Южный федеральный университет

@SINECOR

Линейная регрессия

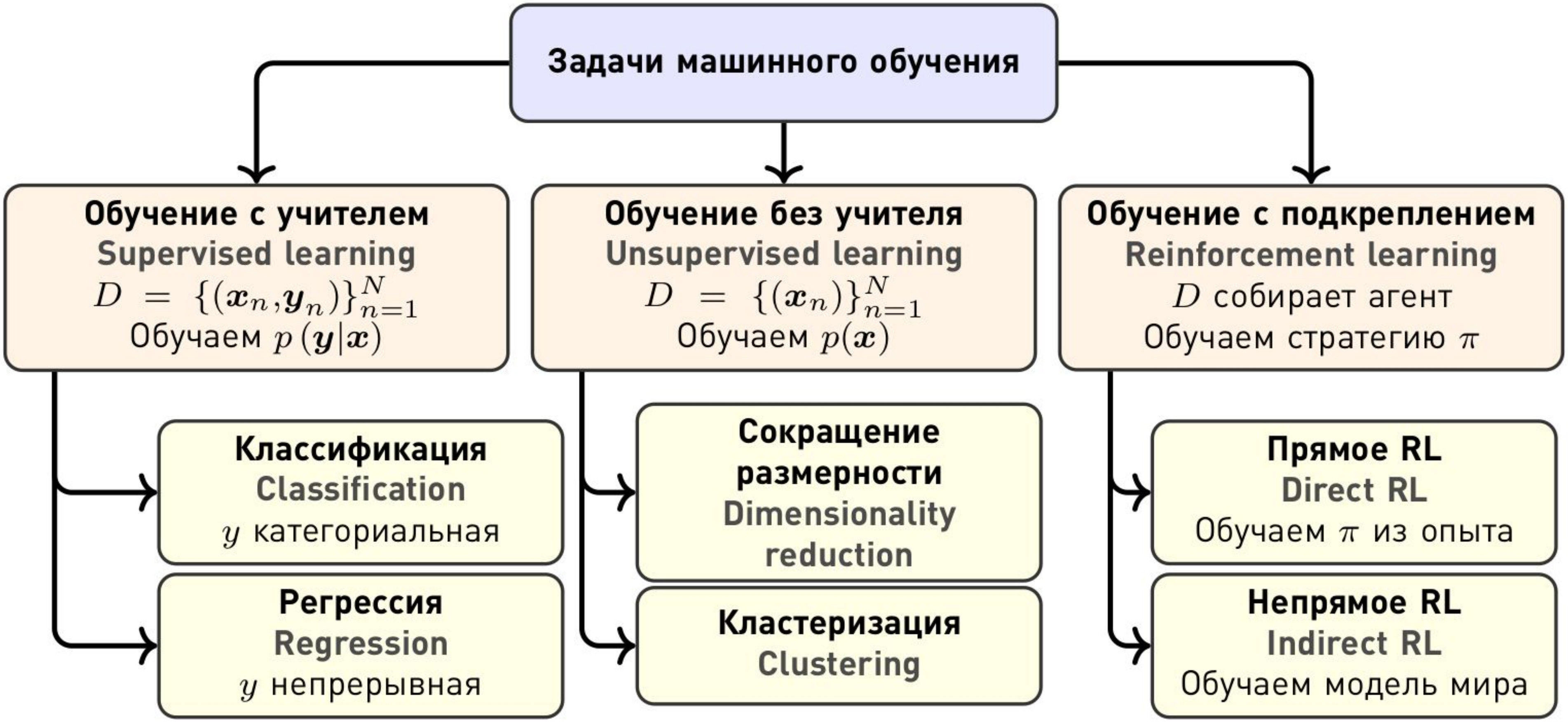
Сергей Николенко
13 ноября 2025 г.

Технологии и фронтиры науки о данных

Метод наименьших квадратов

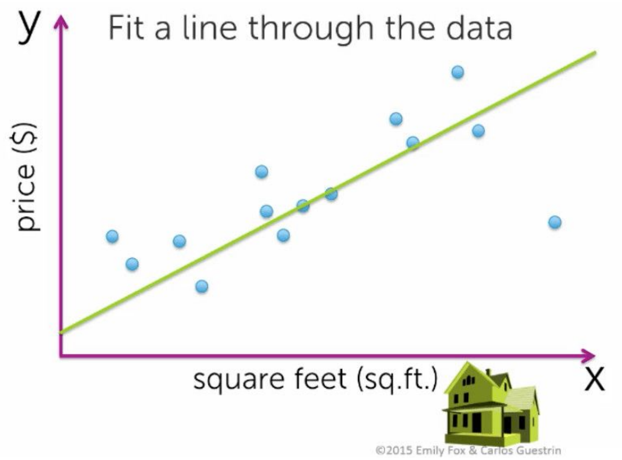
Виды постановок задач

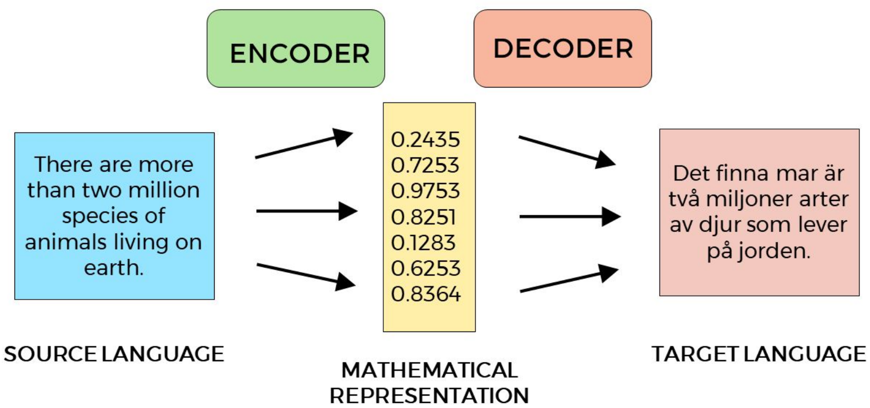
- На этой классификации давайте остановимся поподробнее…

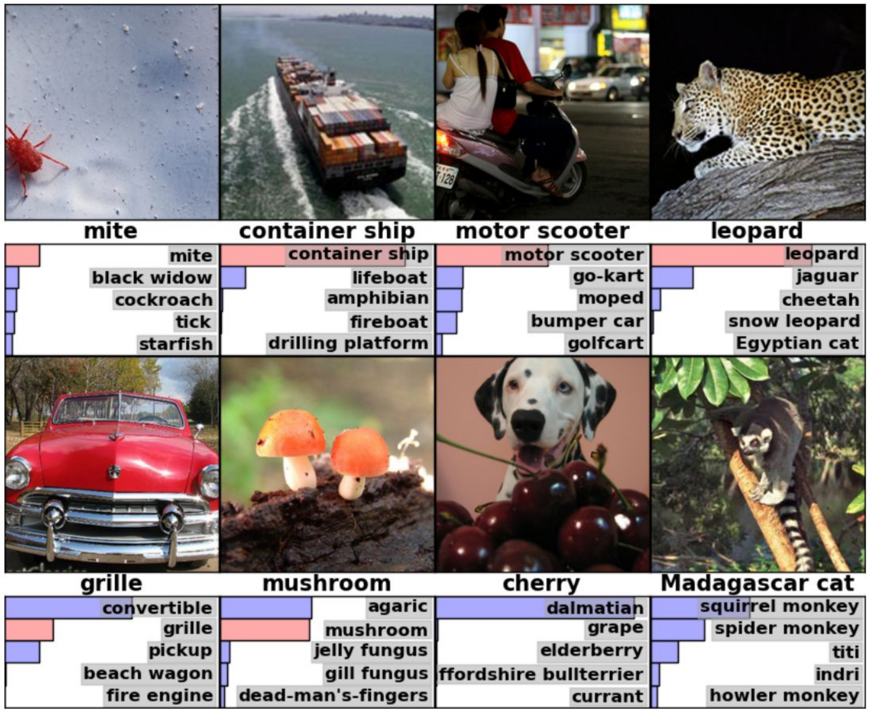


Обучение с учителем

- Примеры задач обучения с учителем:

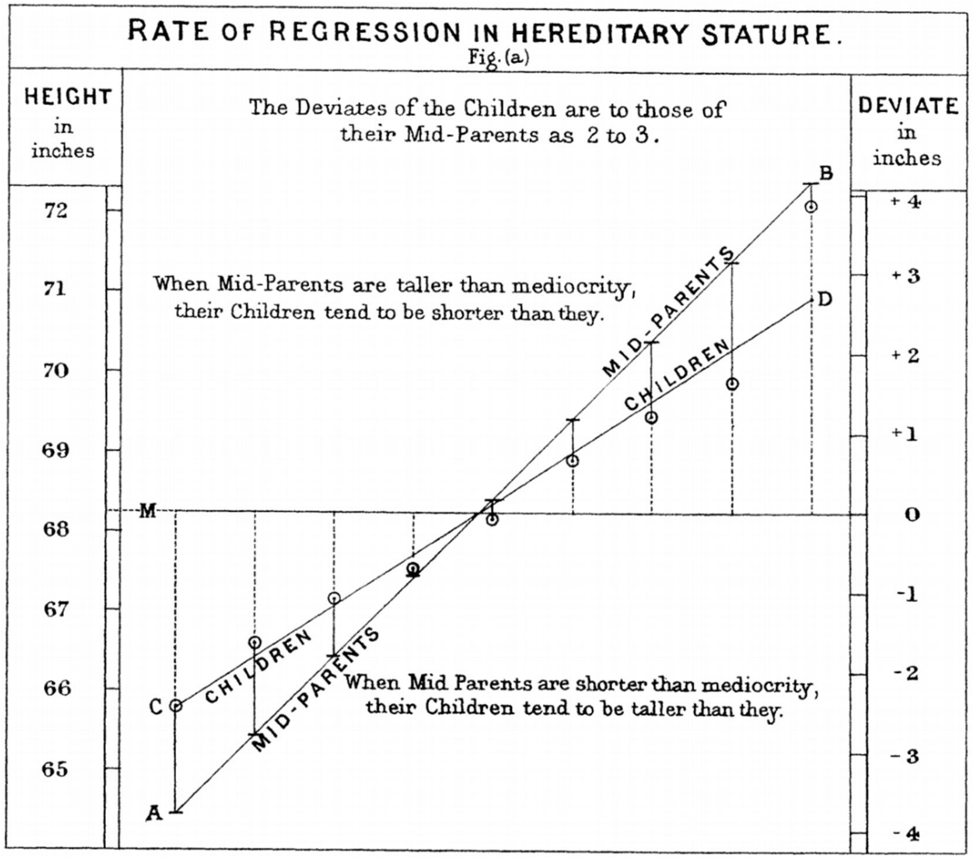


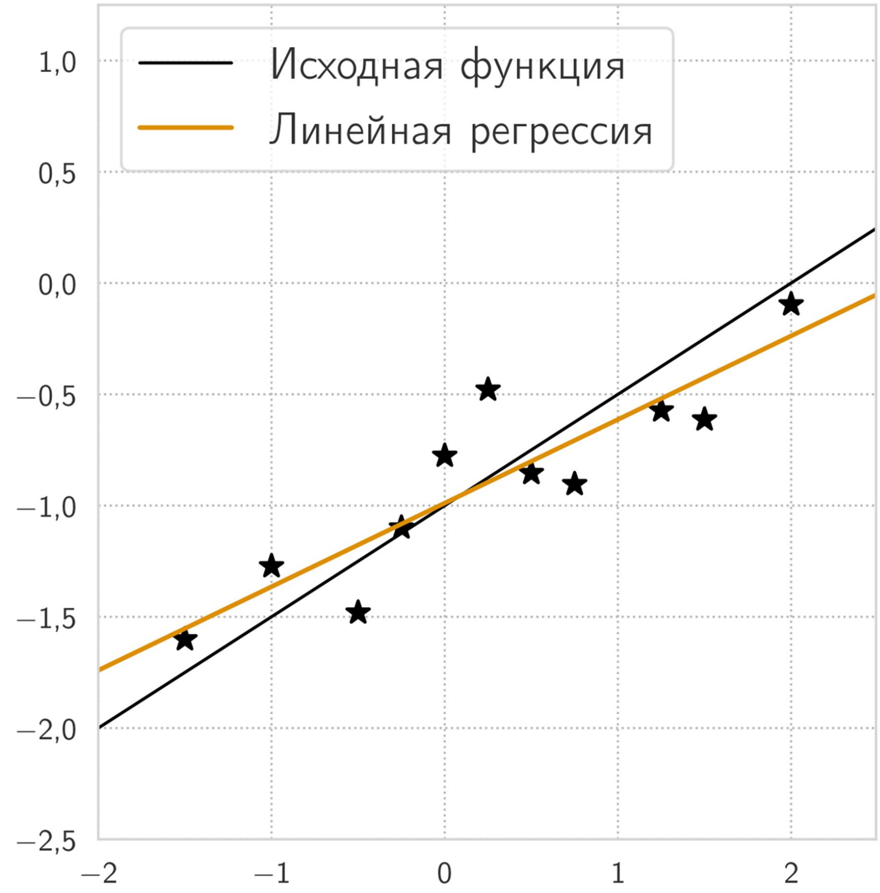




Кстати, почему “регрессия”?

- Почему регрессия называется регрессией? Странное название...





Линейная регрессия

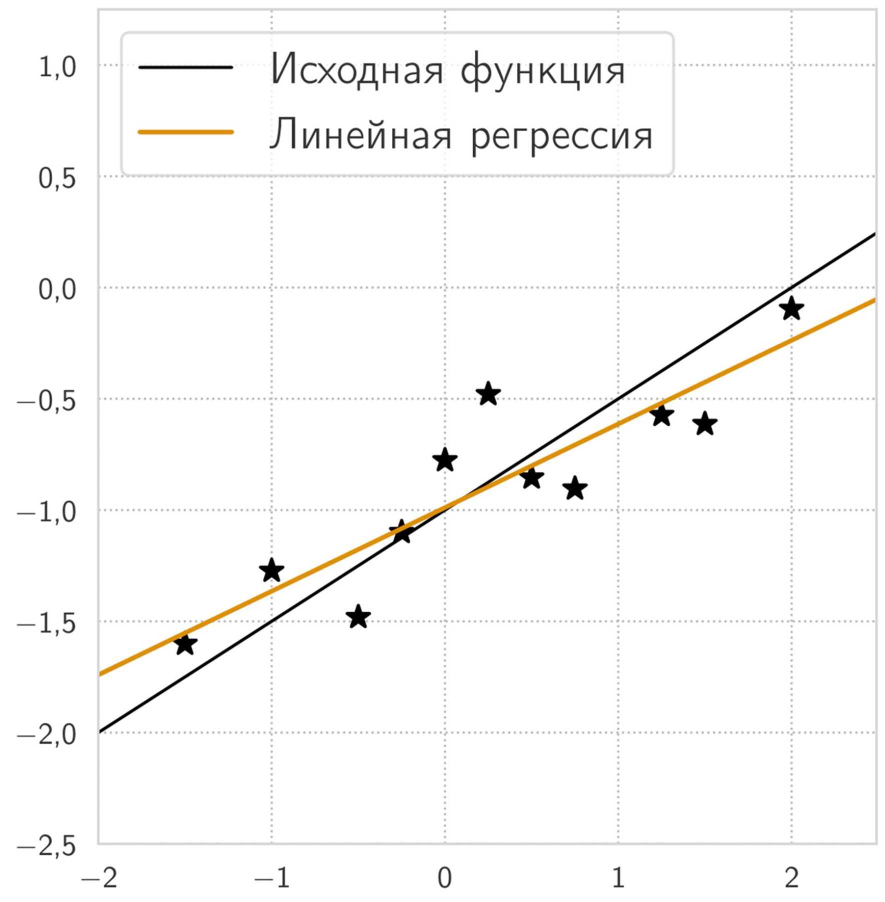
- Рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathbf{x}^\top \mathbf{w}$$

- По вектору входов **x** мы будем предсказывать выход *y* как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

- Как найти оптимальные параметры **w** по данным?



Линейная регрессия

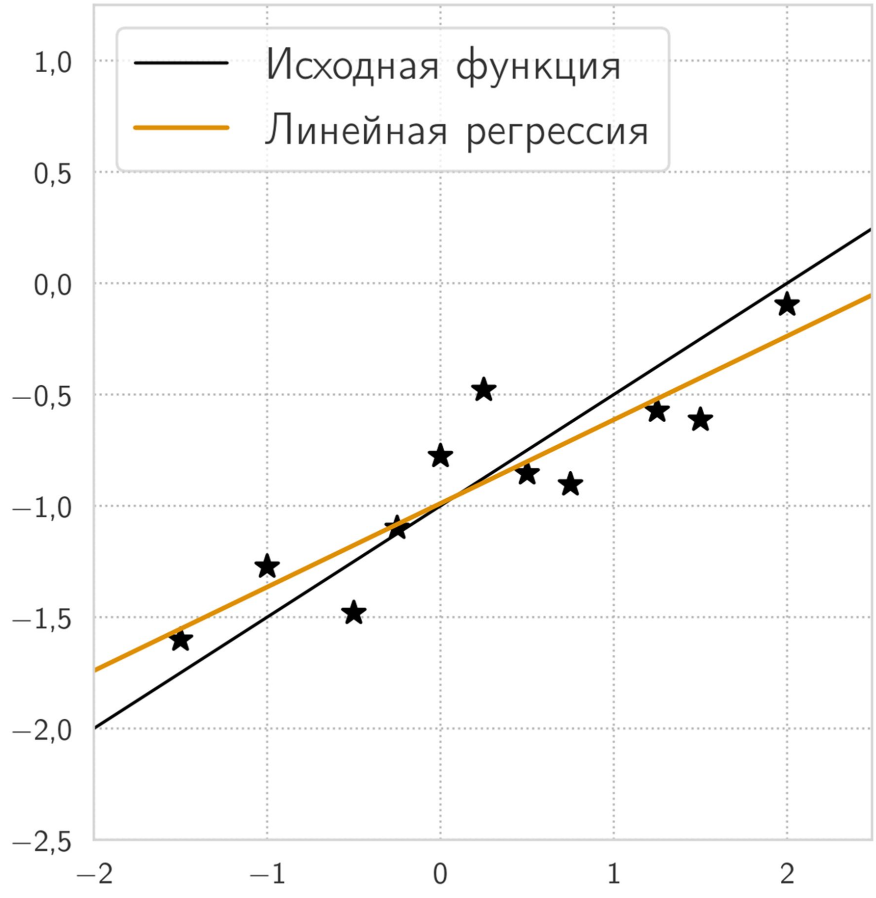
- Рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathbf{x}^\top \mathbf{w}$$

- По вектору входов **x** мы будем предсказывать выход *y* как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

- Как найти оптимальные параметры **w** по данным?
- Можно минимизировать сумму квадратов отклонений...



Линейная регрессия

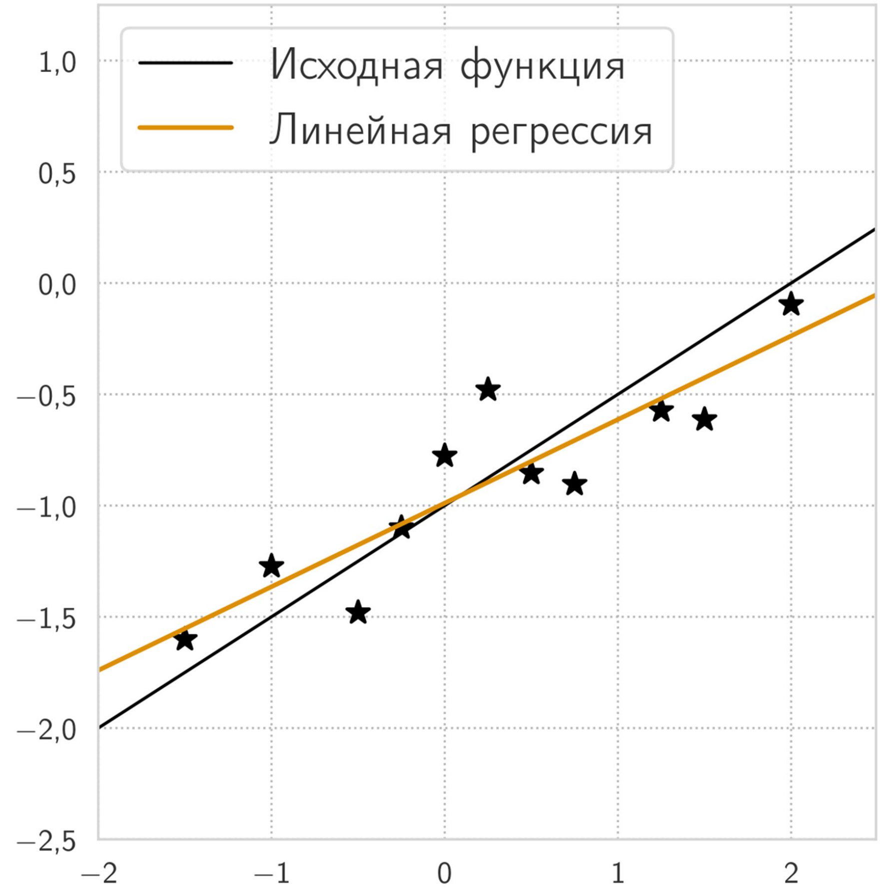
- Можно на самом деле решить задачу точно – записать как

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

продифференцировать, приравнять нулю, и получится

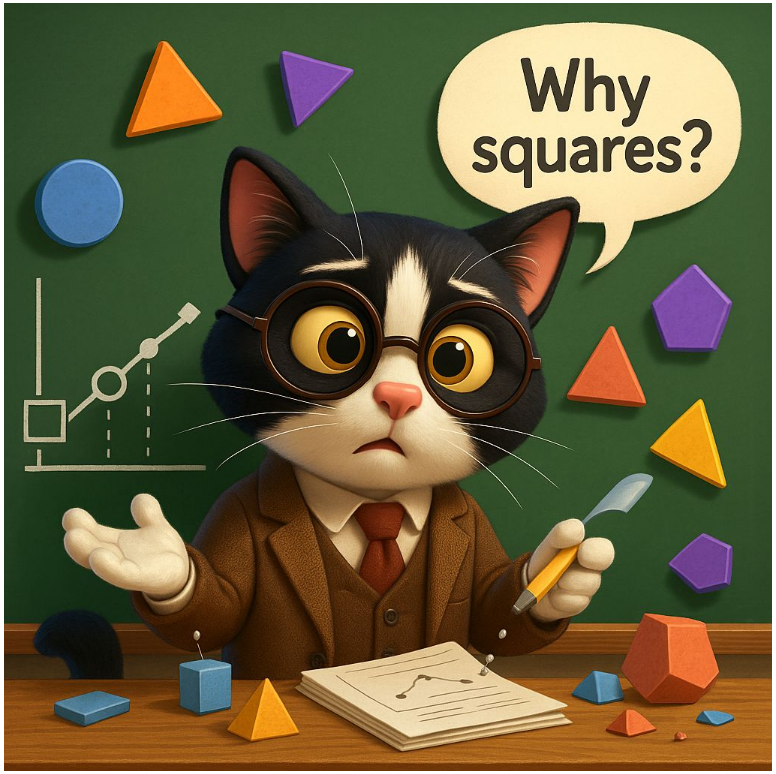
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- Это псевдообратная Мура-Пенроуза (Moore-Penrose pseudoinverse)



Но почему квадрат?

- Но откуда берётся метод наименьших квадратов?
- Почему квадрат?



Но почему квадрат?

- Но откуда берётся метод наименьших квадратов?
- Почему квадрат?
- Основное наше предположение – в том, что шум (ошибка в данных) распределён нормально, т.е.

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

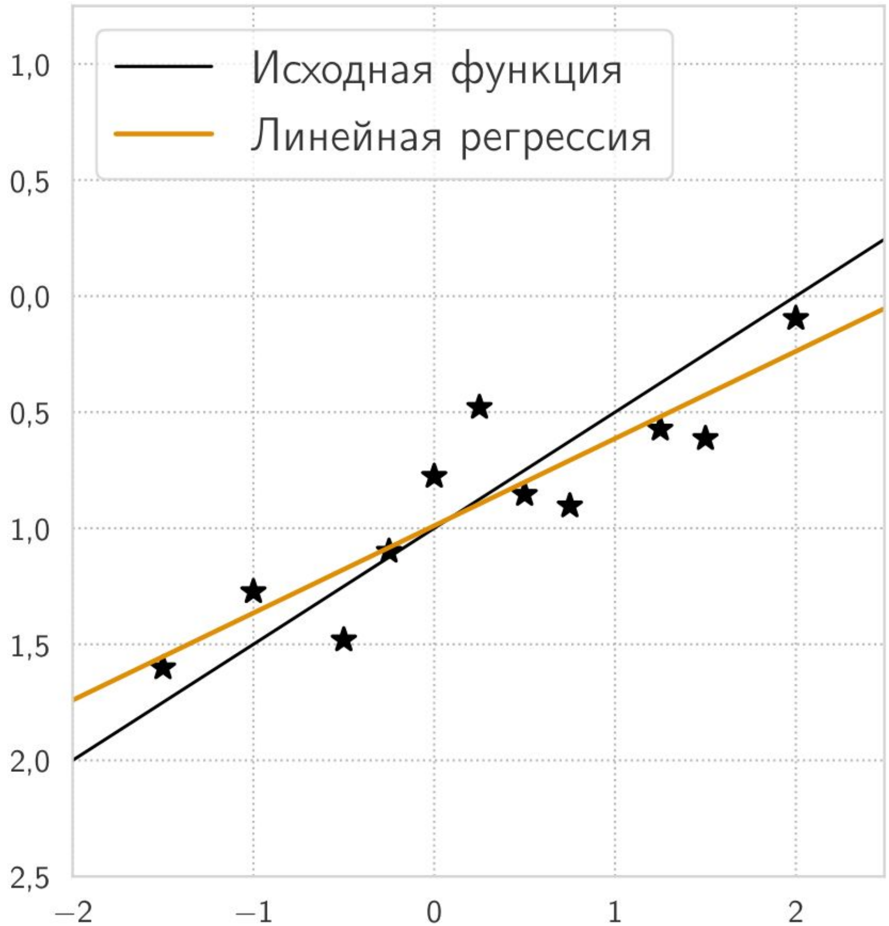
- Иначе говоря,

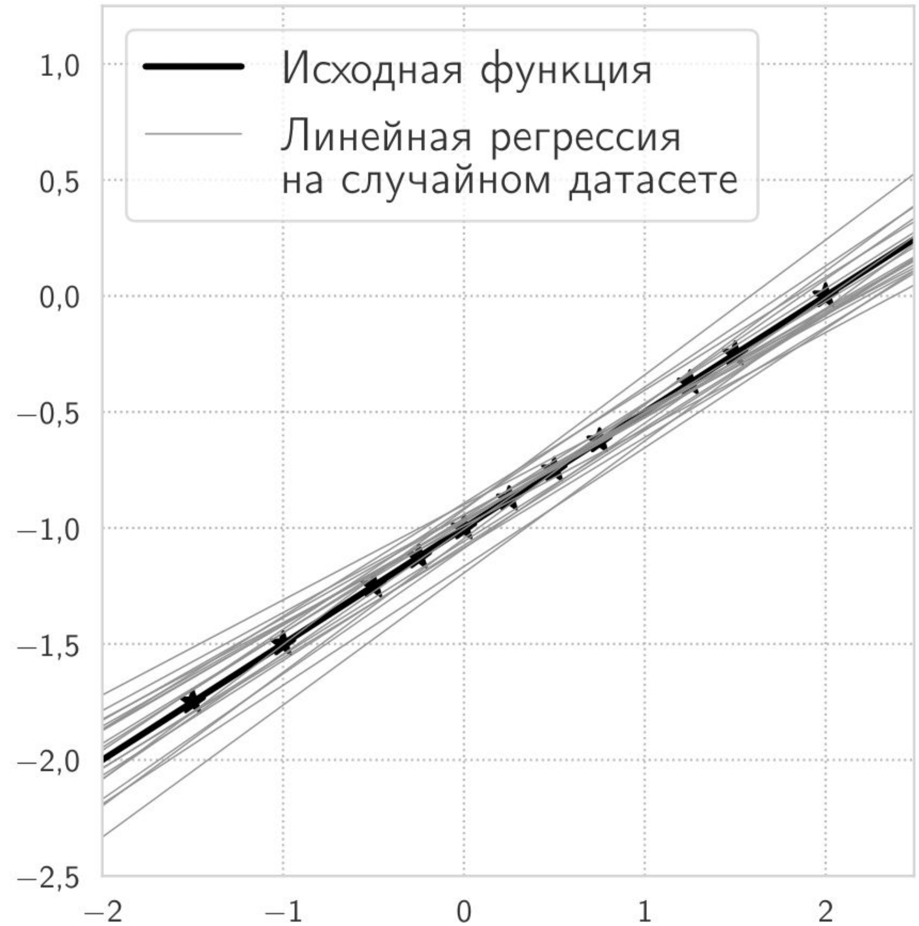
$$p(t \mid \mathbf{x}, \mathbf{w}, \sigma^2) = N(t \mid y(\mathbf{x}, \mathbf{w}), \sigma^2).$$

- И тогда...

Линейная регрессия по-байесовски

- Если набрасывать датасеты согласно нашим предположениям, то в среднем будет получаться то, что надо — но именно в среднем!

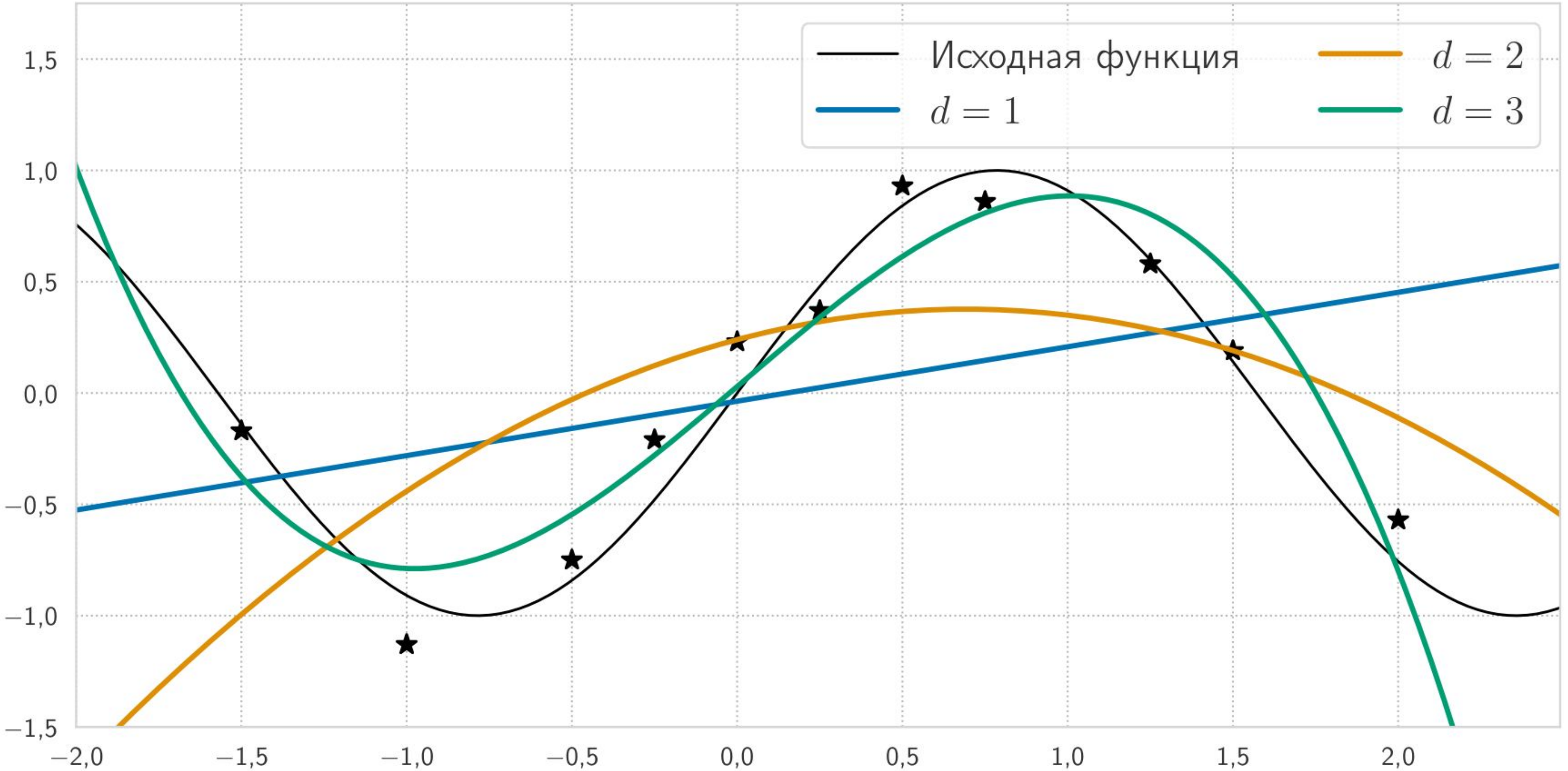




Признаки, оверфиттинг, регуляризация

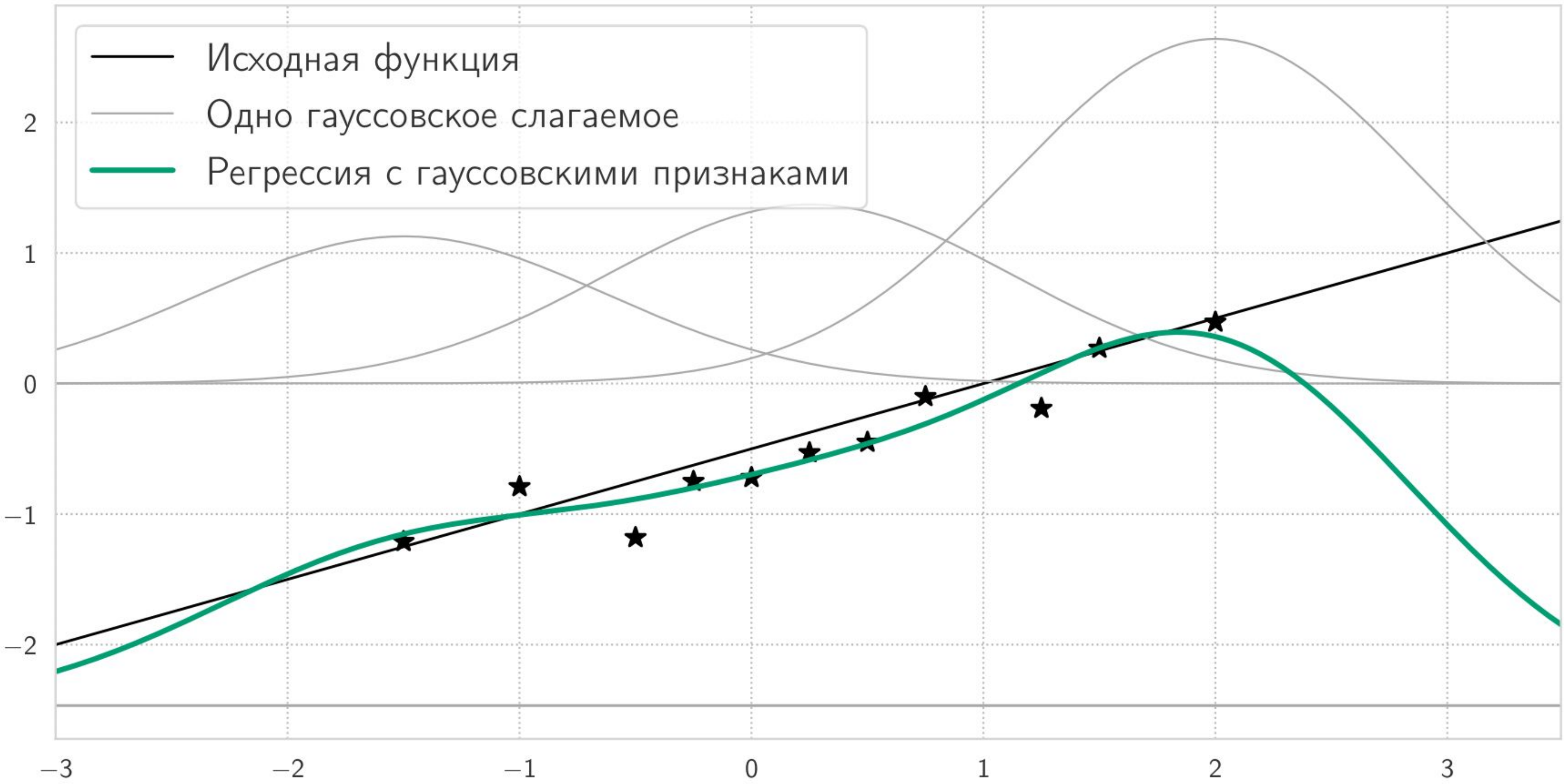
Линейная регрессия с признаками

- К линейной регрессии легко добавить какие угодно признаки:



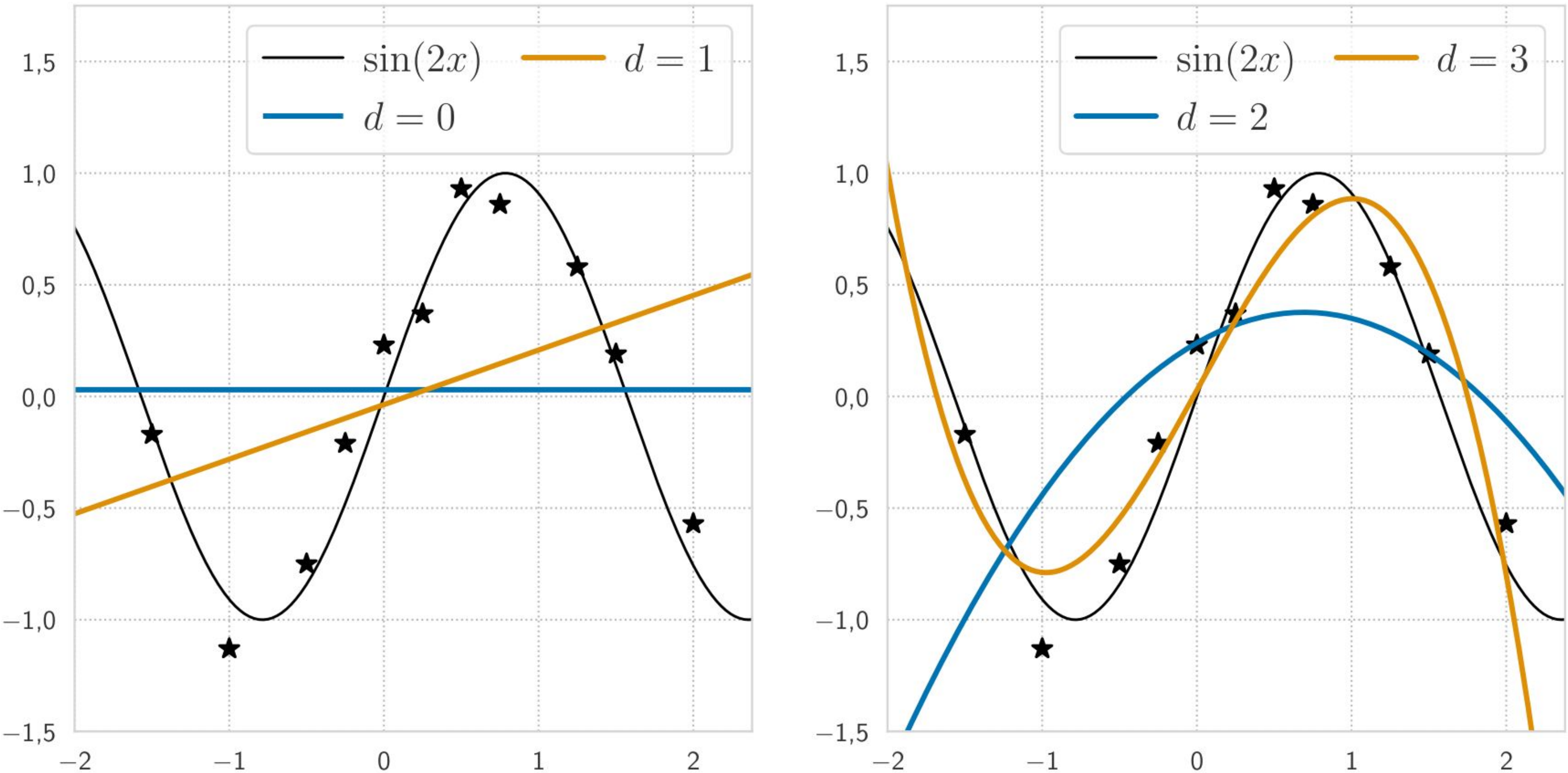
Линейная регрессия с признаками

- К линейной регрессии легко добавить какие угодно признаки:



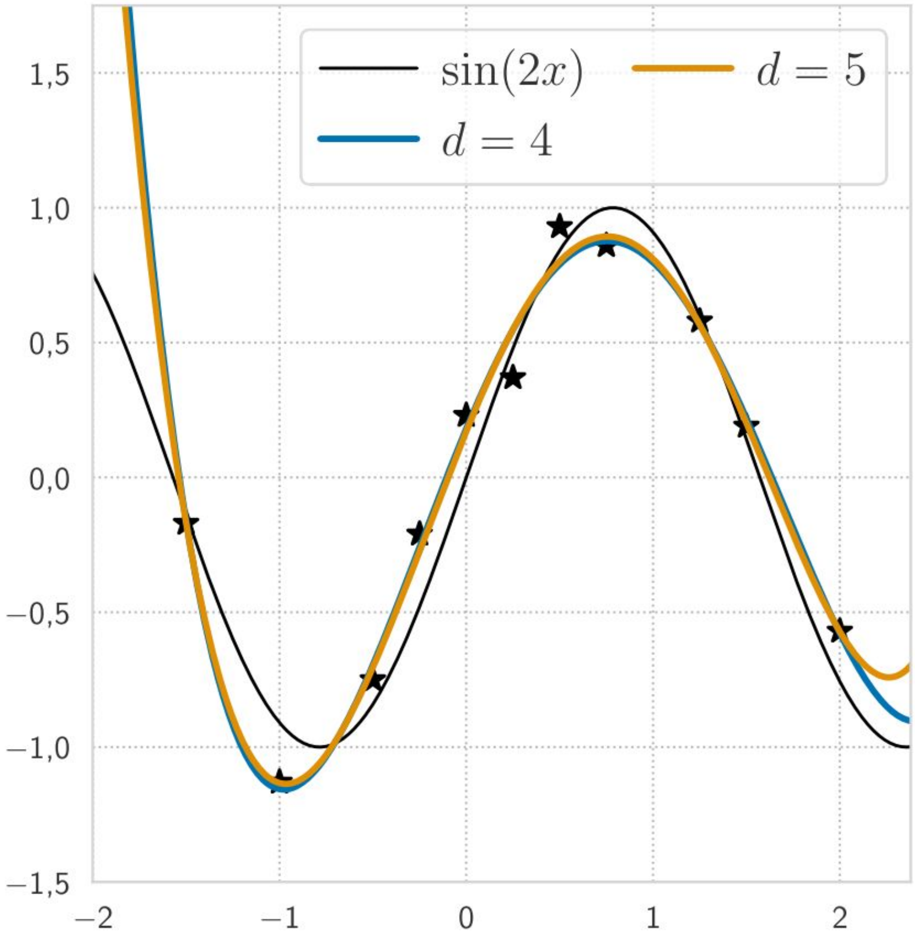
Линейная регрессия с признаками

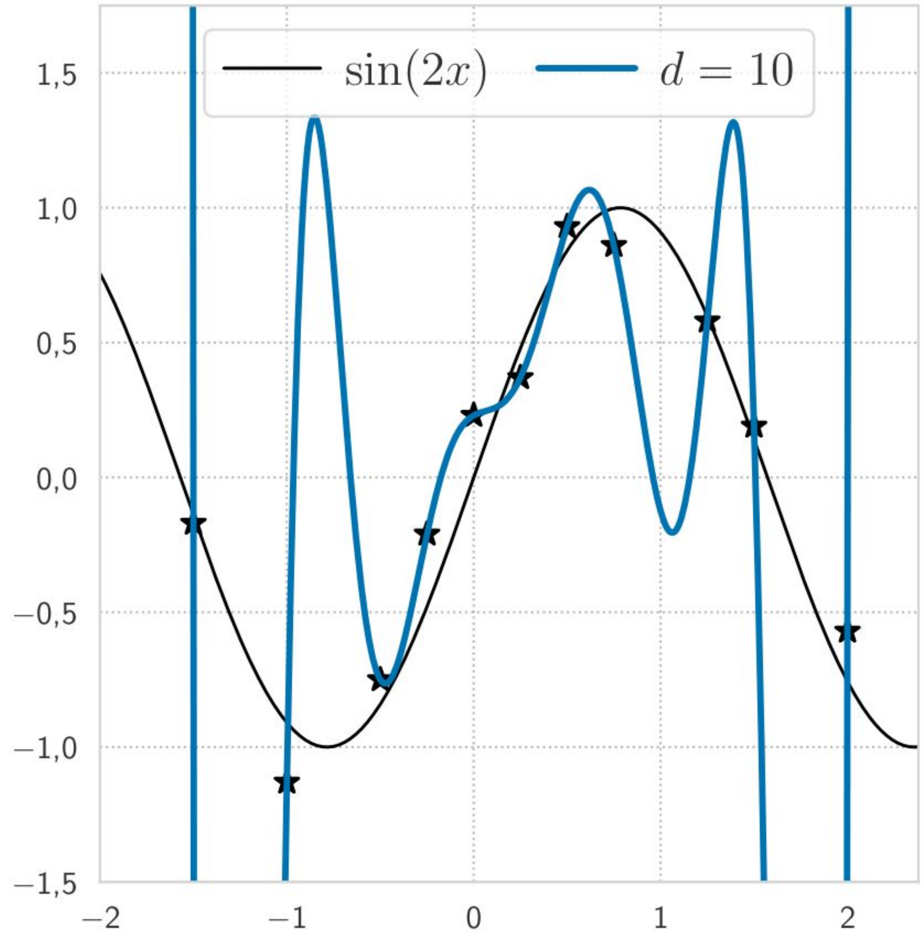
- Но новые признаки — это хорошо только до какого-то момента:



Линейная регрессия с признаками

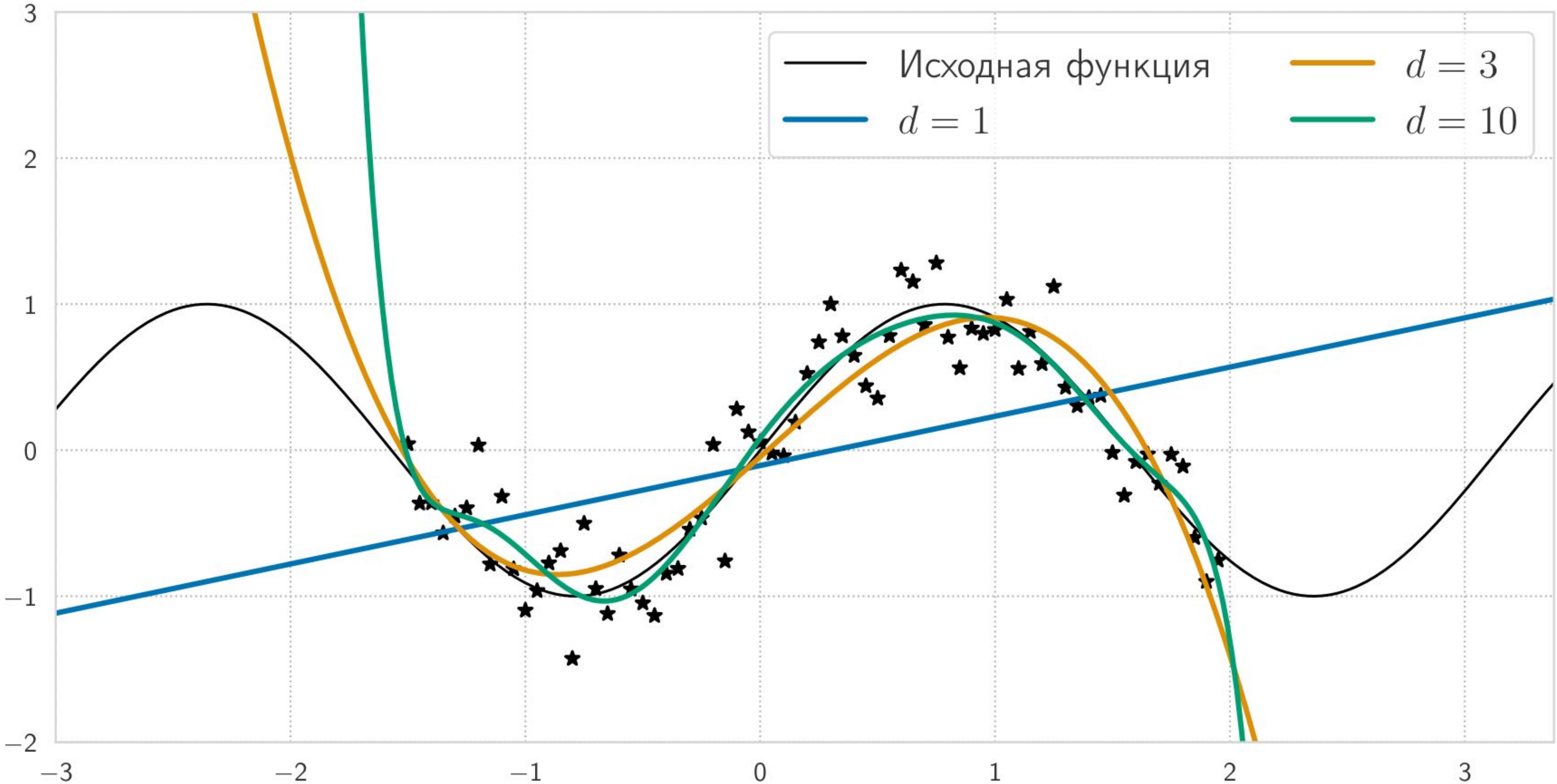
- А потом может начаться жёсткий оверфиттинг:





Линейная регрессия с признаками

- Это, конечно, зависит от количества имеющихся данных:



Южный федеральный университет

@SINECOR

Спасибо за внимание!