National Research Tomsk State University

Обучение регрессионных моделей

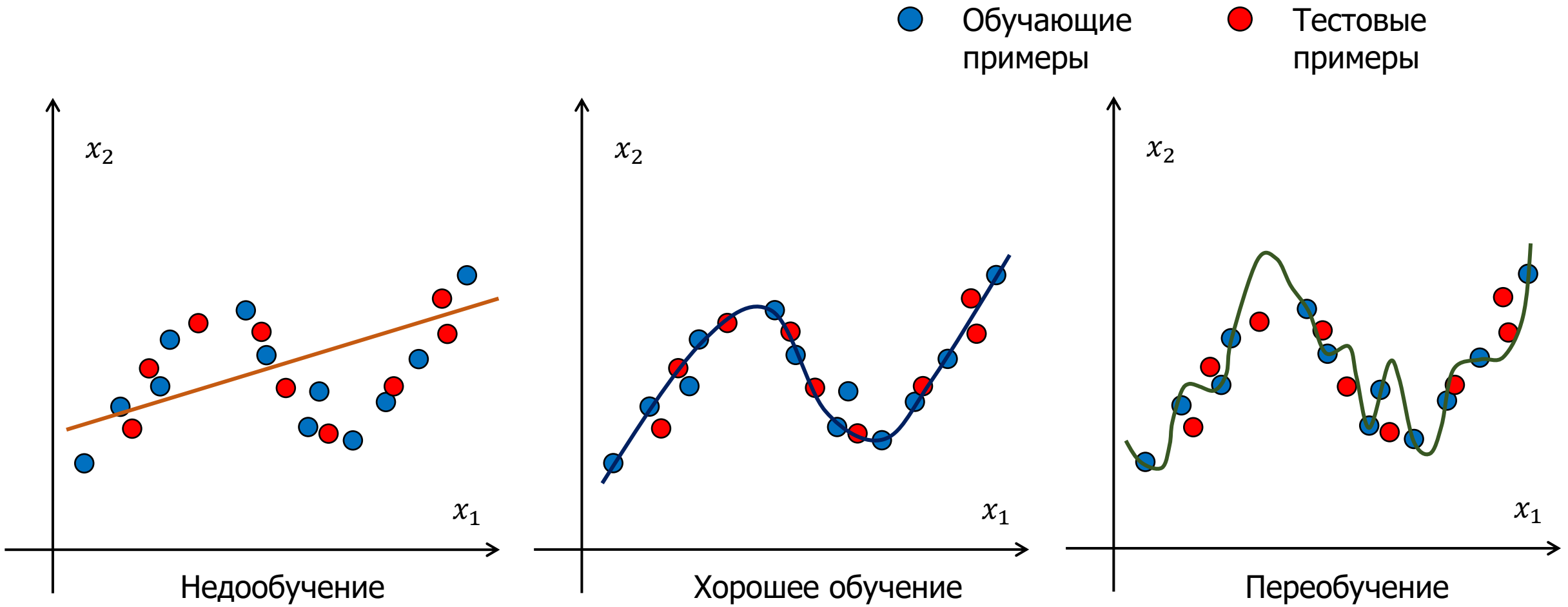
Сергей В. Аксёнов,

к.т.н., доцент кафедры теоретических основ информатики,

Томский государственный университет

Томск-2023

Примеры регрессионных моделей



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: $RAE = \frac{\sum_{i=1}^{n}|y_i - \widetilde{y_i}|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$

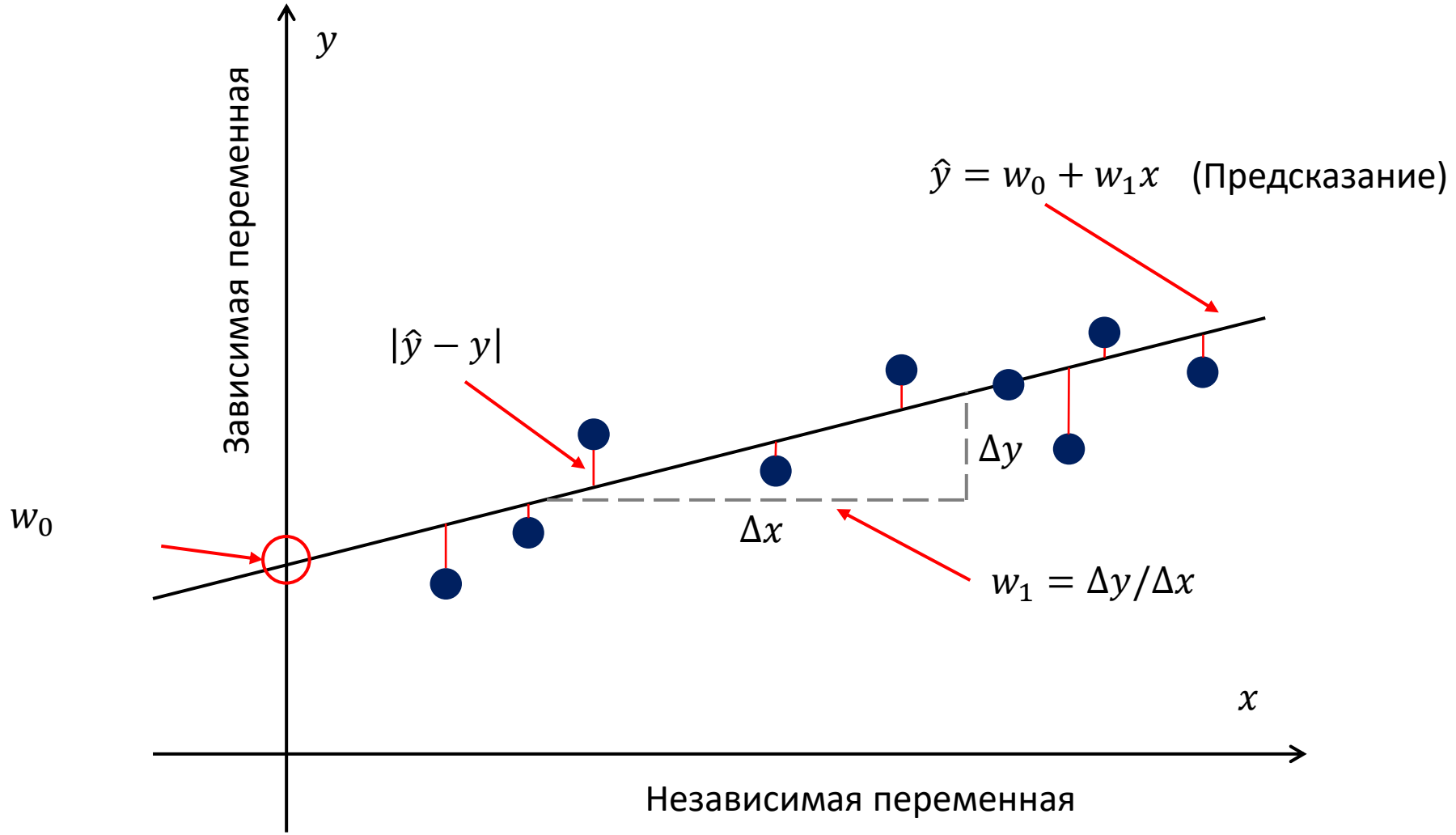
6. Коэффициент детерминации: $R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widetilde{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

y_i - Истинные значения

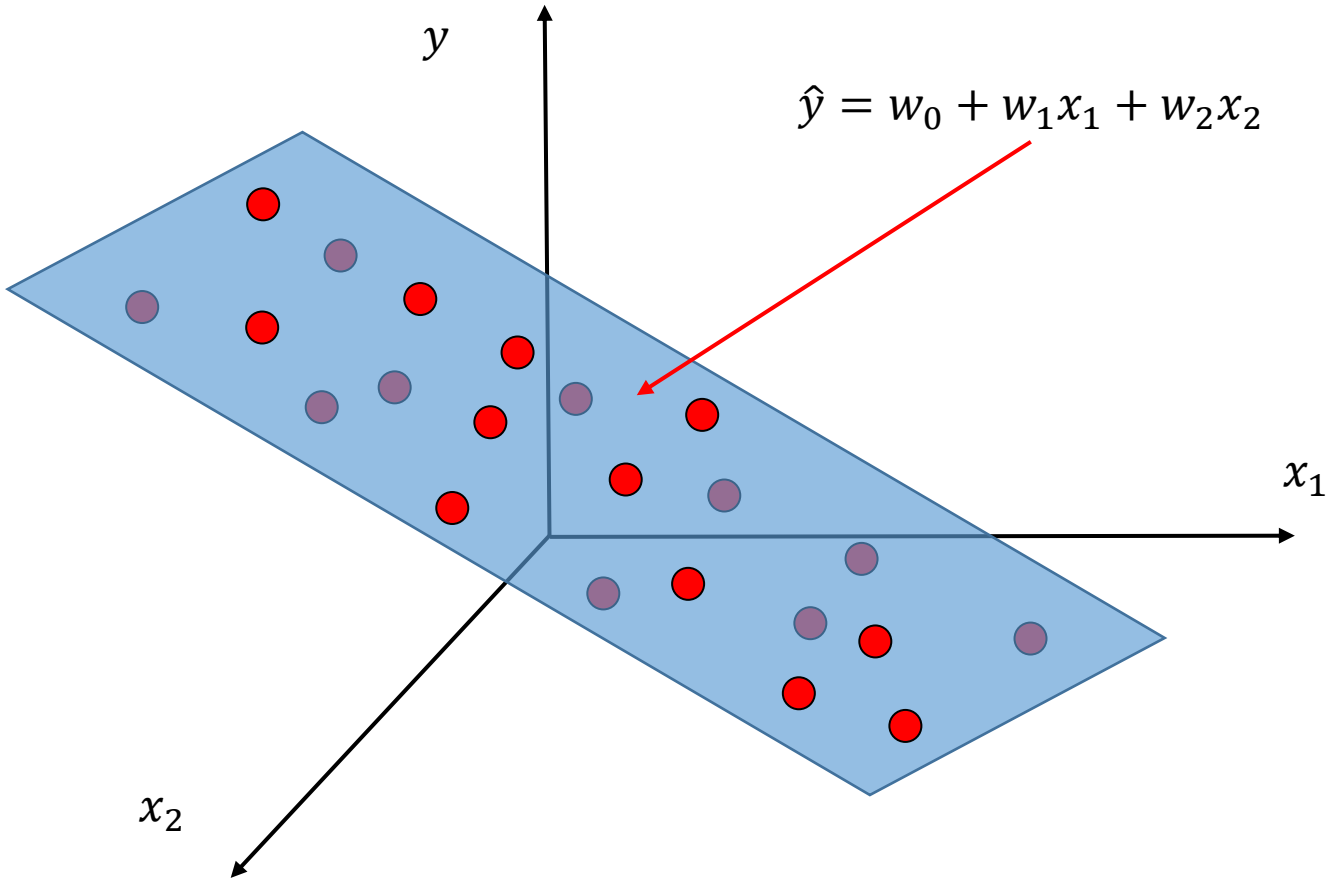
$\widetilde{y_i}$ - Предсказанное значение

$\bar{y}$ - Среднее значение

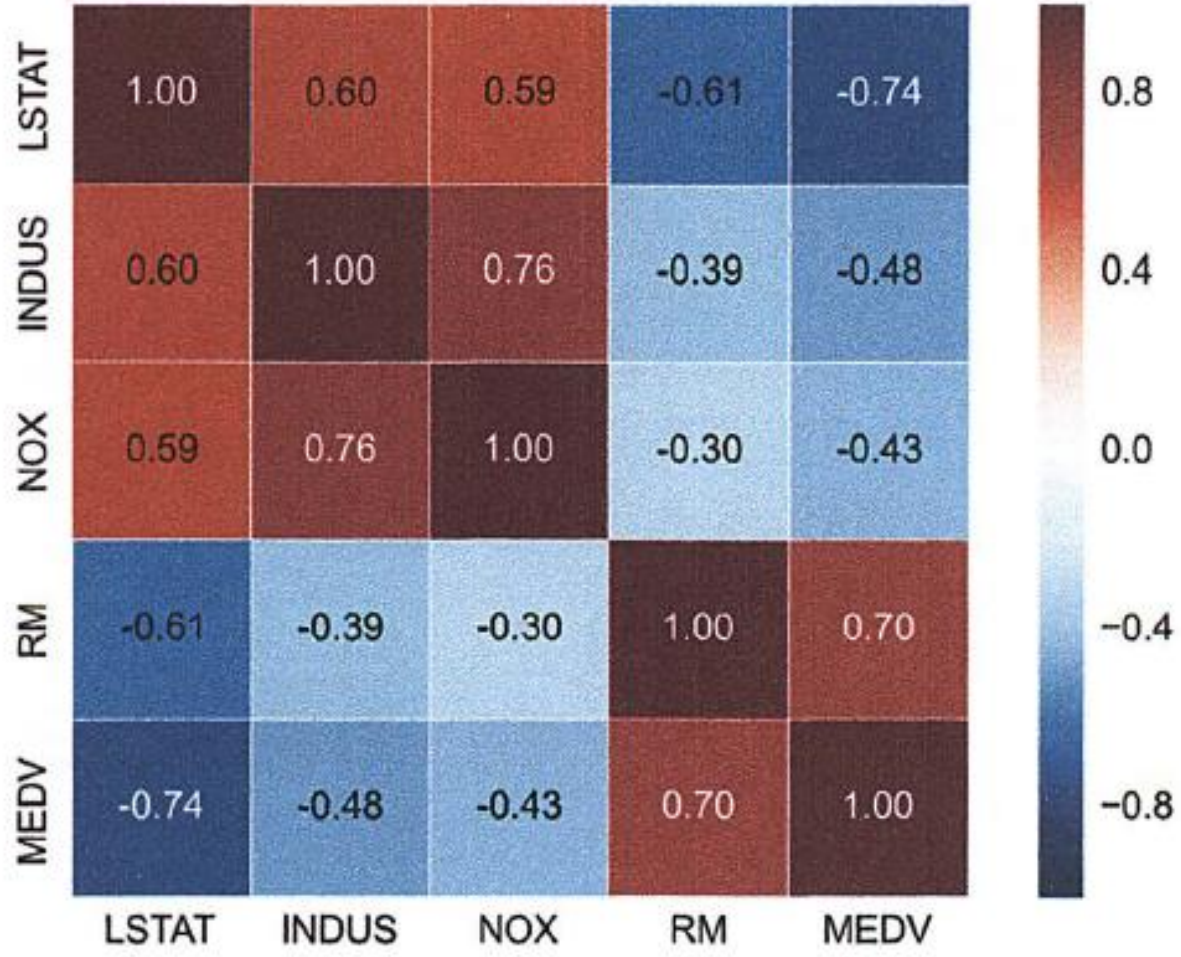
Линейная регрессия: один признак



Линейная регрессия: два признака



Пример тепловой карты и зависимости признаков



Регуляризация в регрессионных моделей

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:

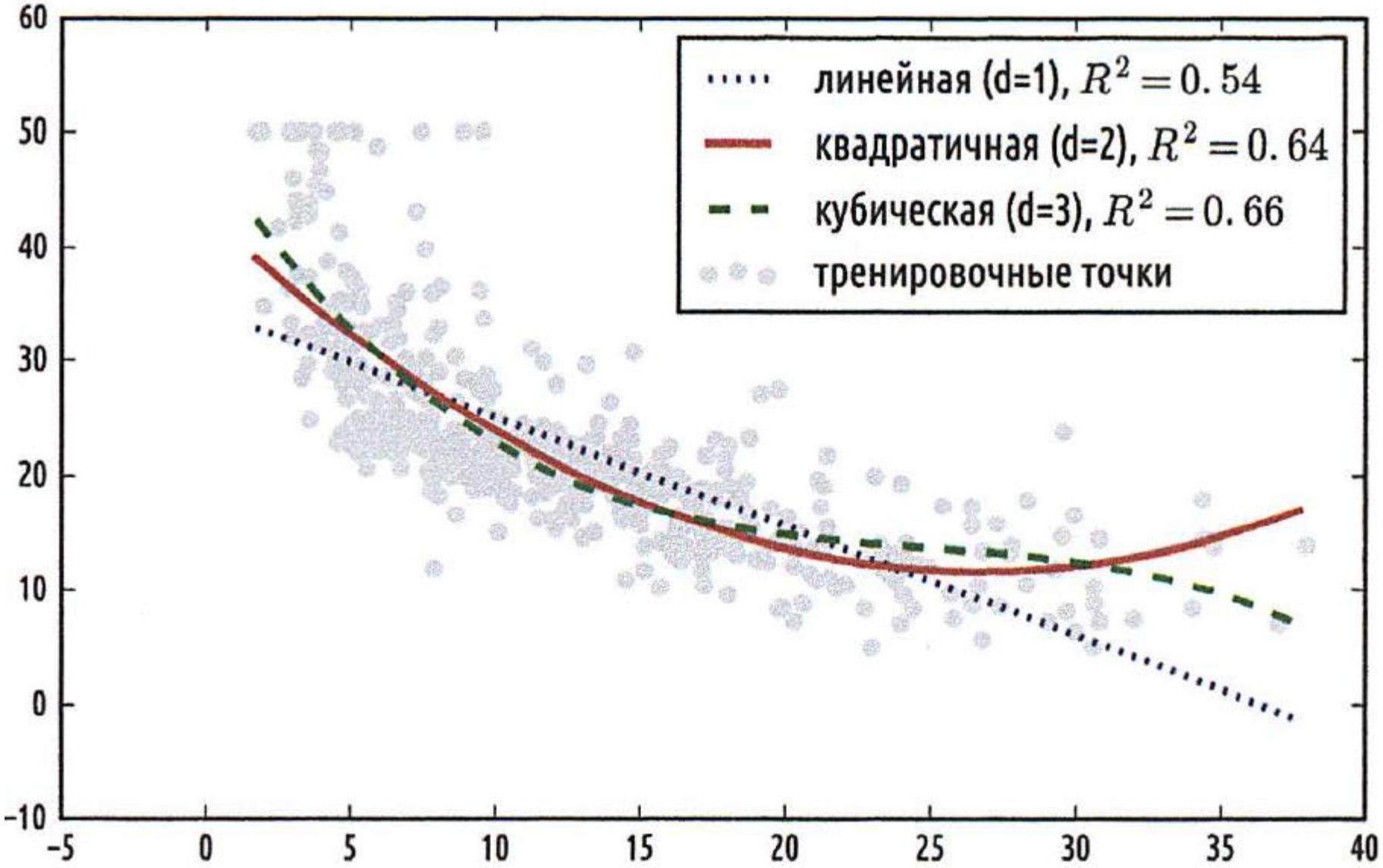
$$J(w)_{Elastic\_Net} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2$$

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



Регрессия с помощью дерева и случайного леса

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:

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

Пример регрессии с помощью дерева