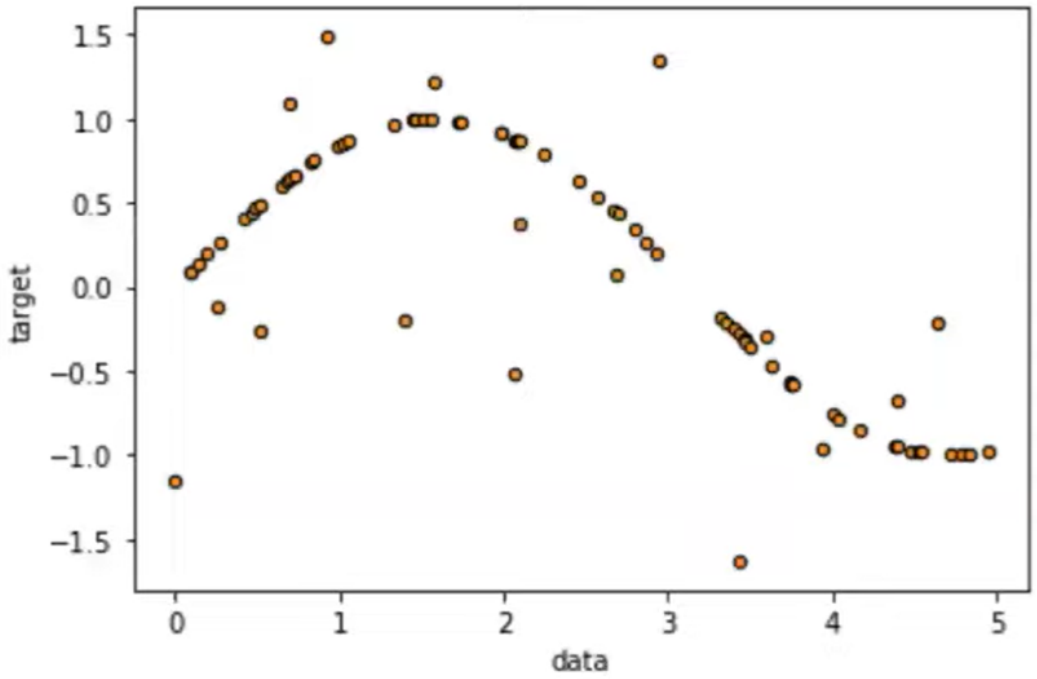
Кодирование категориальных признаков

Площадь (x_1)	ЦАО (x_2)	ЮАО (x_3)	САО (x_4)	Расстояние до метро (x_5)	Стоимость квартиры $(\mathbb{Y})$
460	1	0	0	0.2	195
230	0	1	0	2	130
315	1	0	0	1.2	140
178	0	0	1	5	80
87	0	1	0	0.8	98

$$a(x) = \theta_0 + \theta_1 x_{(\text{площадь})} + \theta_2 x_{(\text{ЦАО})} + \theta_3 x_{(\text{ЮАО})} + \theta_4 x_{(\text{САО})} + \theta_5 x_{(\text{расстояние до метро})}$$

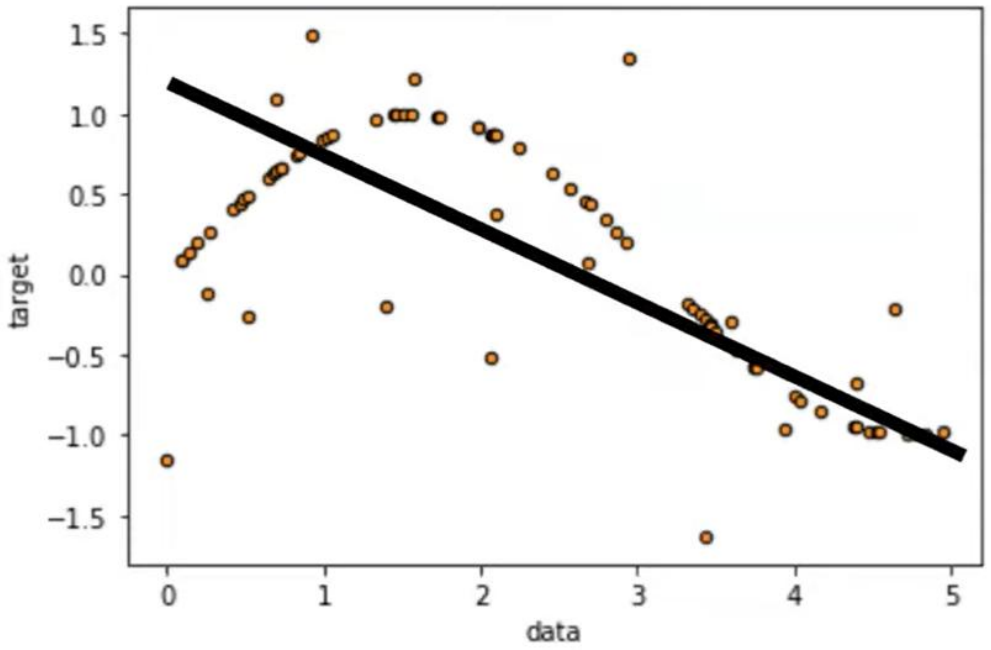
Предсказание стоимости квартиры

$$a(x) = \theta_0 + \theta_1 x_{(\text{площадь})} + \theta_2 x_{(\text{район})} + \theta_3 x_{(\text{расстояние до метро})}$$



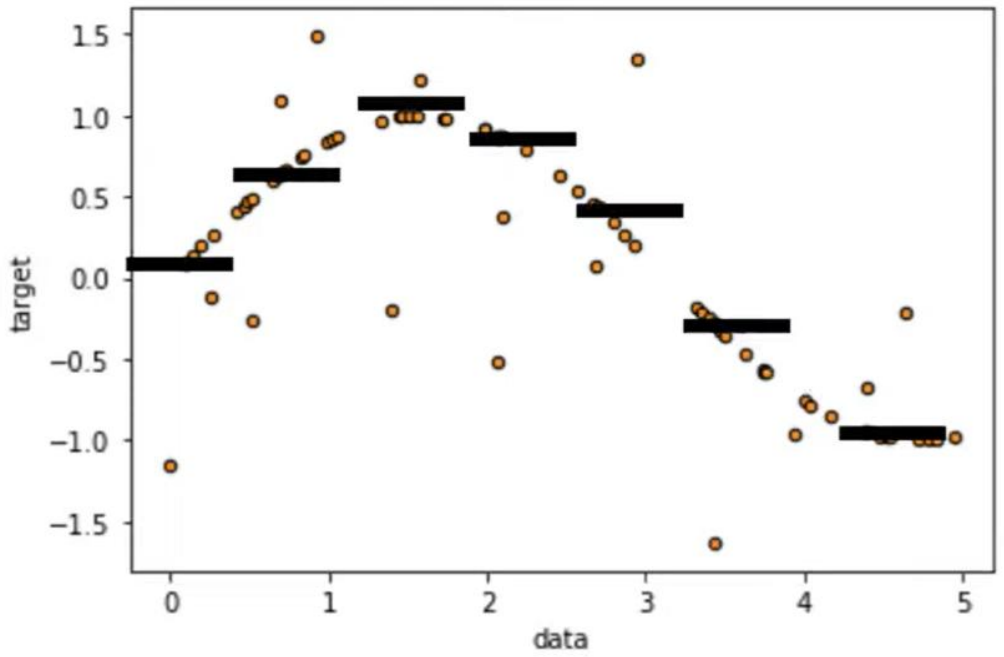
Предсказание стоимости квартиры

$$a(x) = \theta_0 + \theta_1 x_{(\text{площадь})} + \theta_2 x_{(\text{район})} + \theta_3 x_{(\text{расстояние до метро})}$$



Предсказание стоимости квартиры

$$a(x) = \theta_0 + \theta_1 x_{(\text{площадь})} + \theta_2 x_{(\text{район})} + \theta_3 x_{([t_0 \le x_3 \le t_1])} + \cdots + \theta_{3+n} x_{([t_{n-1} \le x_3 \le t_n])}$$



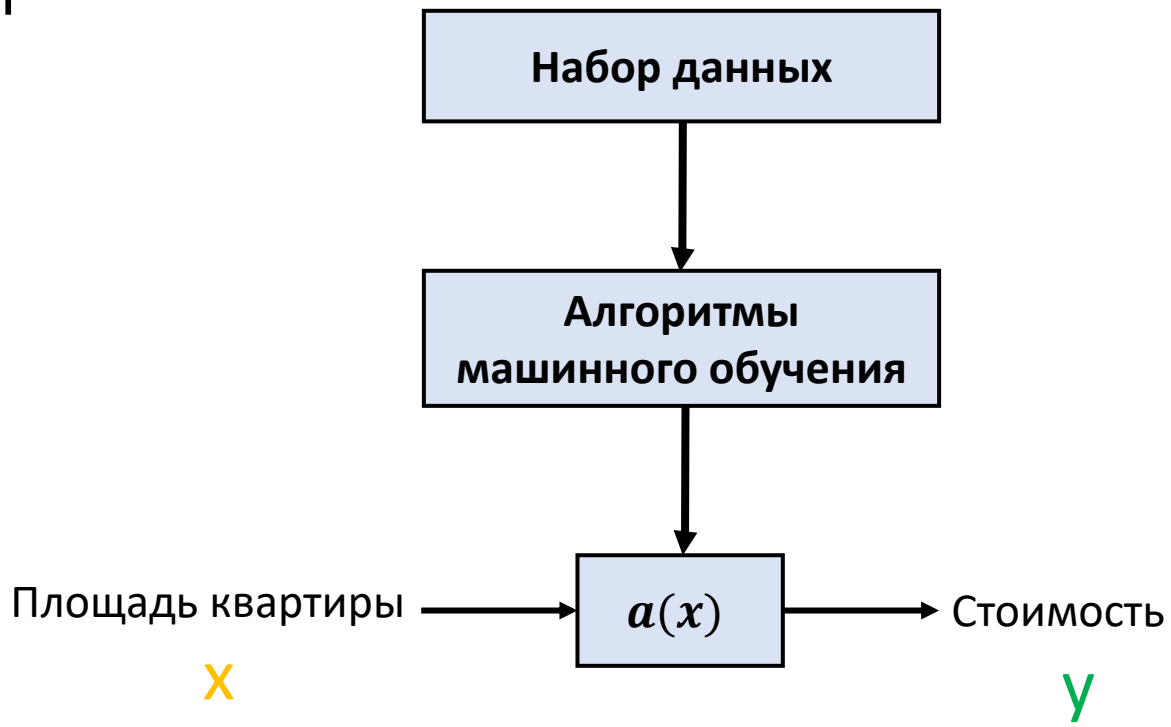
Линейные модели

Модель линейной регрессии хороша, если признаки сделаны специально под неё

Пример: one-hot кодирование категориальных признаков или бинаризация числовых признаков

Задачи регрессии с одной переменной

Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)
70	120
90	140
120	160

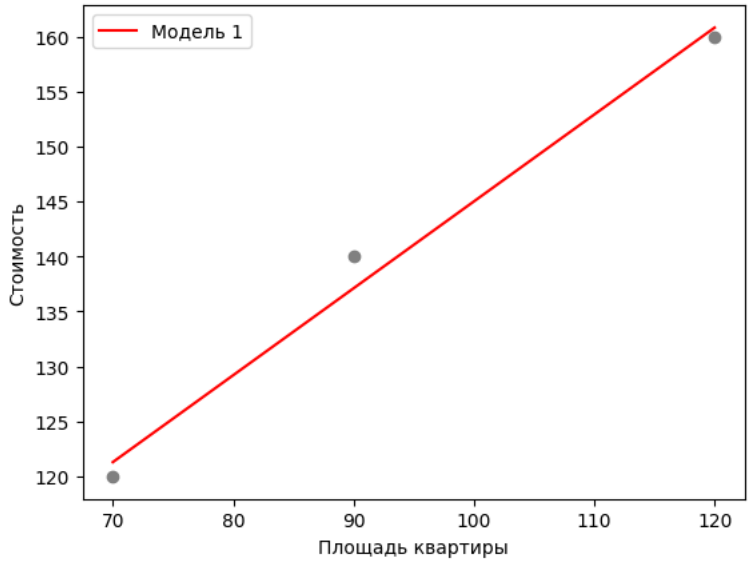


$$a(x, \theta) = \theta_0 + \theta_1 x_{\text{площадь квартиры}}$$

$$\theta_0, \theta_1 = ?$$

Задачи регрессии с одной переменной

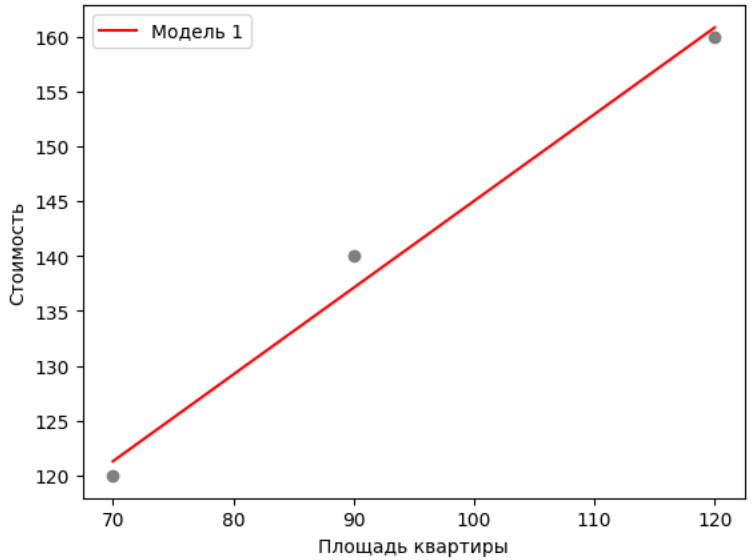
Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость Модель 1
70	120	121.3
90	140	137.1
120	160	160.8



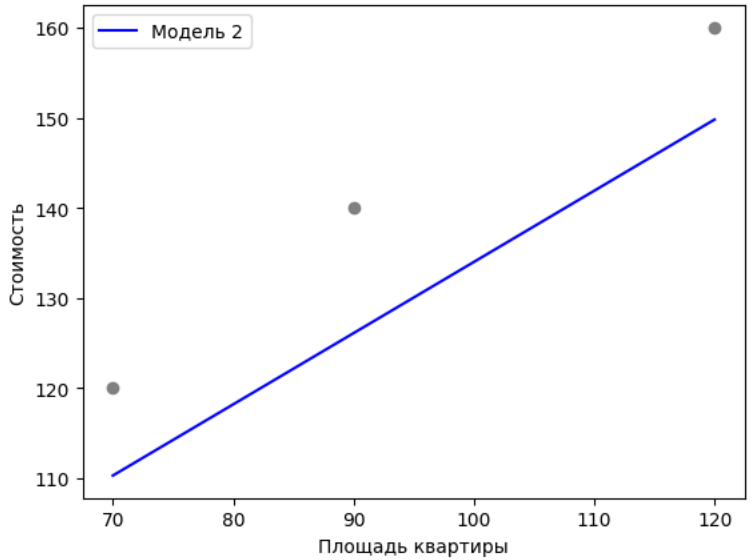
$$\boldsymbol{\theta_0 = 66, \theta_1 = 0.79}$$

Задачи регрессии с одной переменной

Площадь квадрата (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость Модель 1	Стоимость Модель 2
70	120	121.3	110.3
90	140	137.1	126.1
120	160	160.8	149.8



$\boldsymbol{\theta_0 = 66, \theta_1 = 0.79}$



$\boldsymbol{\theta_0 = 55, \theta_1 = 0.79}$

Функция потерь

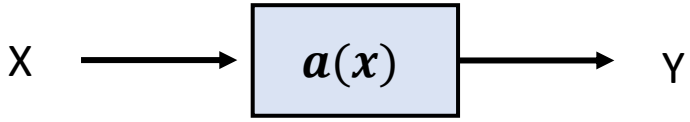
Модель линейной регрессии

$$a(x) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_d x_d$$

Модель линейной регрессии

1. Обучение (Train)

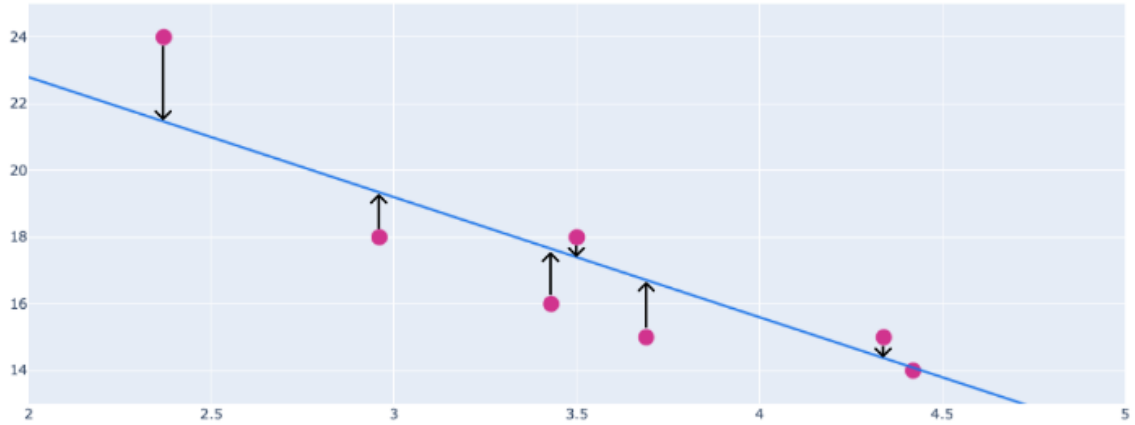
2. Предсказание (Predict)



$$a(x) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_d x_d$$

Функция потерь (loss function)

Хорошо ли работает модель?

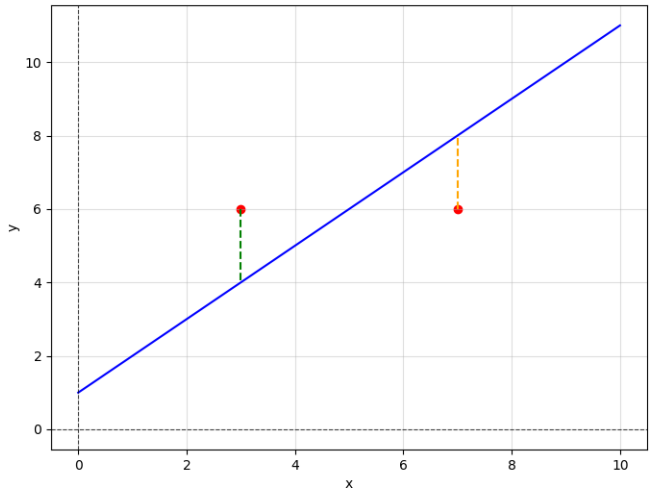


Функция потерь показывает насколько сильно ошибается модель на конкретном объекте

$$L(a(x), y)$$

Расстояние потери

$$L(a(x), y) = a(x) - y$$



1. Потеря фокусируется на расстоянии между значениями, а не на направлении

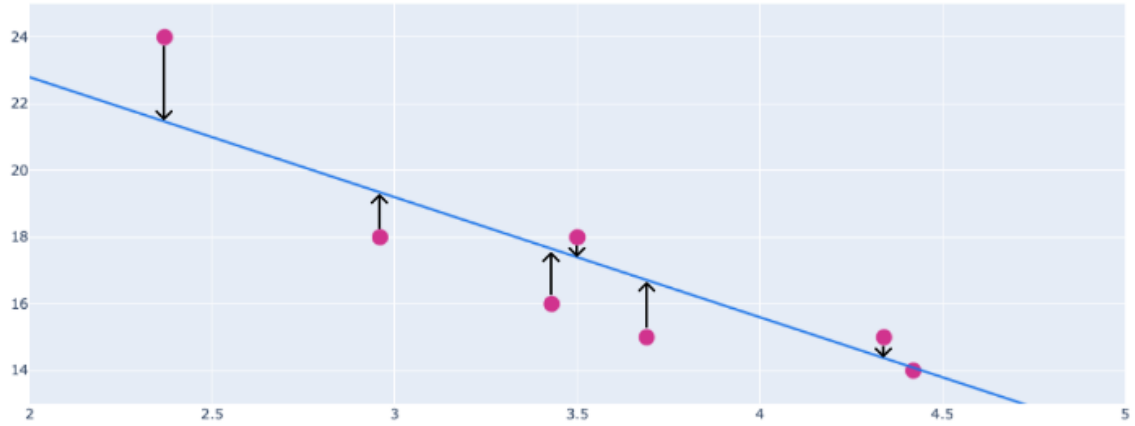
2. Неудобно для метода оптимизации

Функция потерь (loss function)

Функция потерь для задач регрессии:

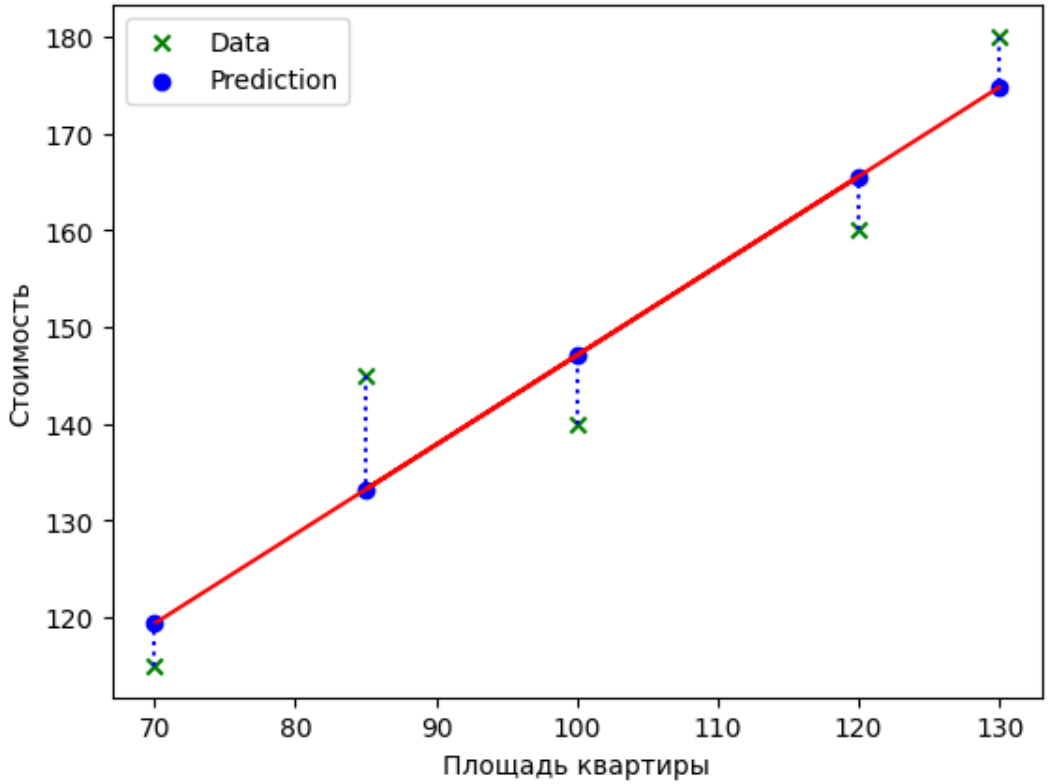
$\mathrm{L(a(x), y)} = (a(x) - y(x))^2$ – Квадратичная ошибка

$\mathrm{L(a(x), y)} = |a(x) - y(x)|$ - Абсолютное значение ошибки



Функция потерь (loss function)

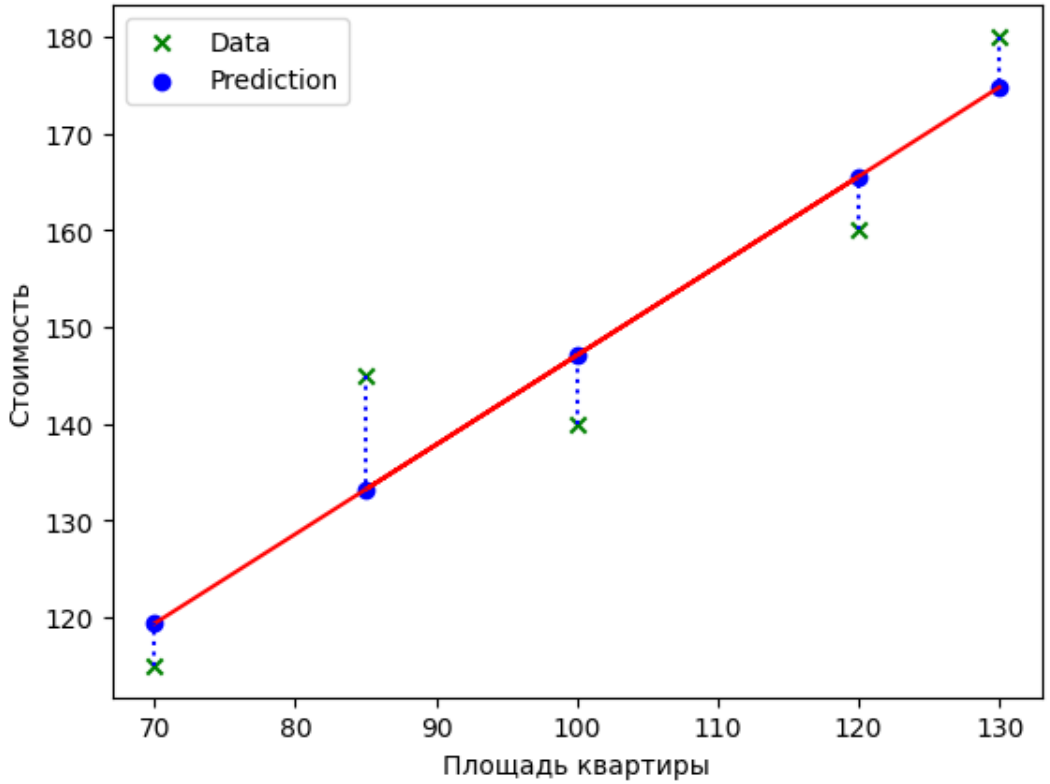
Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость $a(x)$	$L(a(x), y)$
70	115	119.3	
100	140	147.7	
120	160	165.5	
85	145	133.2	
130	180	174.78	



$L(a(x), y) = (a(x) - y(x))^2$ – Квадратичная ошибка

Функция потерь (loss function)

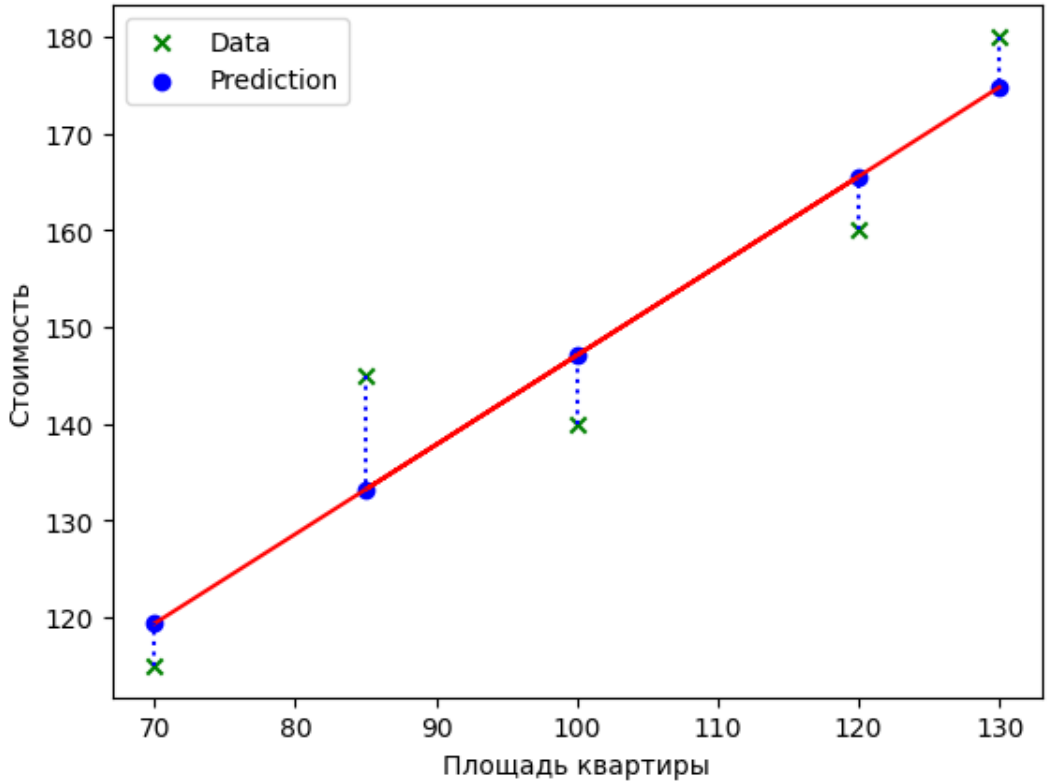
Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость $a(x)$	$L(a(x), y)$
70	115	119.3	19.09
100	140	147.7	
120	160	165.5	
85	145	133.2	
130	180	174.78	



$L(a(x), y) = (a(x) - y(x))^2$ – Квадратичная ошибка

Функция потерь (loss function)

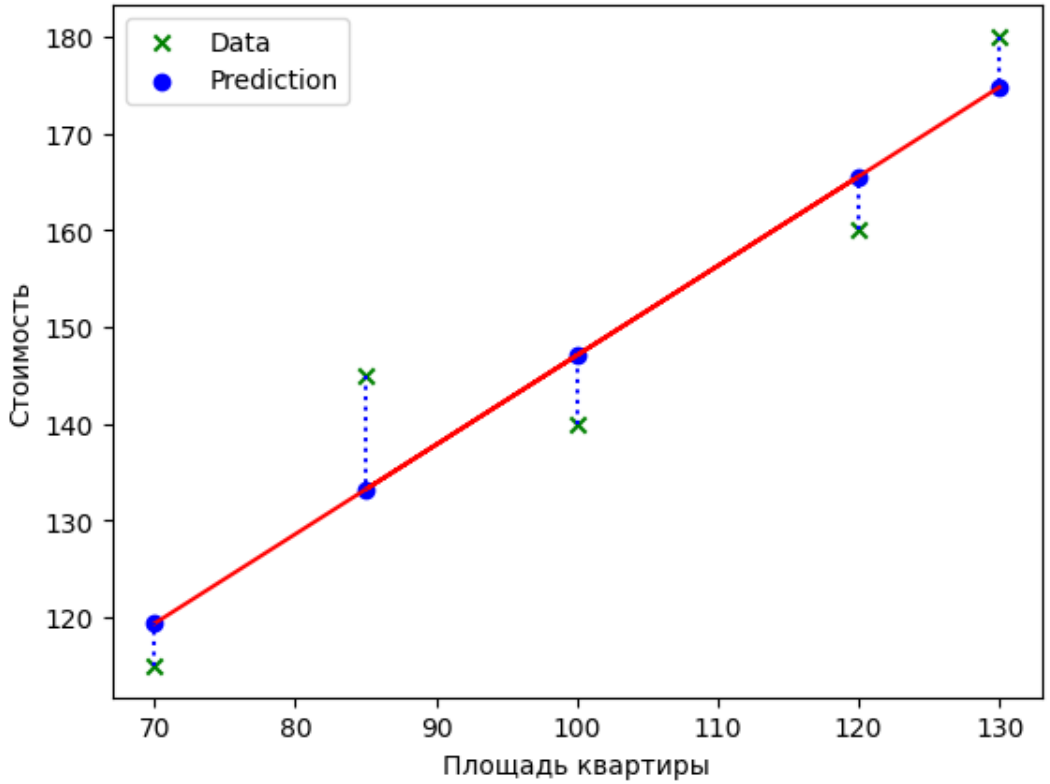
Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость $a(x)$	$L(a(x), y)$
70	115	119.3	19.09
100	140	147.7	50.07
120	160	165.5	
85	145	133.2	
130	180	174.78	



$L(a(x), y) = (a(x) - y(x))^2$ – Квадратичная ошибка

Функция потерь (loss function)

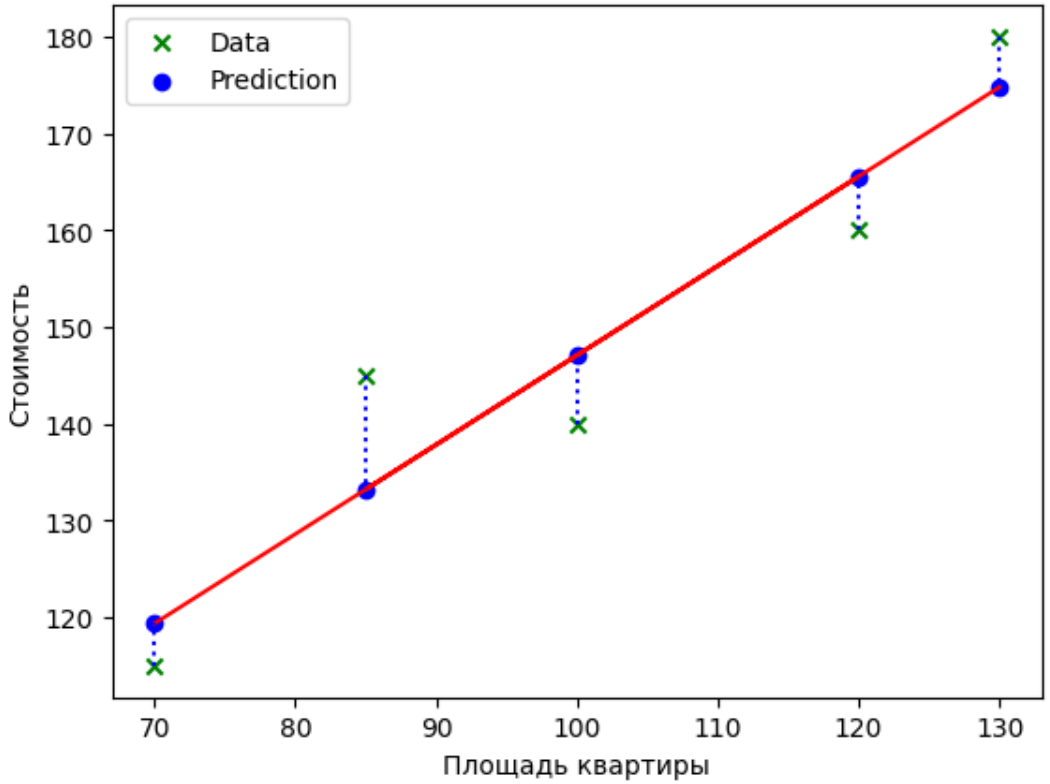
Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость $a(x)$	$L(a(x), y)$
70	115	119.3	19.09
100	140	147.7	50.07
120	160	165.5	30.7
85	145	133.2	
130	180	174.78	



$L(a(x), y) = (a(x) - y(x))^2$ – Квадратичная ошибка

Функция потерь (loss function)

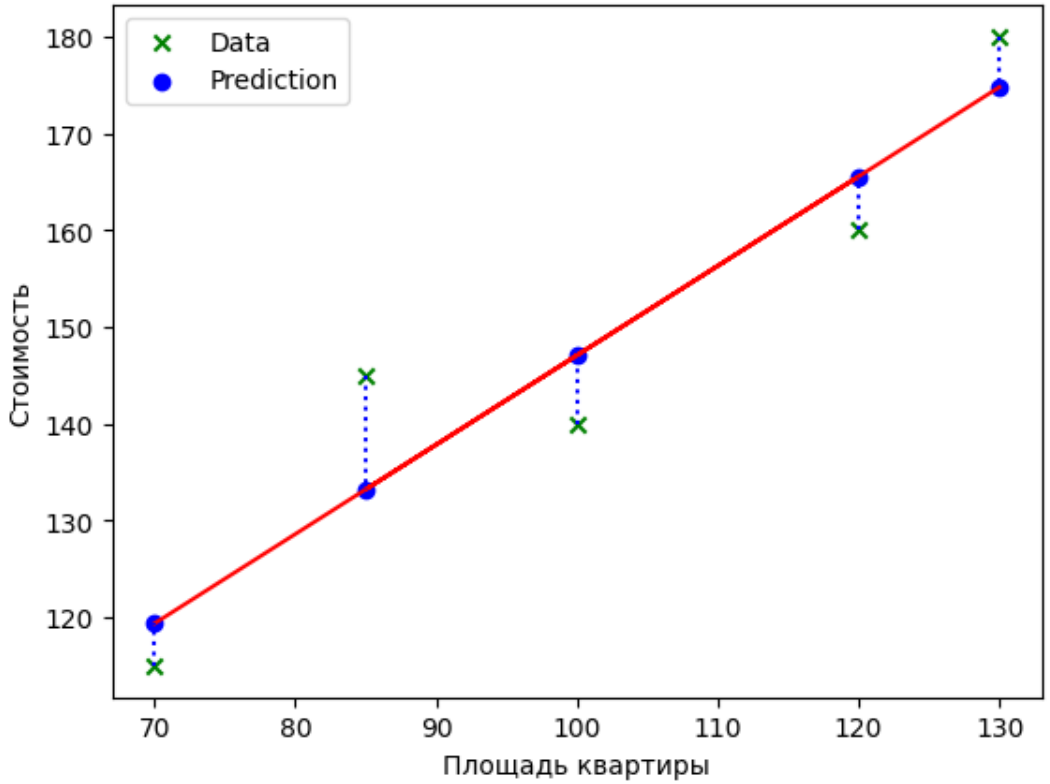
Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость $a(x)$	$L(a(x), y)$
70	115	119.3	19.09
100	140	147.7	50.07
120	160	165.5	30.7
85	145	133.2	138.6
130	180	174.78	



$L(a(x), y) = (a(x) - y(x))^2$ – Квадратичная ошибка

Функция потерь (loss function)

Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость $a(x)$	$L(a(x), y)$
70	115	119.3	19.09
100	140	147.7	50.07
120	160	165.5	30.7
85	145	133.2	138.6
130	180	174.78	27.2



$L(a(x), y) = (a(x) - y(x))^2$ – Квадратичная ошибка

Функционал ошибки

Эмпирический риск – функционал качества алгоритма a на $\mathbb{X}$

$$\mathrm{Q}(\mathrm{a}, \mathbb{X})$$

Среднеквадратичная ошибка (mean squared error, MSE)

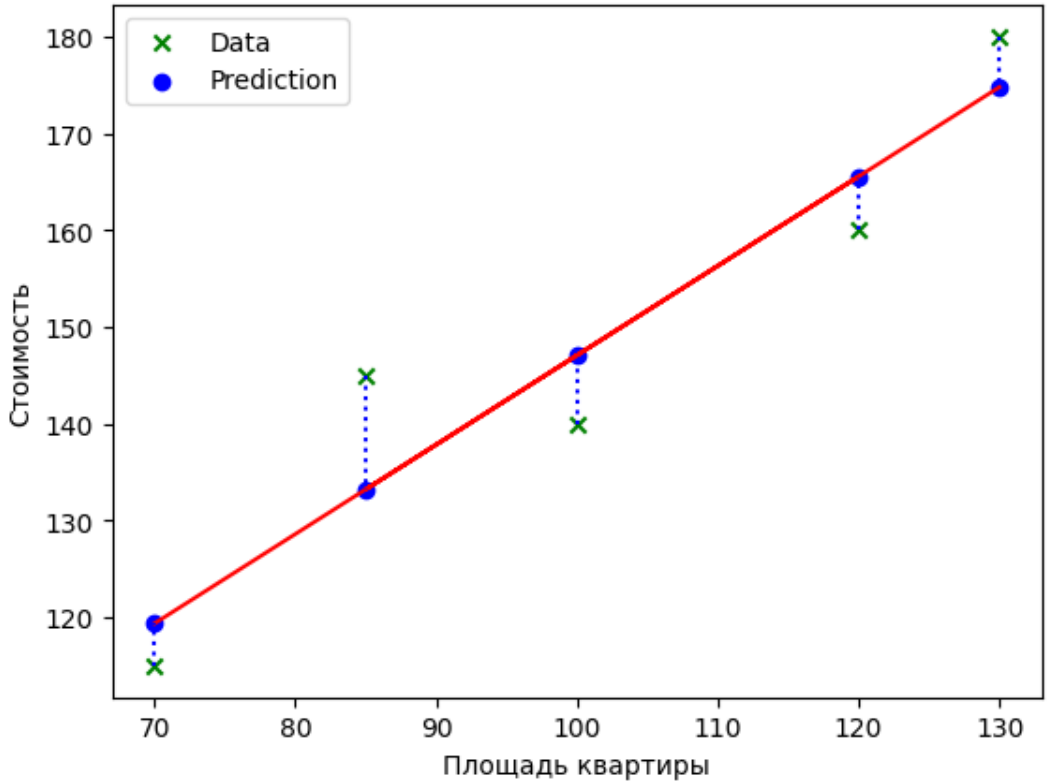
$$\mathrm{Q}(\mathrm{a}, \mathbb{X}) = \frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2$$

Средняя абсолютная ошибка (Mean Absolute Error, MAE)

$$\mathrm{Q}(\mathrm{a}, \mathbb{X}) = \frac{1}{\ell}\sum_{i=1}^{\ell}|a(x_i) - y_i|$$

Функционал ошибки

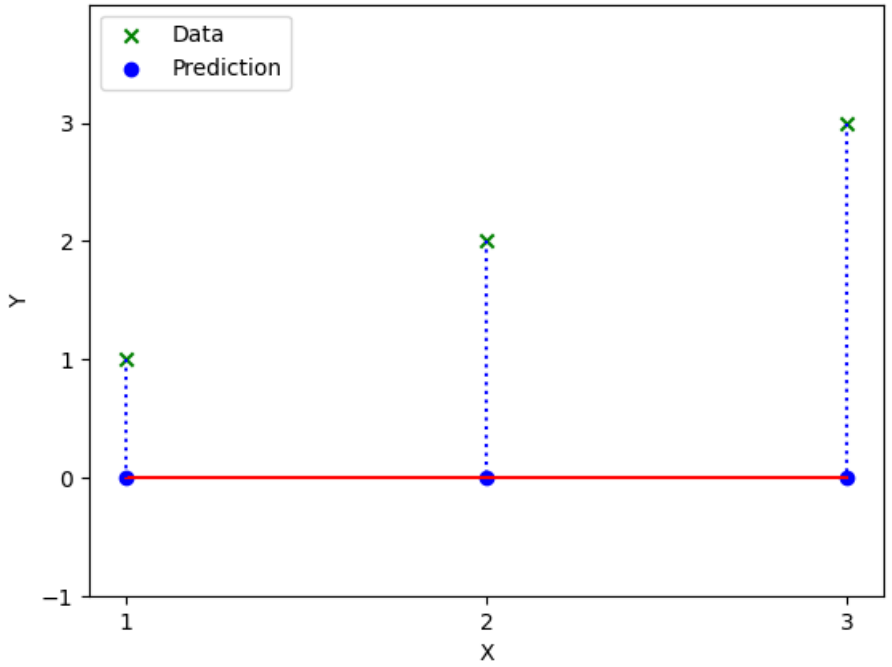
Площадь квартиры (x_1)	Стоимость квартиры ($\mathbb{Y}$)	Стоимость $a(x)$	$L(a(x), y)$
70	115	119.3	19.09
100	140	147.7	50.07
120	160	165.5	30.7
85	145	133.2	138.6
130	180	174.78	27.2

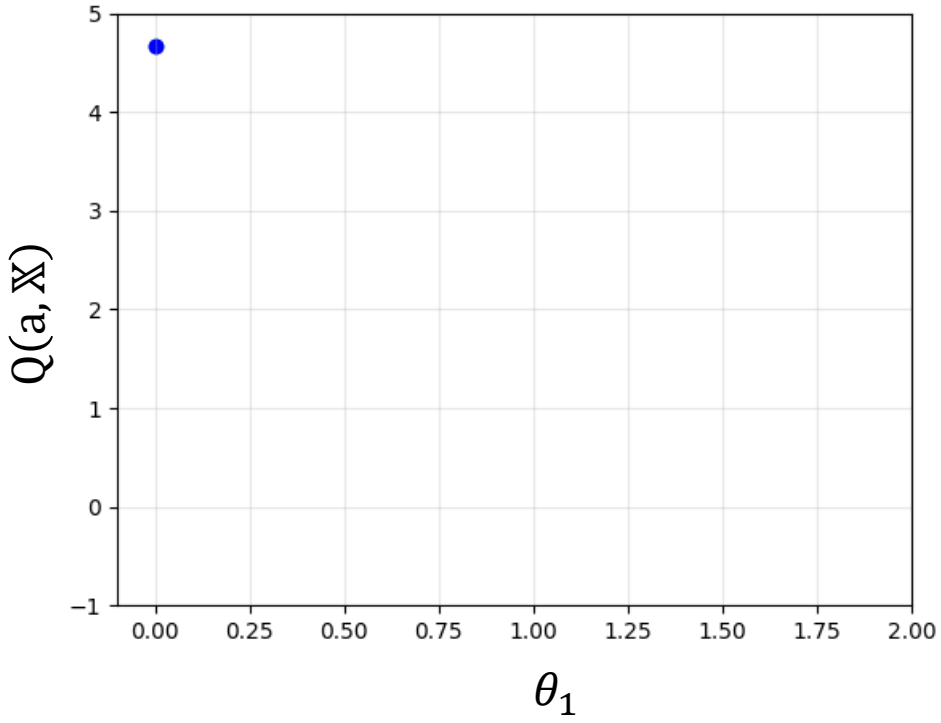


Среднеквадратичная ошибка $$Q(a, \mathbb{X}) = \frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2$$

$$Q(a, \mathbb{X}) = \frac{1}{5}(19.09 + 50.07 + 30.7 + 138.6 + 27.2) = 53.13$$

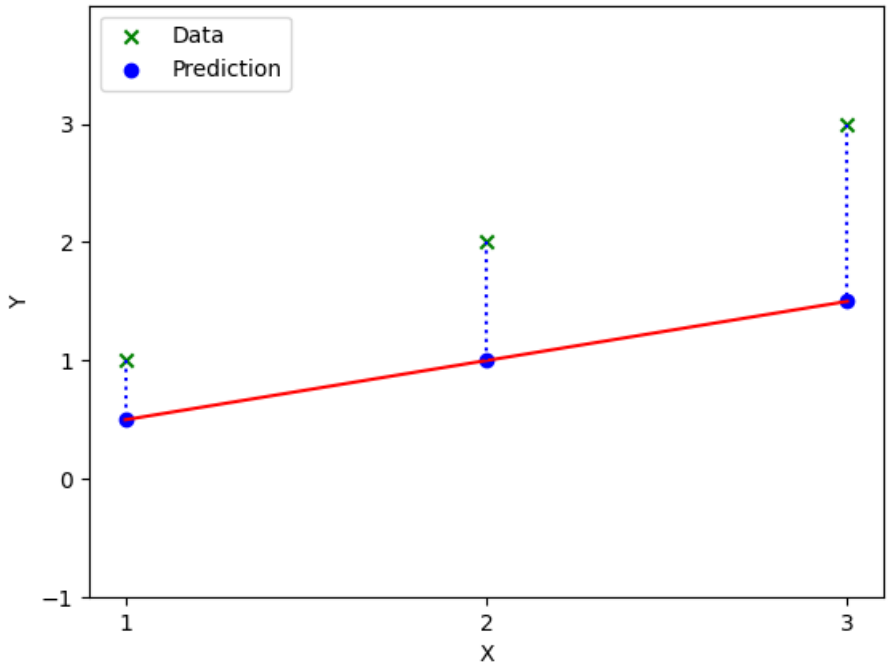
Функция потерь (Lost function)

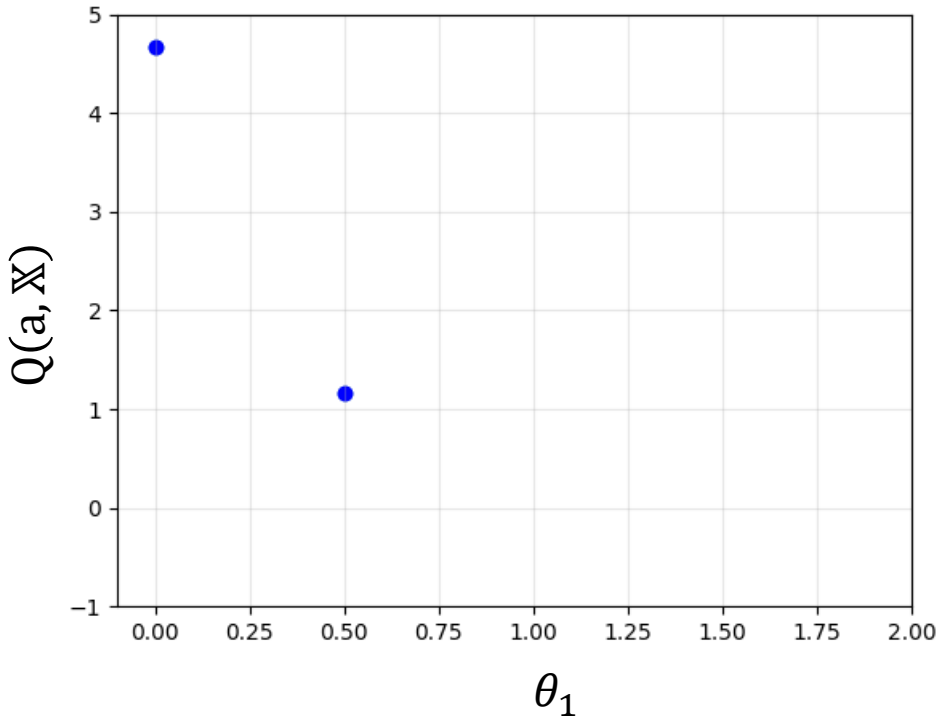




$\boldsymbol{\theta_1 = 0}$ **MSE**$\boldsymbol{= 4.6}$

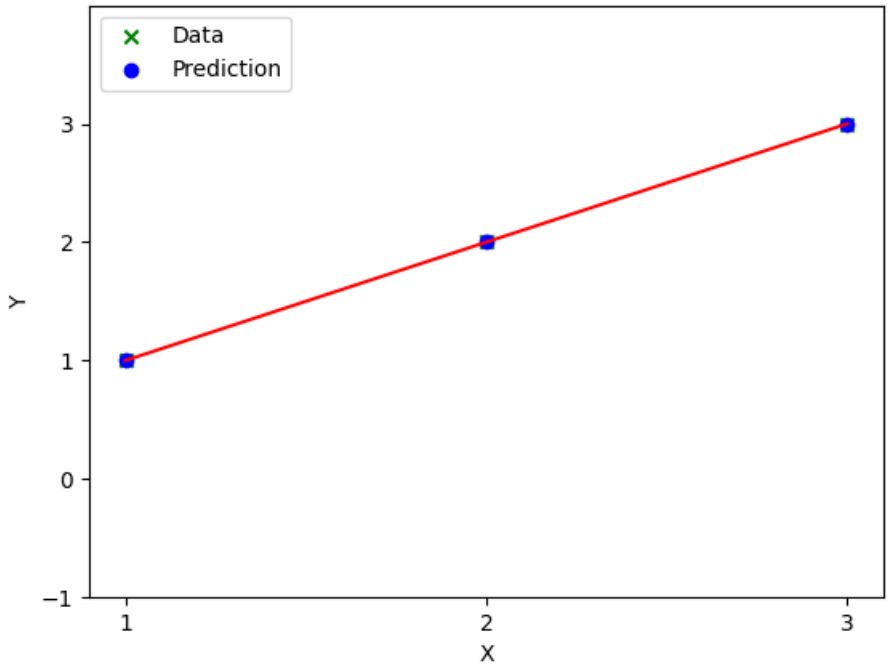
Функция потерь (Lost function)

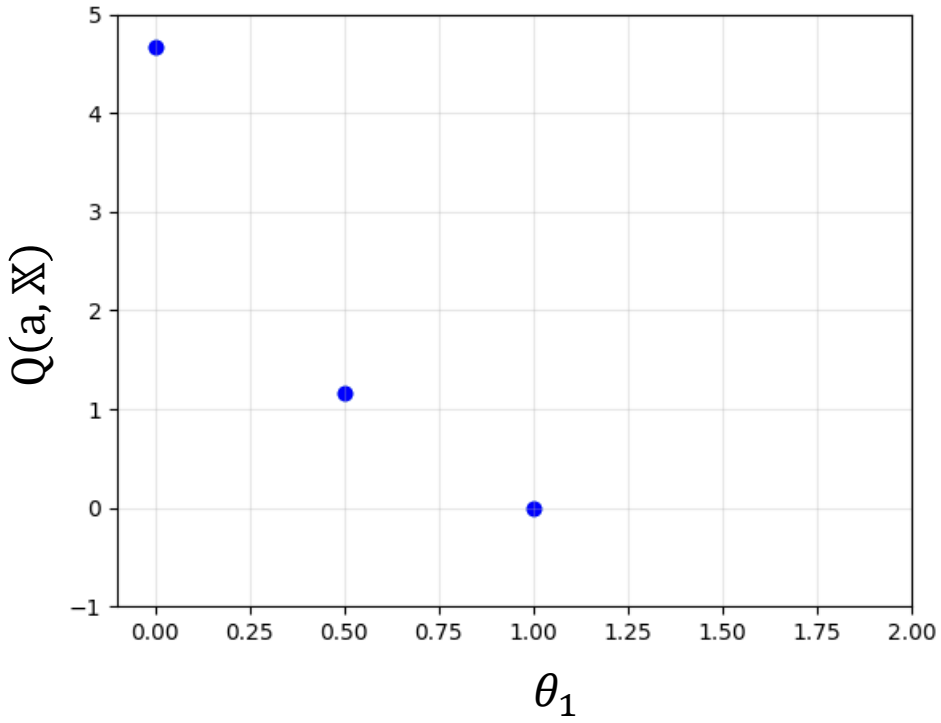




$\boldsymbol{\theta_1 = 0.5}$ **MSE**$\boldsymbol{= 1.1}$

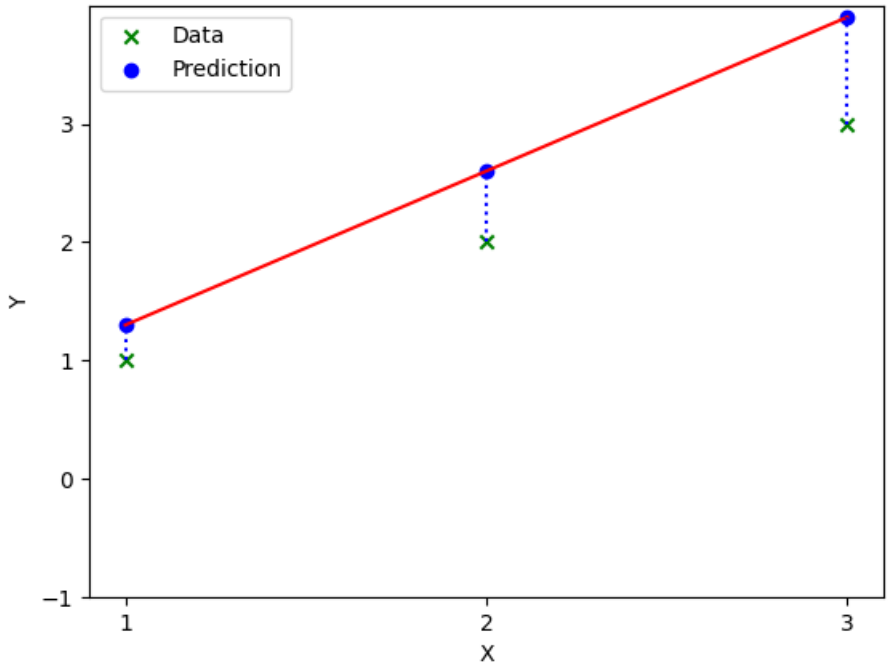
Функция потерь (Lost function)

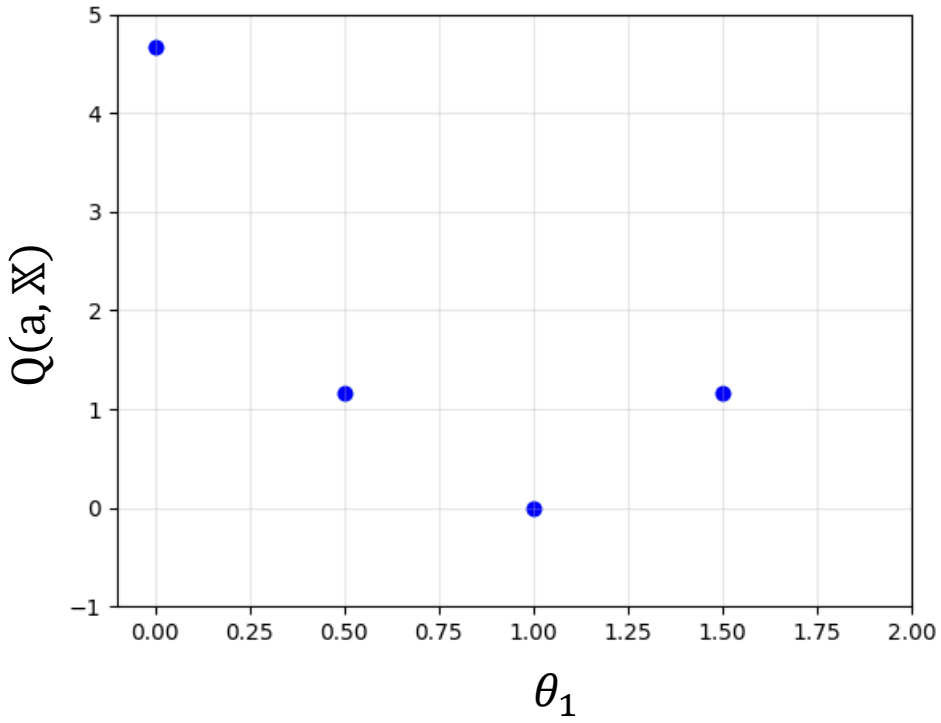




$\boldsymbol{\theta_1 = 1}$ **MSE= 0**

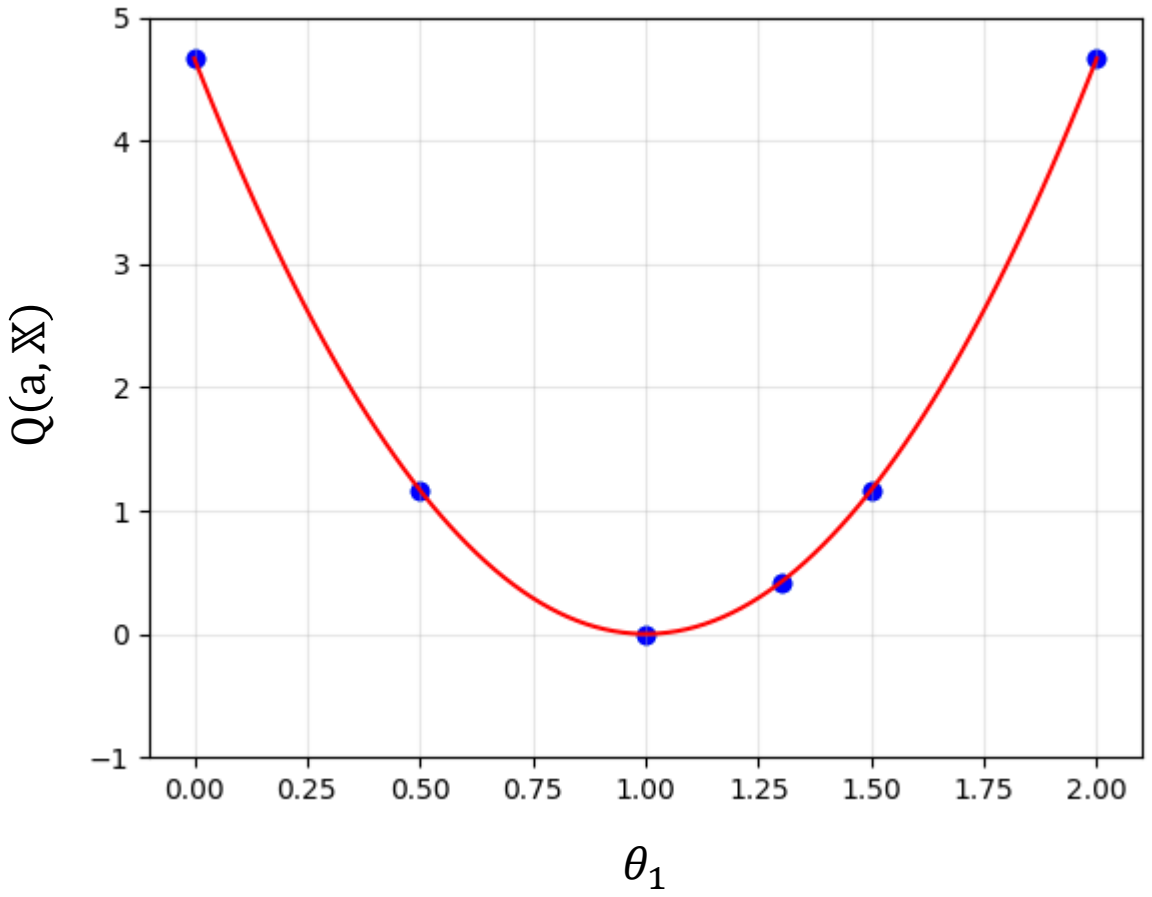
Функция потерь (Lost function)



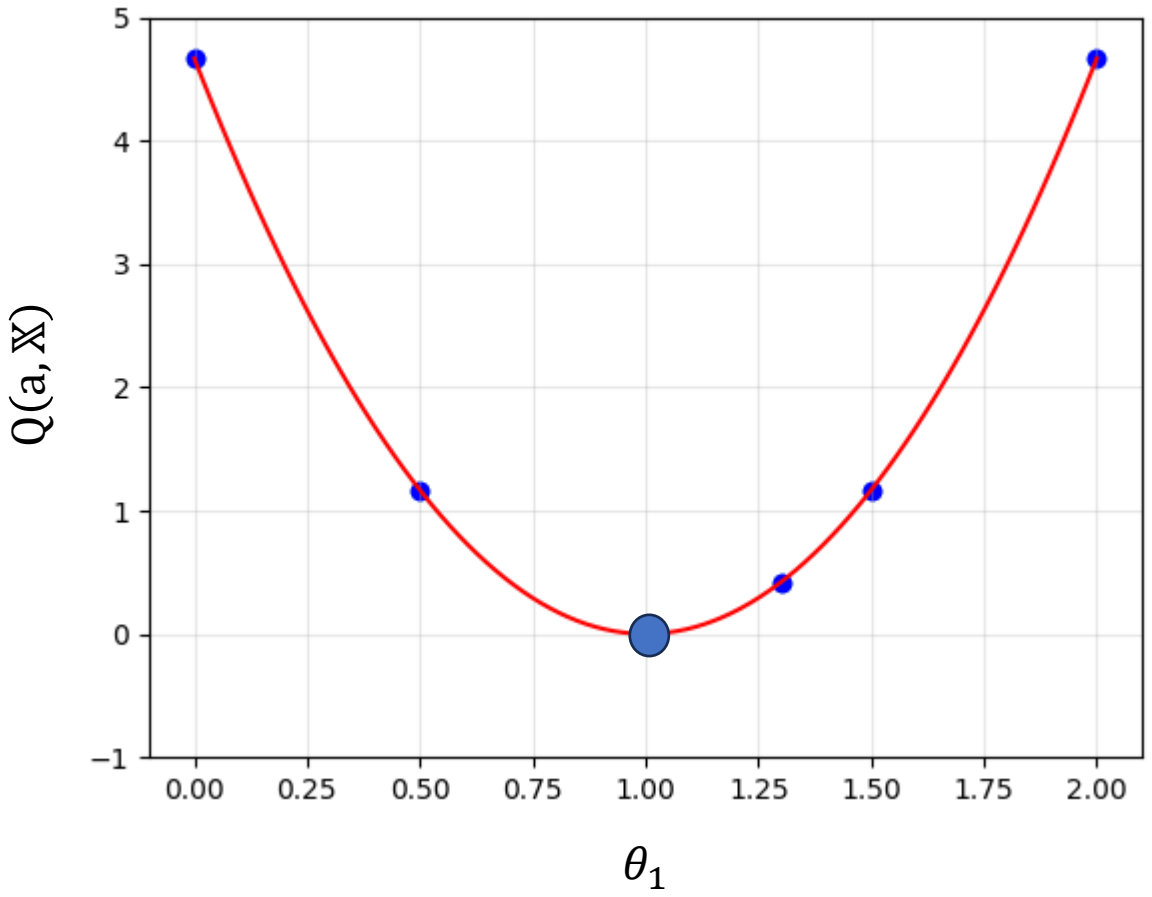


$\boldsymbol{\theta_1 = 1.5}$ **MSE**$\boldsymbol{= 1.1}$

Функция потерь (Lost function)



Функция потерь (Lost function)



Функция потерь (Lost function)

Модель

$$\boldsymbol{a(x) = \theta_0 + \theta_1 x_1}$$

Параметры модели

$$(\boldsymbol{\theta_0, \theta_1})$$

Функционал ошибки

$$Q(\mathrm{a}, \mathbb{X}) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Цель

$$\underset{\theta}{minimize}\, Q(\mathrm{a}, \mathbb{X})$$