SCHOOL
OF ENERGY
& POWER ENGINEERING

Обучение регрессоров

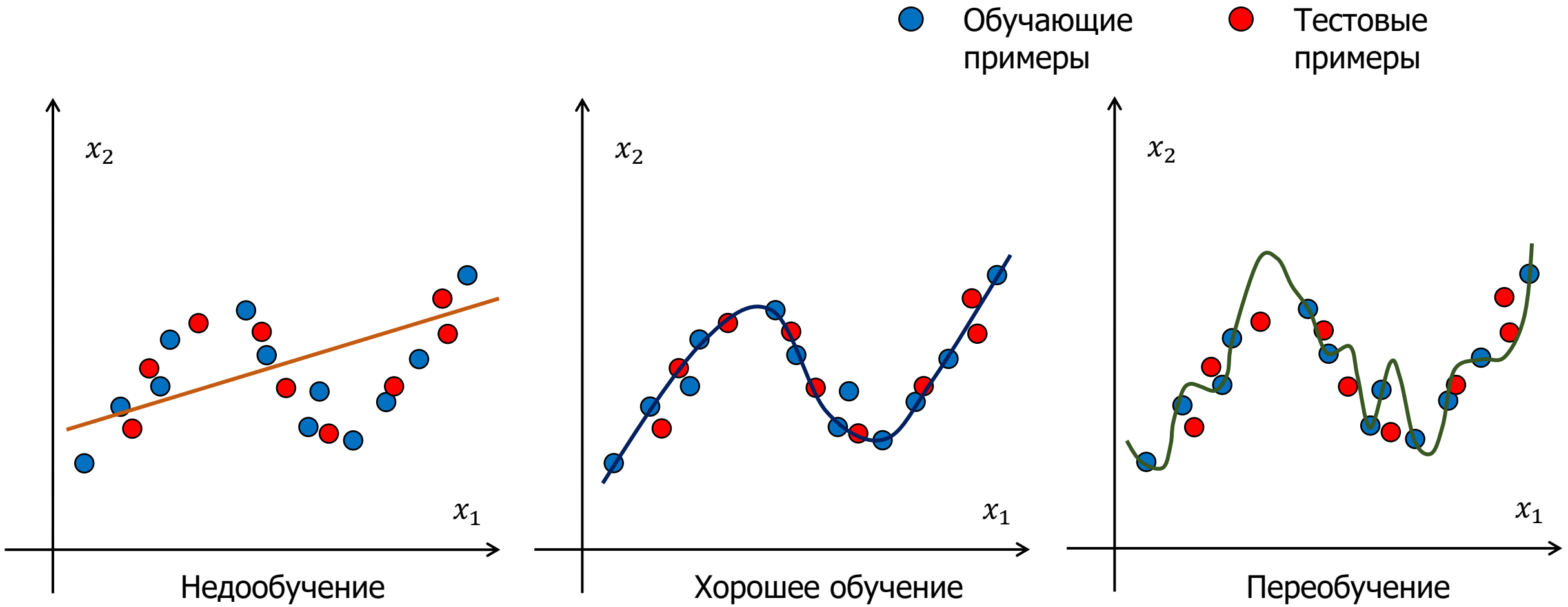
Сергей Владимирович Аксёнов,

Доцент отделения информационных технологий ИШИТР,

Томский политехнический университет

Томск-2023

Плохое и хорошее обучение



Метрики -1

1. Средняя квадр. Ошибка (СКО): $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

2. Квадрат СКО: $$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}$$

3. Относит. квадр. ошибка (ОКО): $$RSE = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

4. Корень ОКО: $$RRSE = \sqrt{RSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

y_i - Истинные значения

$\tilde{y}_i$ - Предсказанное значение

$\bar{y}$ - Среднее значение

Метрики-2

5. Средняя абс. ошибка: $$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widetilde{y_i}|$$

6. Относит. абс. ошибка: $$RAE = \frac{\sum_{i=1}^{n}|y_i - \widetilde{y_i}|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$$

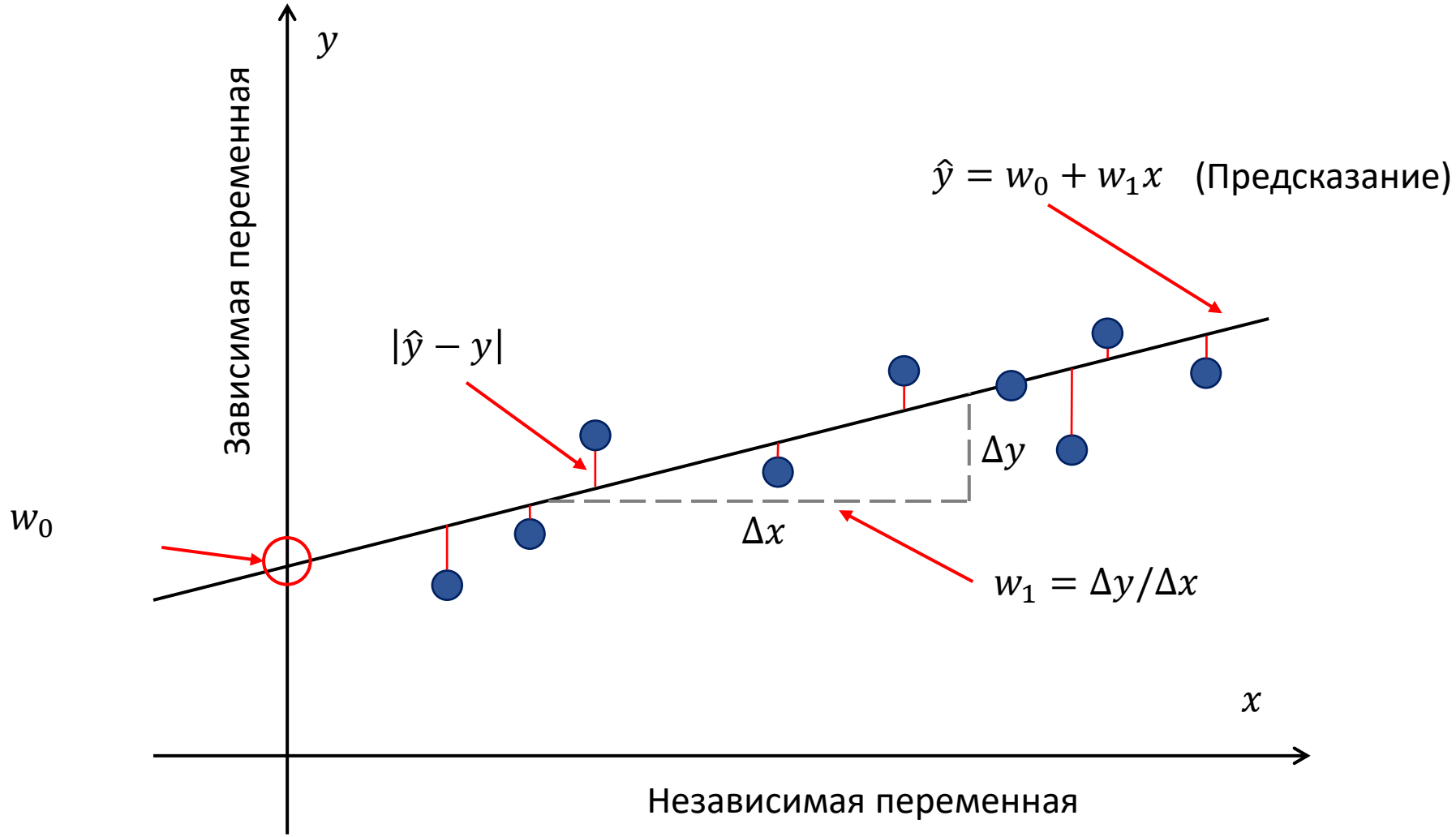
6. Коэффициент детерминации: $$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widetilde{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

y_i - Истинные значения

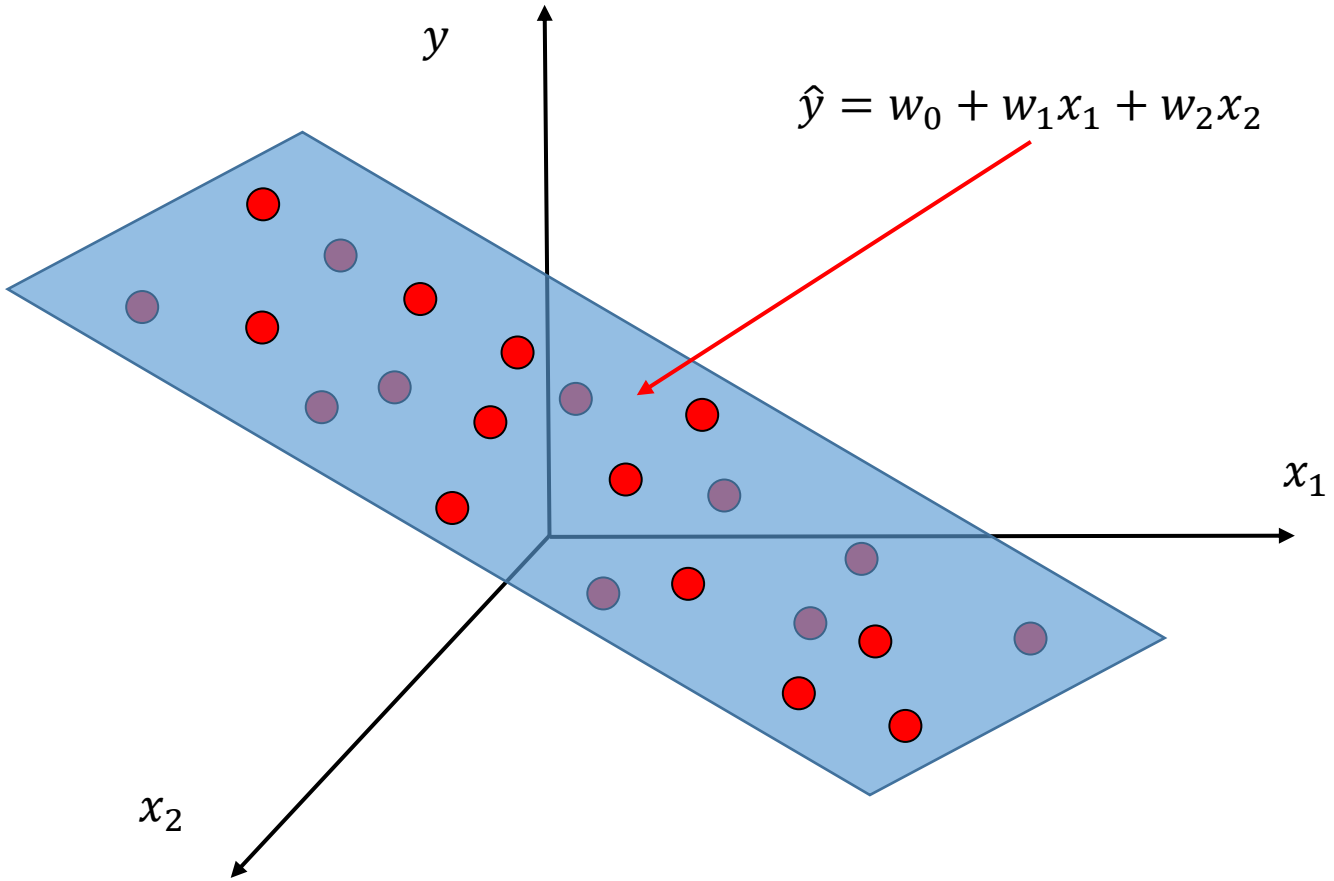
$\widetilde{y_i}$ - Предсказанное значение

$\bar{y}$ - Среднее значение

Линейная регрессия: один признак



Линейная регрессия: два признака



Линейные корреляции

	Цемент	Супер-пластификатор	Вода	Прочность бетона
Цемент	1	0.09	-0.08	0.5
Супер-пластификатор	0.09	1	-0.66	0.37
Вода	-0.08	-0.66	1	-0.29
Прочность бетона	0.5	0.37	-0.29	1

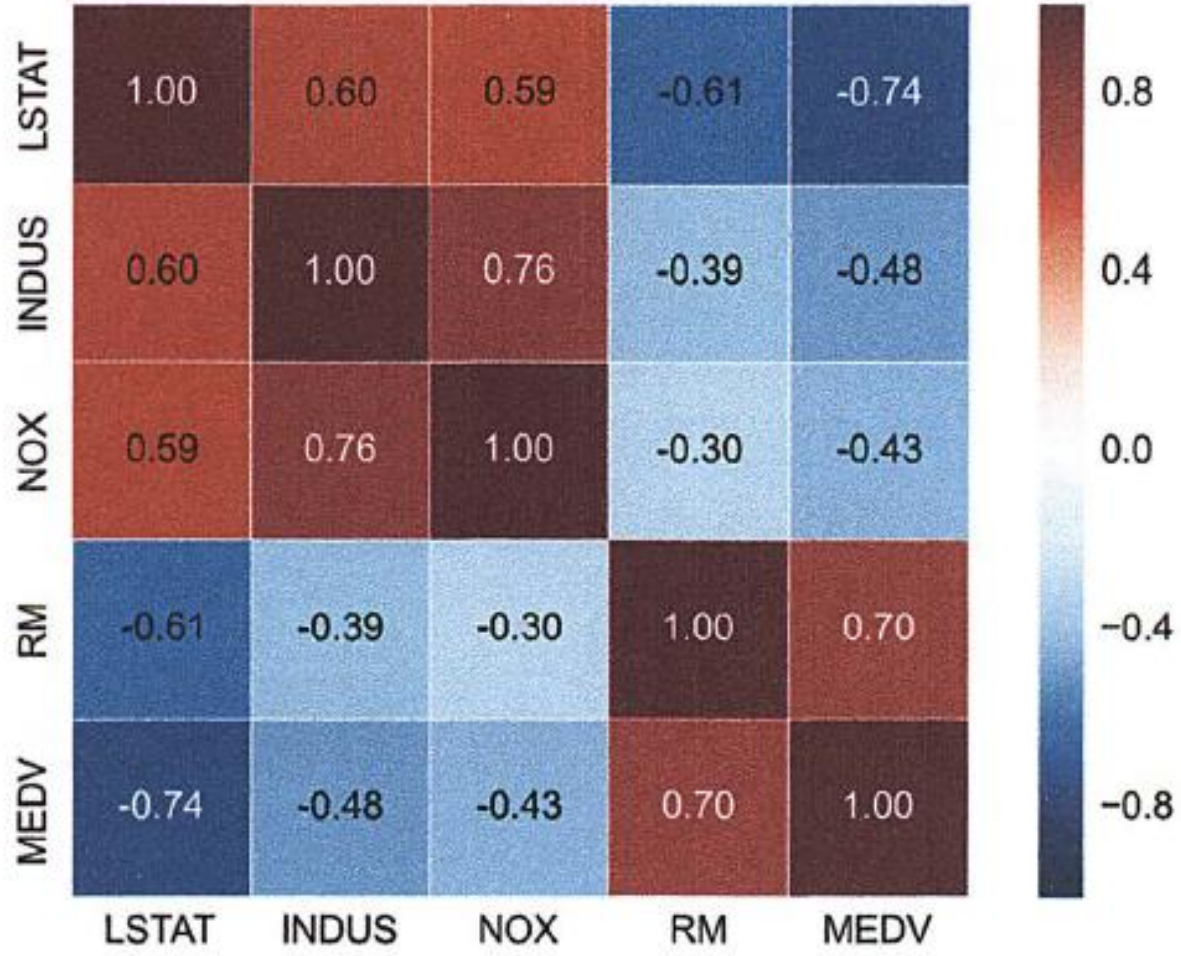
Коэффициент корреляции Пирсона:

$$r = \frac{\sum_{i=1}^{n}\left[(a_i - \overline{a})\left(b_i - \overline{b}\right)\right]}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2} \cdot \sqrt{\sum_{i=1}^{n}(b_i - \overline{b})^2}}$$

a, b – Признаки

$\overline{a}, \overline{b}$ – Выборочное среднее для a,b

Тепловая карта



Регуляризация в регрессионных моделях

Гребневая регрессия:

$$J(w)_{Ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\|w\|_2^2 \qquad L2\colon \lambda\|w\|_2^2 = \lambda\sum_{j=1}^{m} w_j^2$$

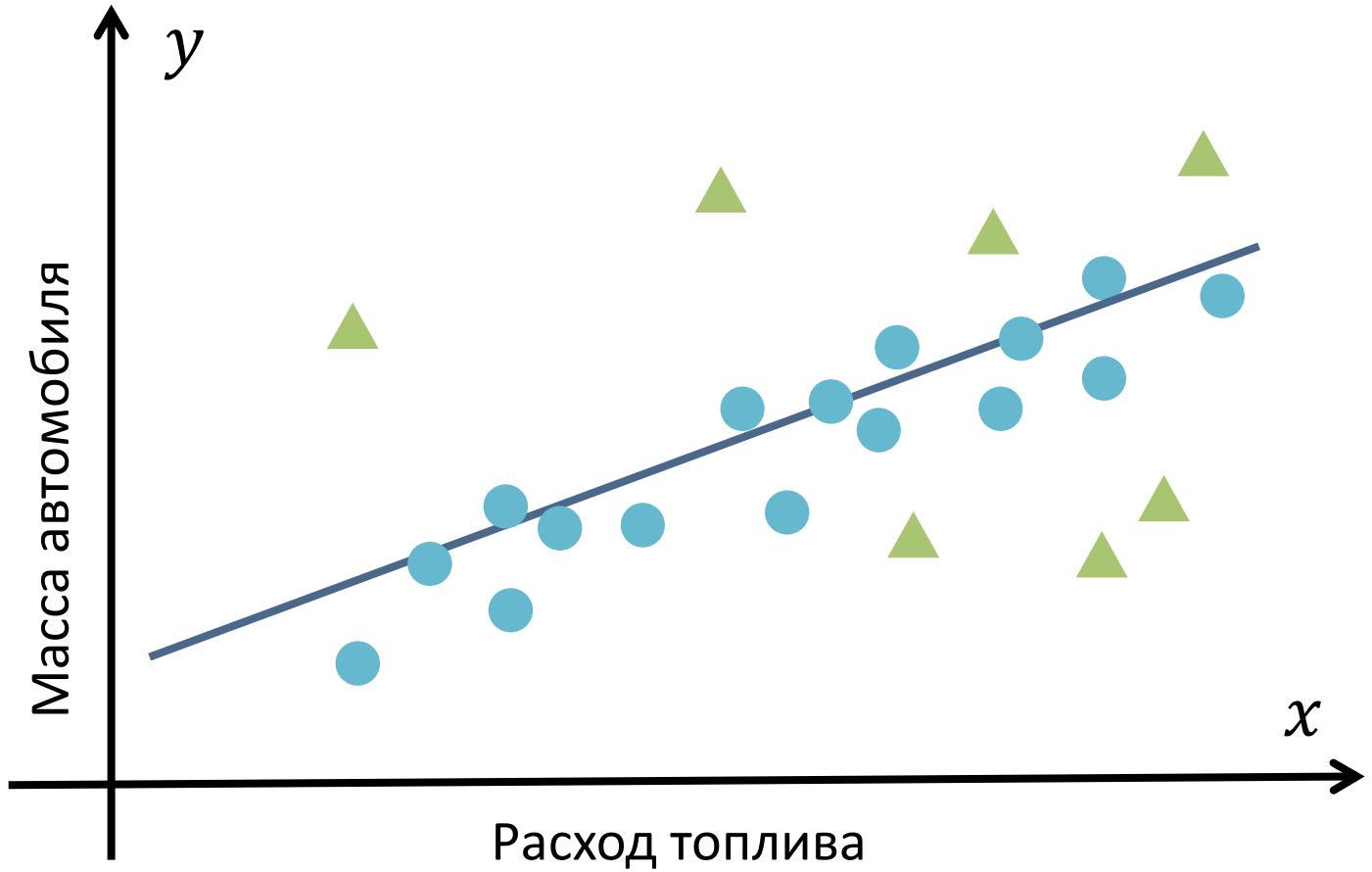
Метод Lasso:

$$J(w)_{Lasso} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\|w\|_1 \qquad L1\colon \lambda\|w\|_1 = \lambda\sum_{j=1}^{m}|w_j|$$

Метод эластичной сети:

$$J(w)_{Elastic\_Net} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2$$

Учет выбросов. RANSAC



– Объекты, использующиеся при получении модели

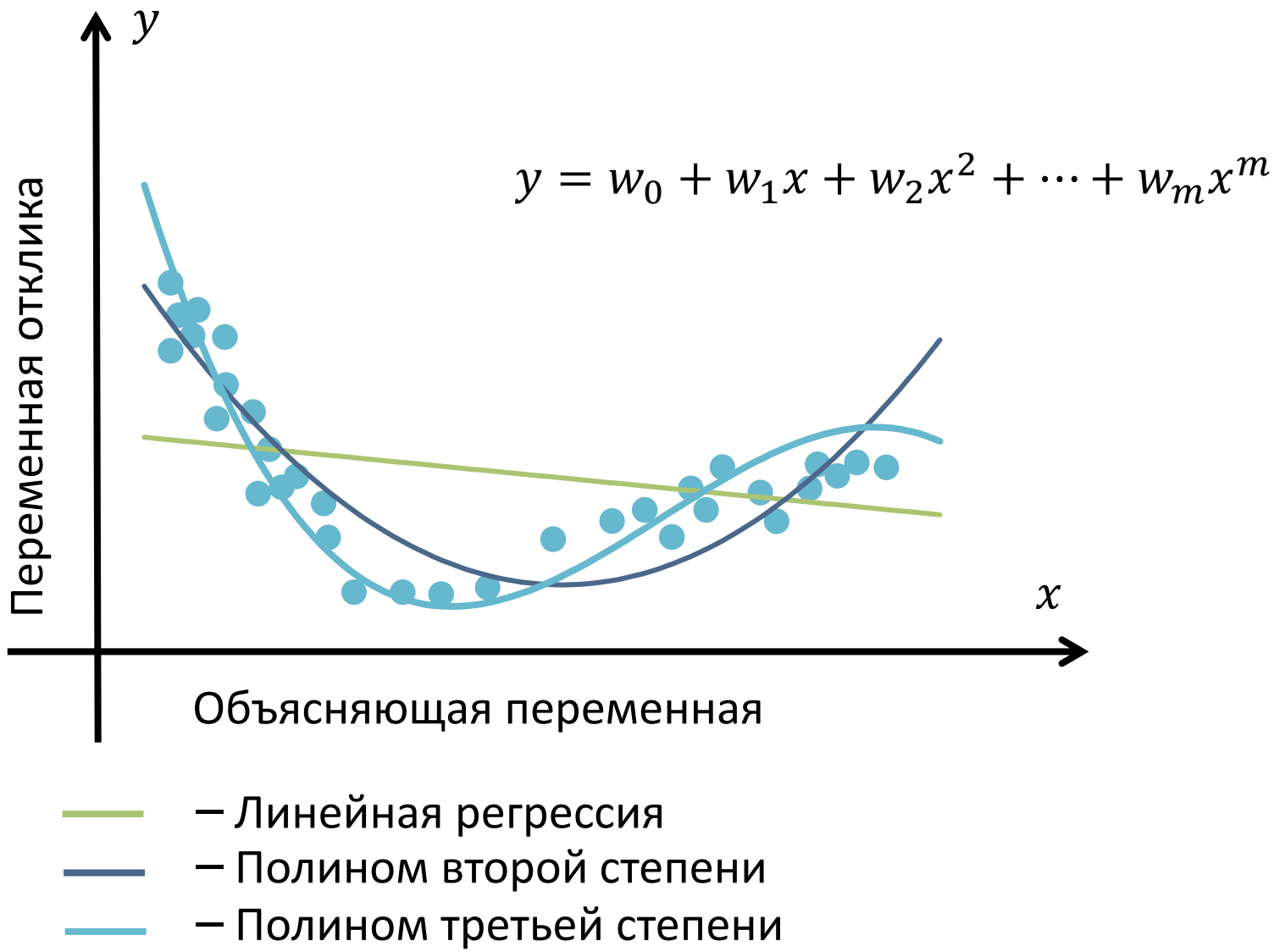
– Выбросы, не влияющие на модель

Полиномиальная регрессия

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_m x^m$$

Примеры:	Начальный набор:	Новый набор:
Квадратичная регрессия (Степень=2):	x	x, x^2
Кубическая регрессия (Степень=3):	x	x, x^2, x^3
Квадратичная регрессия (Степень=2):	x_1, x_2	$x_1, x_2, x_1 x_2, x_1^2, x_2^2$
Кубическая регрессия (Степень=3):	x_1, x_2	$x_1, x_2, x_1 x_2, x_1^2, x_2^2, x_1 x_2^2, x_2 x_1^2, x_1^3, x_2^3$

Сравнение регрессионных моделей. Пример



Регрессия с помощью дерева

Прирост информации, использующийся для бинарного расщепления:

$$IG(D_p, x) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

Мера неоднородности (энтропия) для регрессии:

$$I(t) = MSE(t) - \frac{1}{N_t} \sum_{i \in D_t}^{n} \left(y^{(i)} - \hat{y}_t\right)^2$$

Предсказанное целевое значение для узла дерева:

$$\hat{y}_t = \frac{1}{N} \sum_{i \in D_t} y^{(i)}$$

$\boldsymbol{D_p}, \boldsymbol{D_{left}}, \boldsymbol{D_{right}}$ – Набор образцов в родительском, левом и правом дочерних узлах после расщепления

$\boldsymbol{N_p}, \boldsymbol{N_{left}}, \boldsymbol{N_{right}}$ – Количество образцов в узлах

Пример регрессии с помощью дерева