Проклятие размерности. Регуляризация

Квантитативный анализ текста

Кирилл Александрович Маслинский

04.04.2022 / 06

НИУ ВШЭ Санкт-Петербург

Проклятие размерности

Sparse data problem

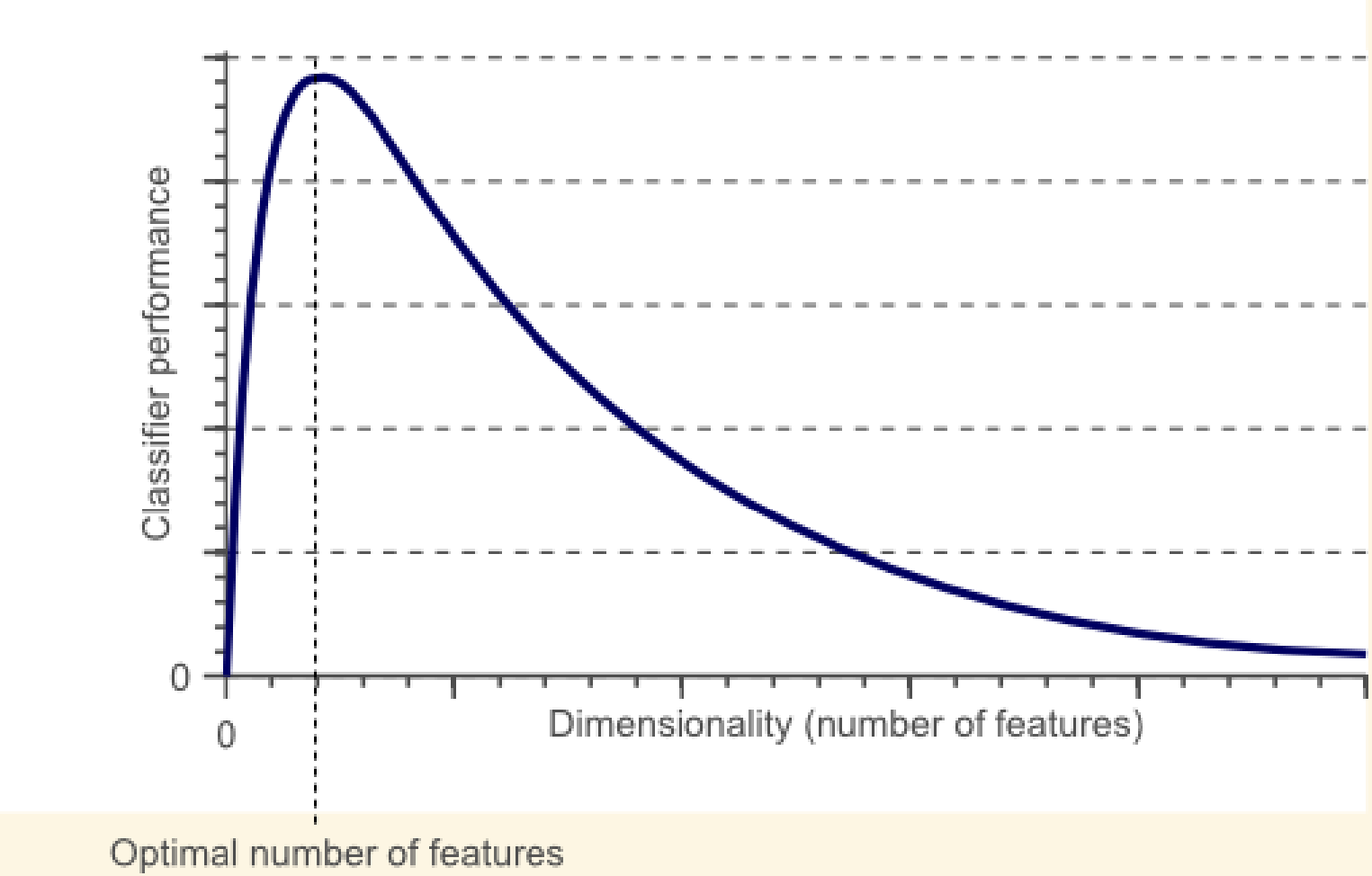
	Terms				
Docs	выгребать	выгребной	выгружать	выгрузка	выгрыза
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	
5	0	0	0	0	

Проклятие размерности

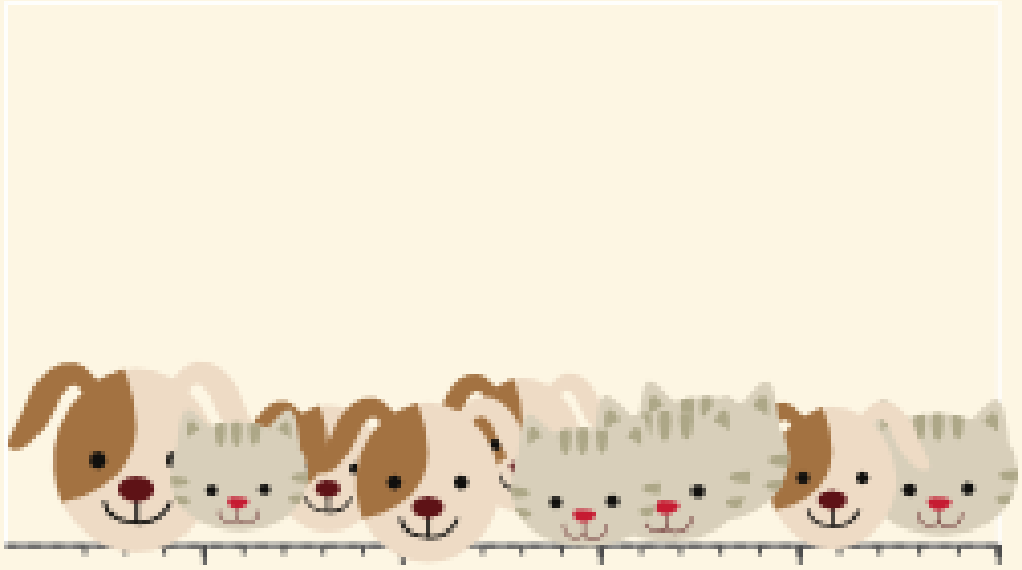
```
A document-term matrix (1530 documents, 13322 terms

Non-/sparse entries: 68859/20313801
Sparsity           : 100%
Maximal term length: 66
Weighting          : term frequency (tf)
```

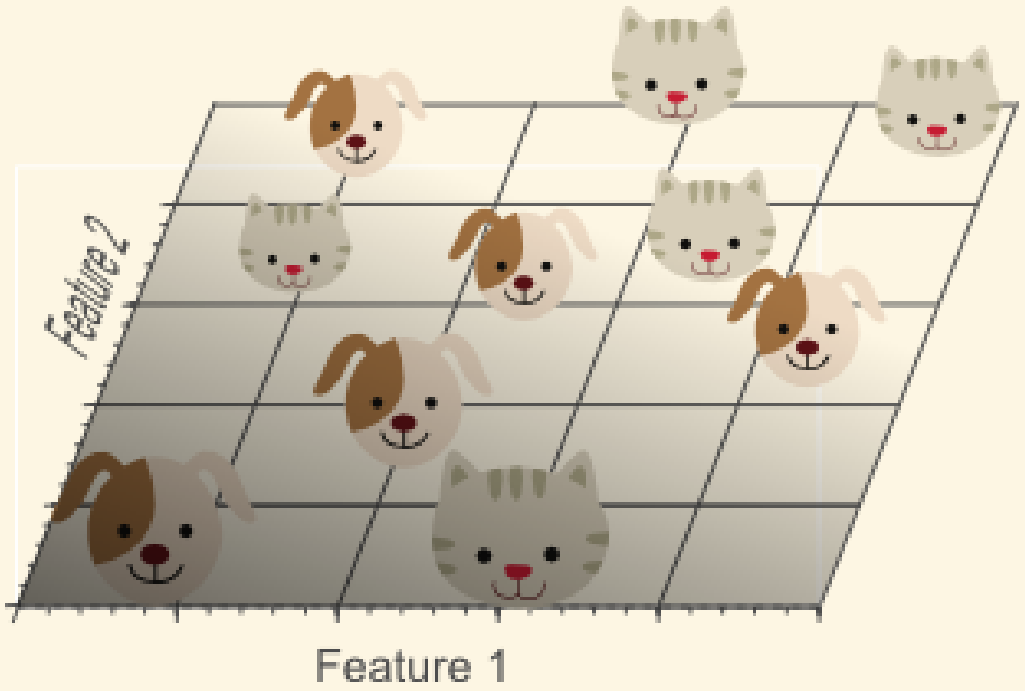
Hughes phenomenon



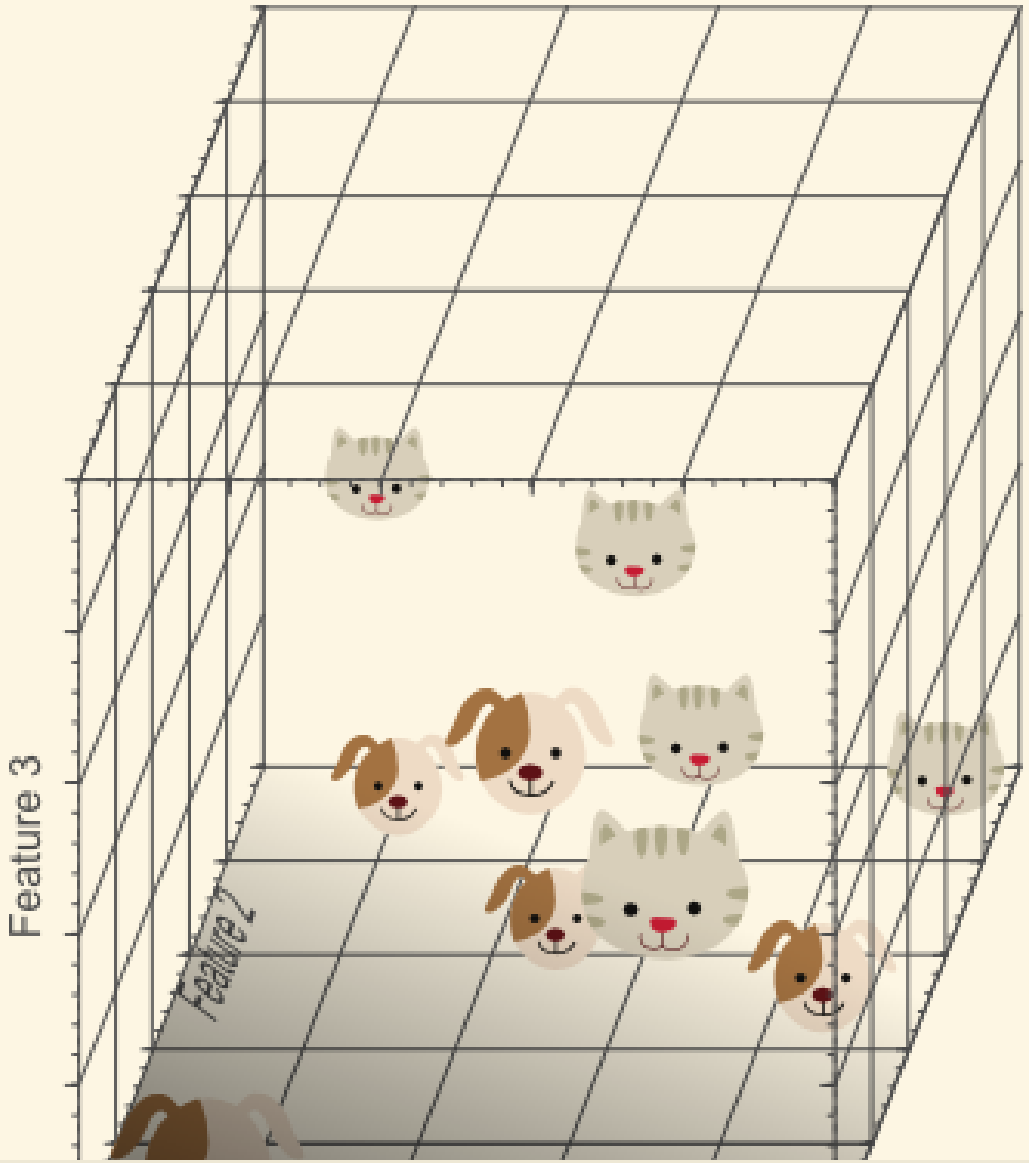
Проклятие размерности на кошках



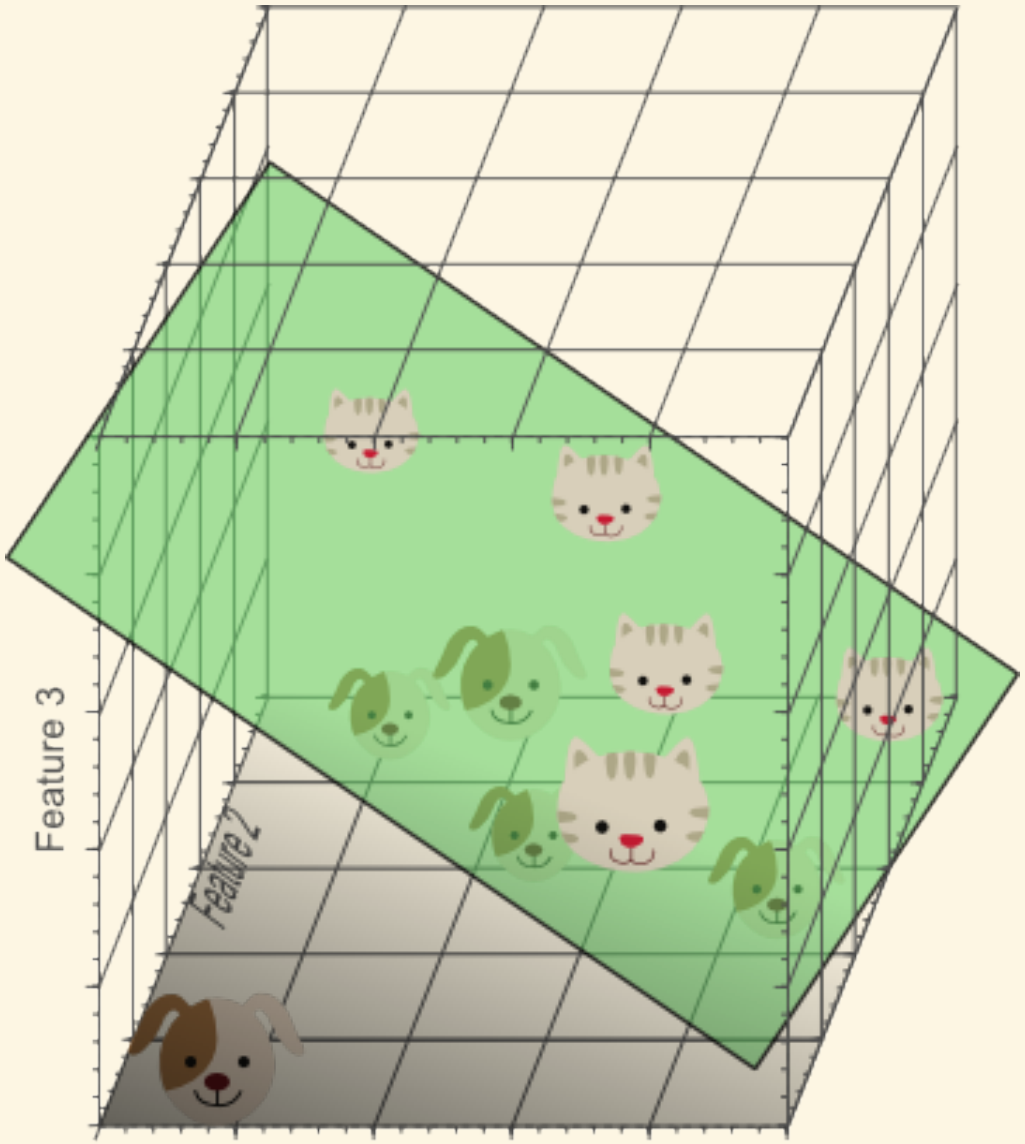
Проклятие размерности на кошках



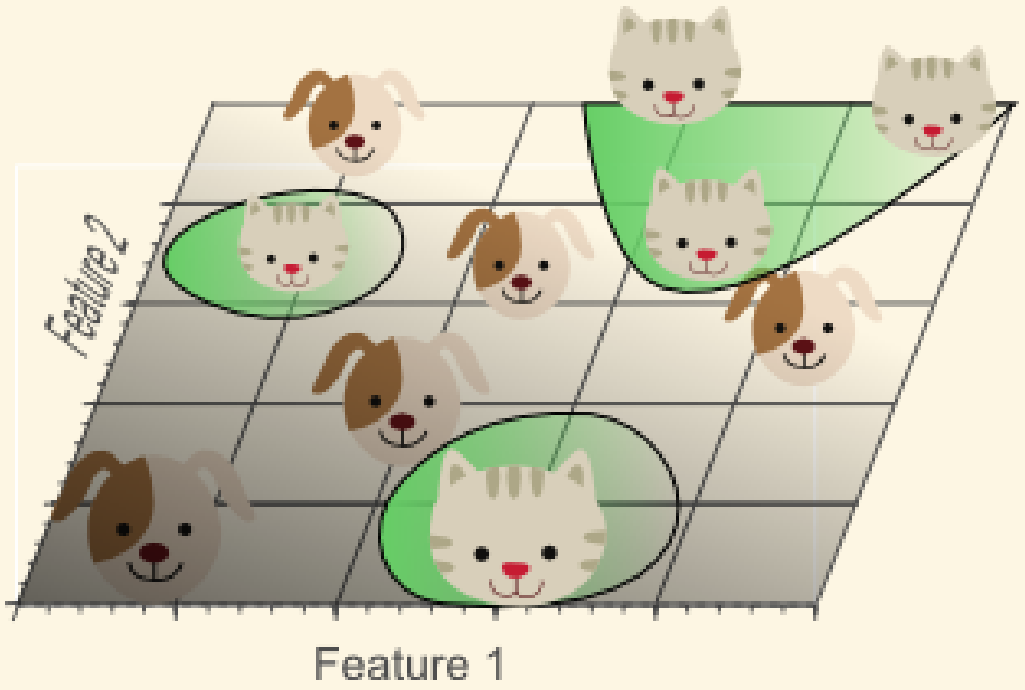
Проклятие размерности на кошках



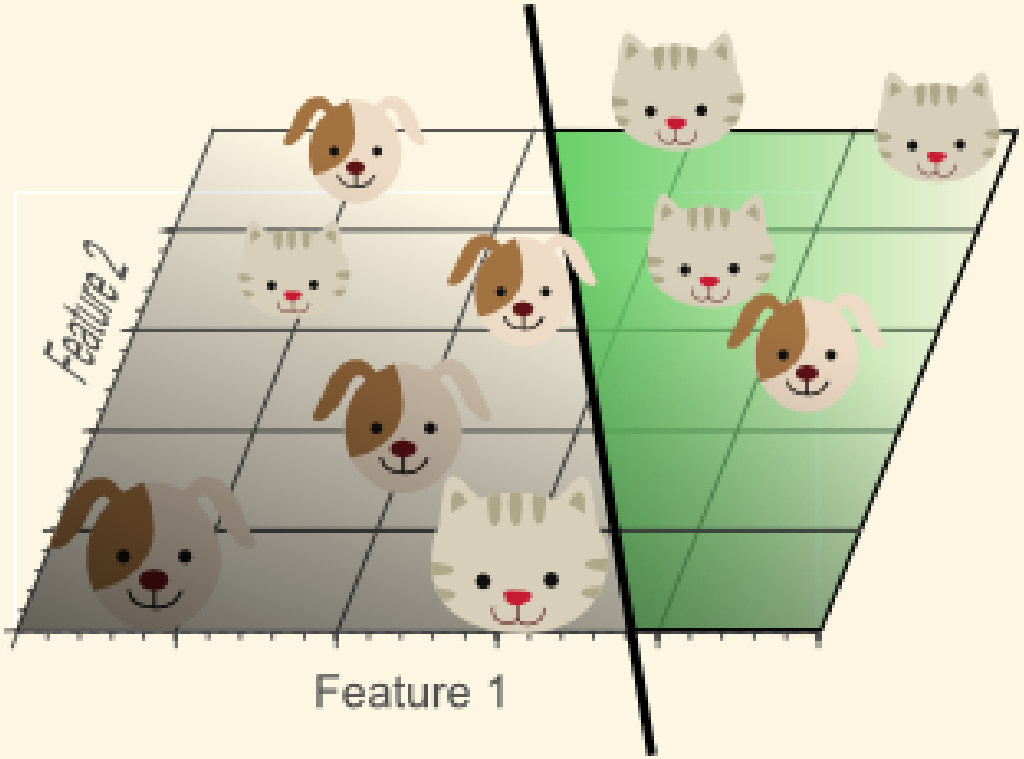
Проклятие размерности на кошках



Переобучение



Переобучение



Снижение размерности: простые способы

Снижение размерности

- Матрица терминов-документов очень большая и редкая
- Близкие по смыслу слова не обязательно встречаются в одних и тех же документах:
  - синонимия
  - полисемия
  - шум
- Нужно сократить размерность матрицы (сделать меньше столбцов).

Снижение размерности

- Матрица терминов-документов очень большая и редкая
- Близкие по смыслу слова не обязательно встречаются в одних и тех же документах:
  - синонимия
  - полисемия
  - шум
- Нужно сократить размерность матрицы (сделать меньше столбцов).

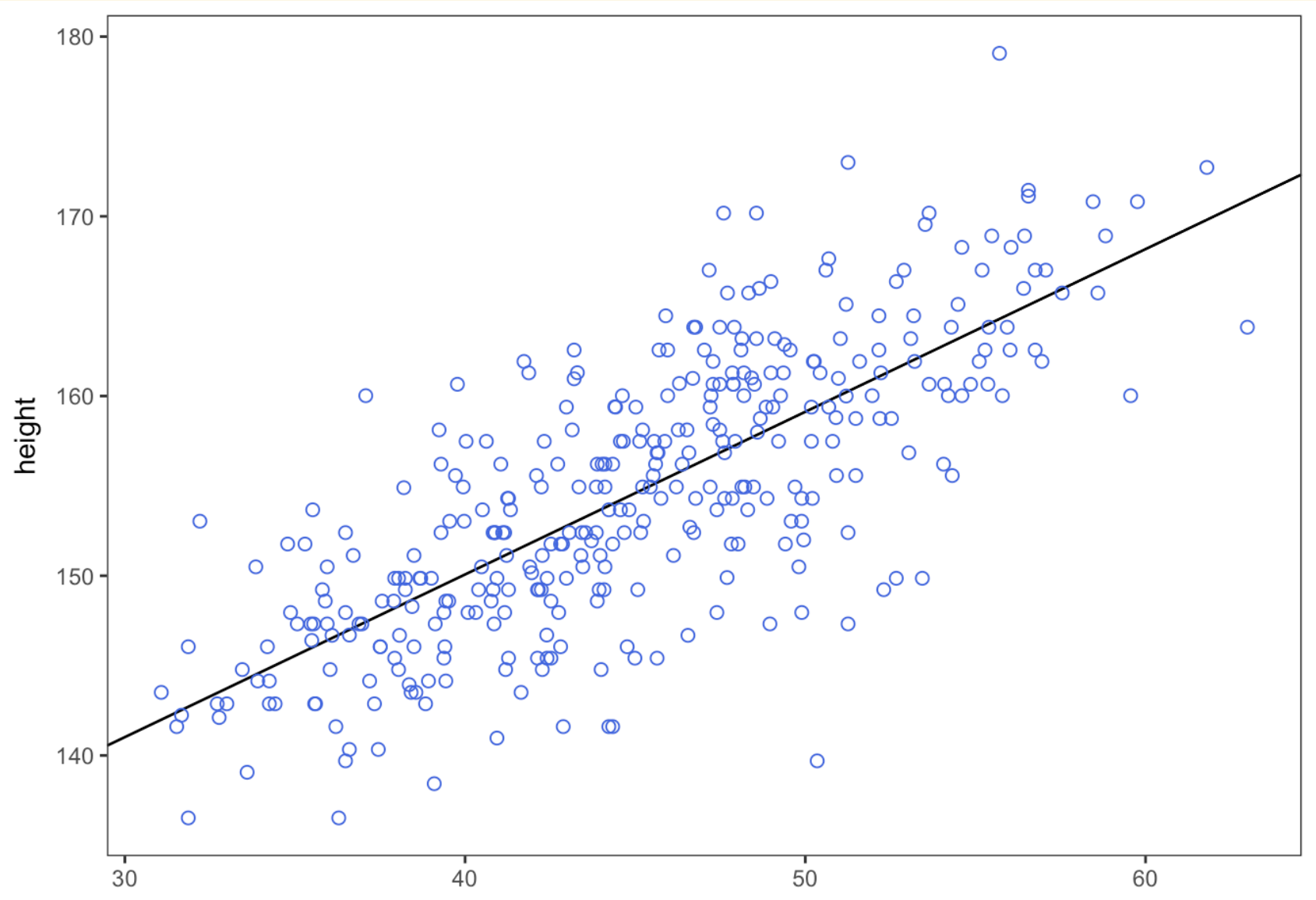
Стоп-слова

Простейший способ уменьшить число столбцов — просто удалить лишние слова:

- Статический список:
  без более бы был была были было быть в вам вас весь во вот все всего всех вы где да даже для ...
- Динамический список:
  - Слишком частотные (N самых частотных)
  - Слишком редкие (порог: не менее чем в F документов)
  - Слишком короткие (меньше M букв)

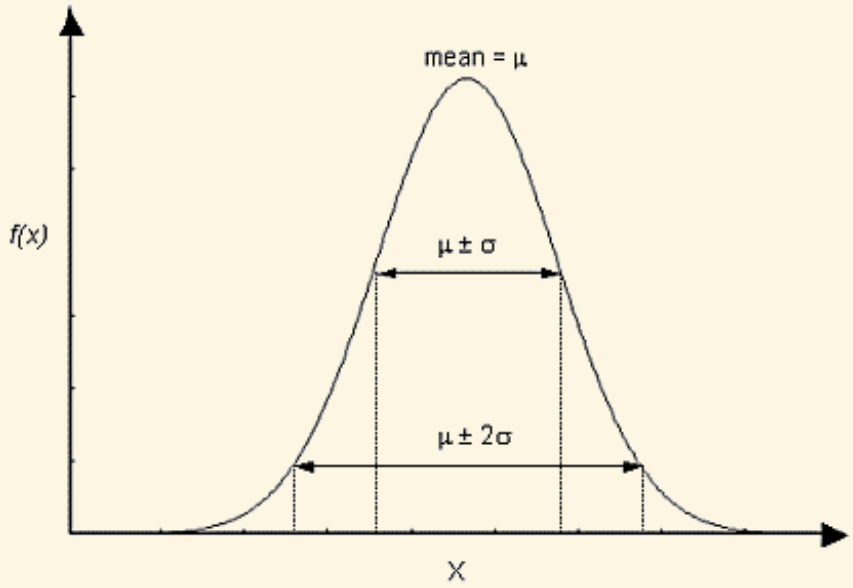
Алгоритм классификации: Логистическая регрессия

Простая линейная регрессия



Линейная регрессия: генеративная формулировка

$$\begin{aligned} H_i &\sim \mathcal{N}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta W_i \end{aligned}, \quad (1)$$

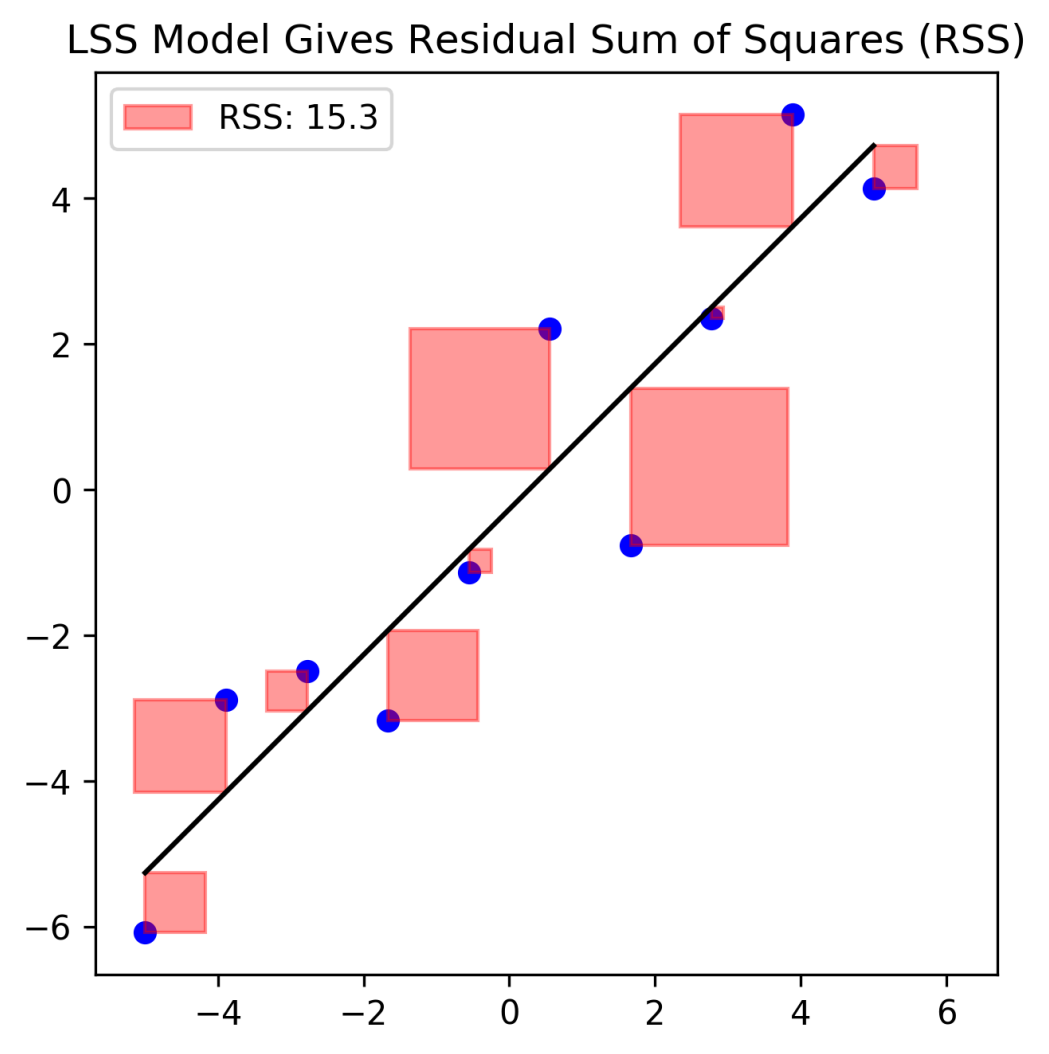


где

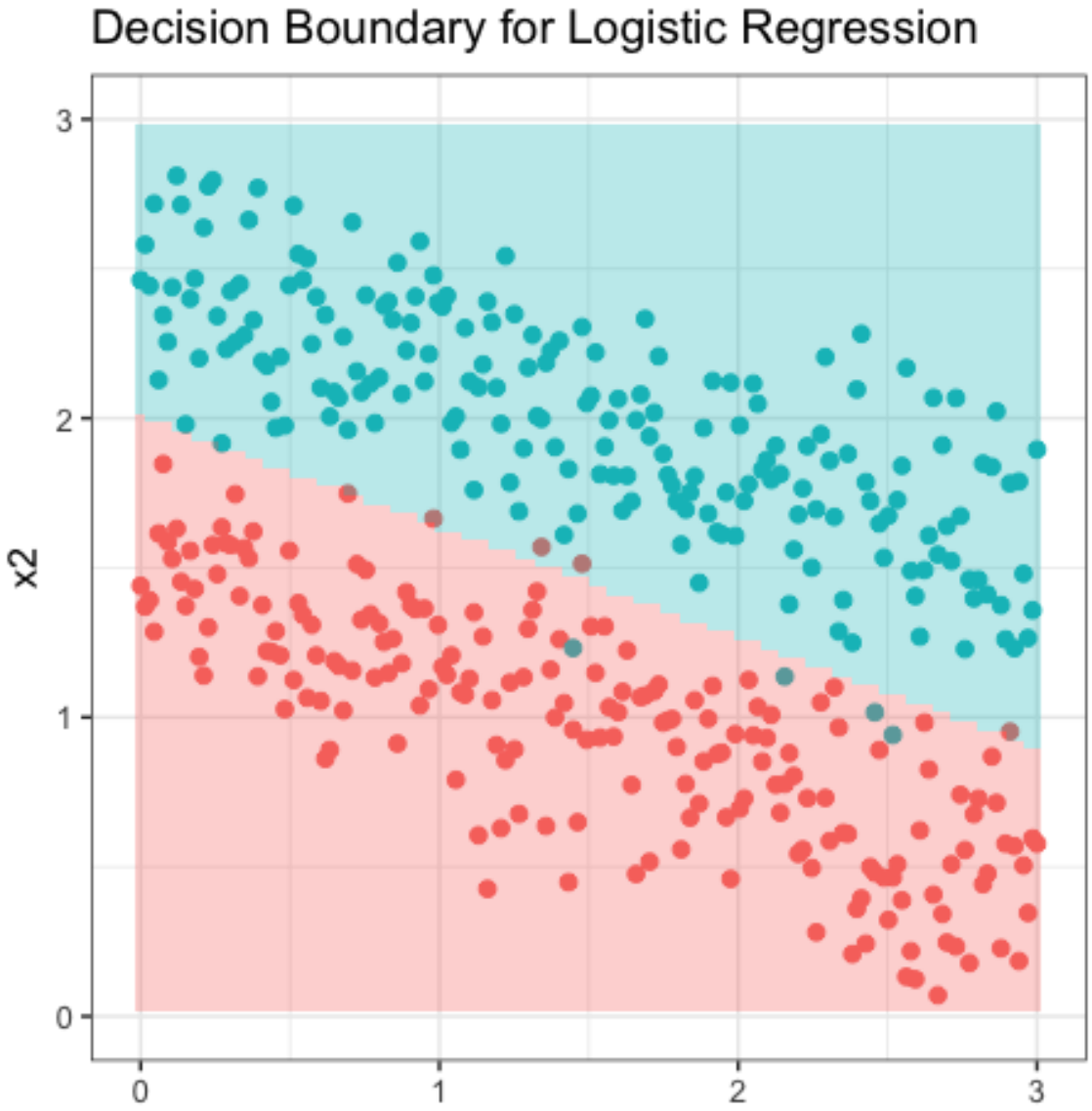
H_i рост индивида i

W_i вес индивида i

RSS — функция потерь



Логистическая регрессия: бинарный классификатор

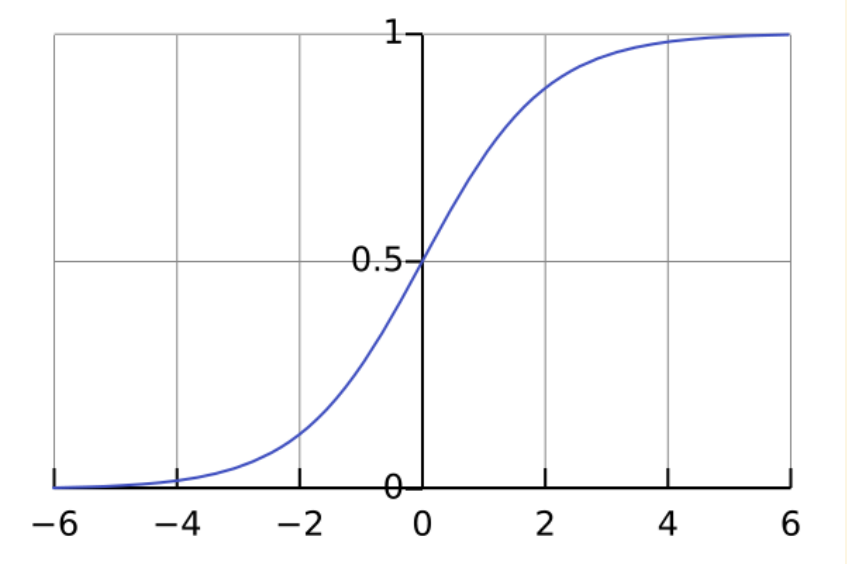


Логистическая регрессия: генеративная формулировка

$$Y_i \sim \text{Binomial}(1, p_i)$$
$$\text{logit}(p_i) = \alpha + \beta_1 x_1 + \beta_2 x_2, \quad (2)$$
$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i}$$

где

- Y_i класс индивида i {0,1}
- p_i вероятность позитивного класса (1) для индивида i



Логистическая регрессия: генеративная формулировка

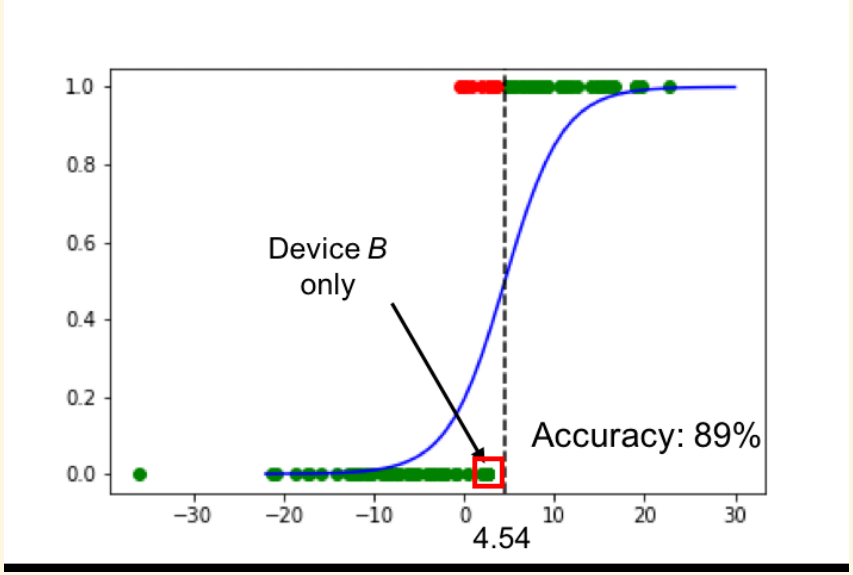
$$Y_i \sim \text{Binomial}(1, p_i)$$

$$\text{logit}(p_i) = \alpha + \beta_1 x_1 + \beta_2 x_2, \quad (2)$$

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i}$$

где

- Y_i класс индивида i $\{0,1\}$
- p_i вероятность позитивного класса (1) для индивида i



Логистическая регрессия — функция потерь

$$LS = -(ylog(p) + (1 - y)log(1 - p)) \tag{3}$$

Примеры:

$$y = 1; p = 0.8$$
$$LS = -(1 * log(0.8) + (1 - 1) * log(0.2)) = \tag{4}$$
$$= 0.22$$

$$y = 0; p = 0.8$$
$$LS = -(0 * log(0.8) + (1 - 0) * log(0.2)) = \tag{5}$$
$$= 1.6$$

Логистическая регрессия — функция потерь

$$LS = -(ylog(p) + (1-y)log(1-p)) \tag{3}$$

Примеры:

y = 1 ; p = 0.8

$$LS = -(1 * log(0.8) + (1-1) * log(0.2)) = \tag{4}$$
$$= 0.22$$

$$y = 0; p = 0.8$$
$$LS = -(0 * log(0.8) + (1-0) * log(0.2)) = \tag{5}$$
$$= 1.6$$

Логистическая регрессия — функция потерь

$$LS = -(ylog(p) + (1 - y)log(1 - p)) \tag{3}$$

Примеры:

$$y = 1; p = 0.8$$

$$LS = -(1 * log(0.8) + (1 - 1) * log(0.2)) = \tag{4}$$

$$= 0.22$$

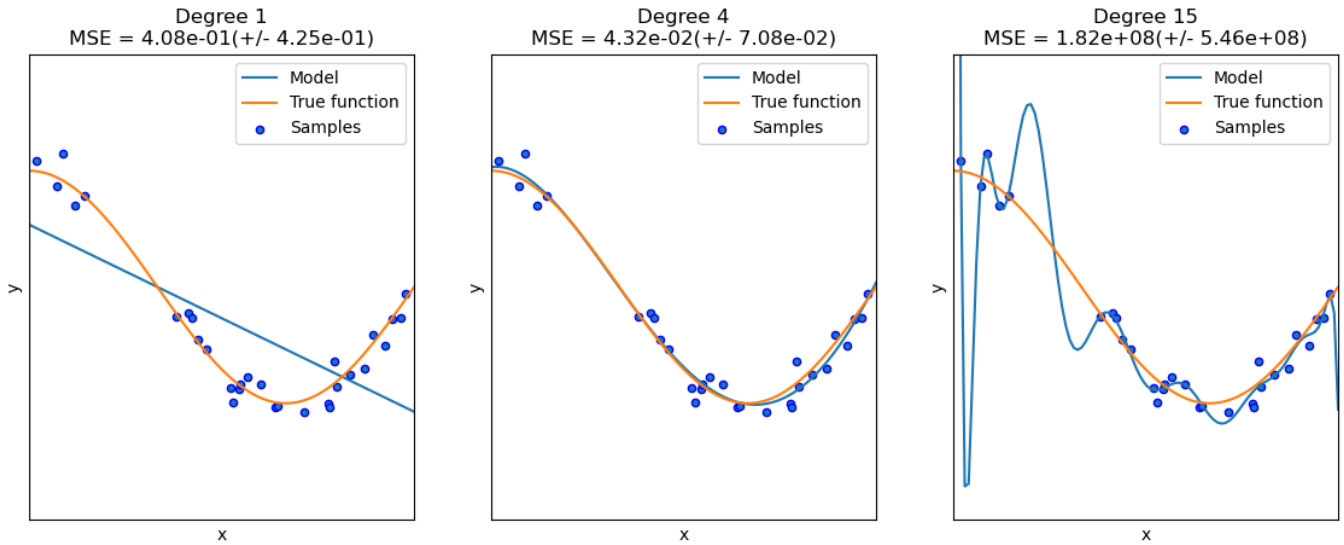
y = 0 ; p = 0.8

$$LS = -(0 * log(0.8) + (1 - 0) * log(0.2)) = \tag{5}$$

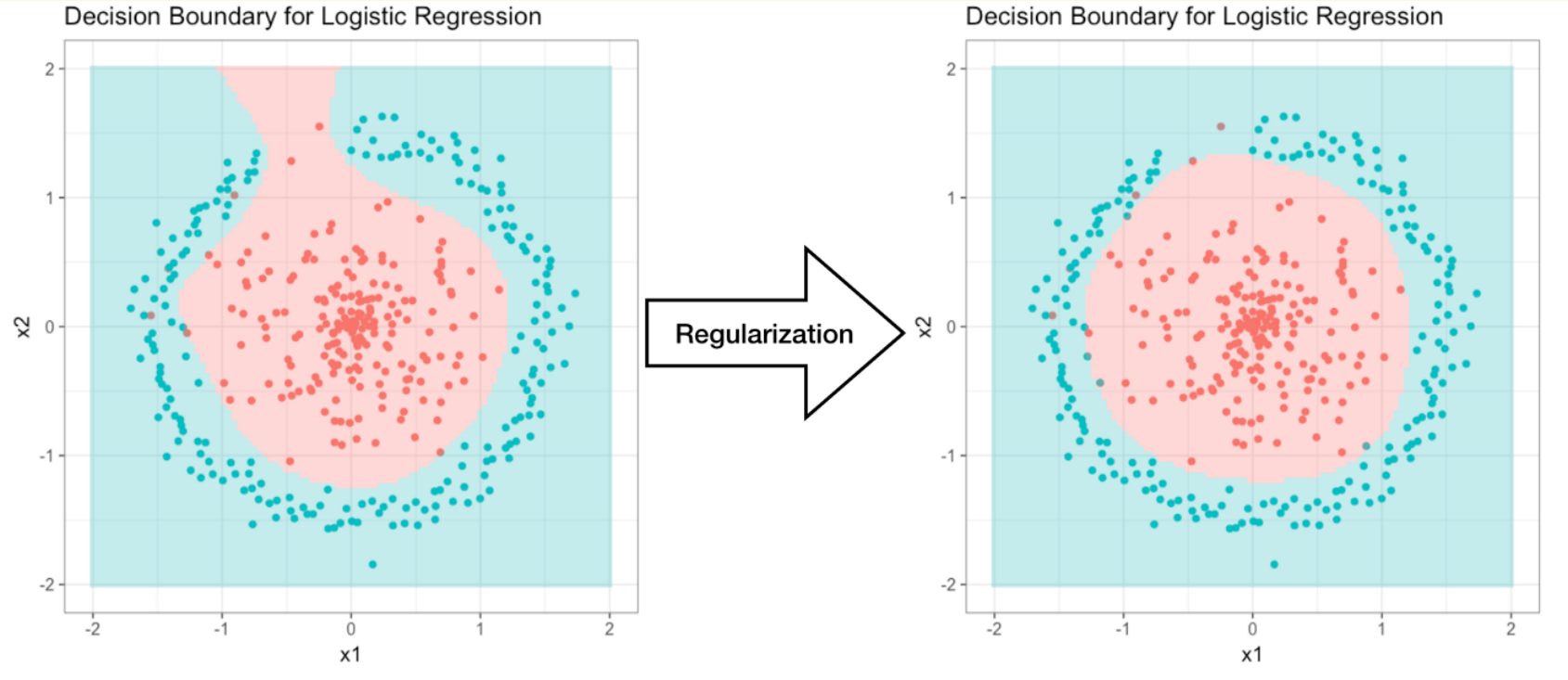
$$= 1.6$$

Снижение размерности: Регуляризация

Underfitting/Overfitting



Регуляризация



Регуляризованная функция потерь

Обычная регрессия:

$$Loss = Error(y, \hat{y}) \tag{6}$$

L1 Loss (LASSO regression):

$$Loss = Error(y, \hat{y}) + \lambda \sum |\beta_j| \tag{7}$$

L2 Loss (RIDGE regression):

$$Loss = Error(y, \hat{y}) + \lambda \sum \beta_j^2 \tag{8}$$

Регуляризованная функция потерь

Обычная регрессия:

$$Loss = Error(y, \hat{y}) \tag{6}$$

L1 Loss (LASSO regression):

$$Loss = Error(y, \hat{y}) + \lambda \sum |\beta_j| \tag{7}$$

L2 Loss (RIDGE regression):

$$Loss = Error(y, \hat{y}) + \lambda \sum \beta_j^2 \tag{8}$$

Регуляризованная функция потерь

Обычная регрессия:

$$Loss = Error(y, \hat{y}) \tag{6}$$

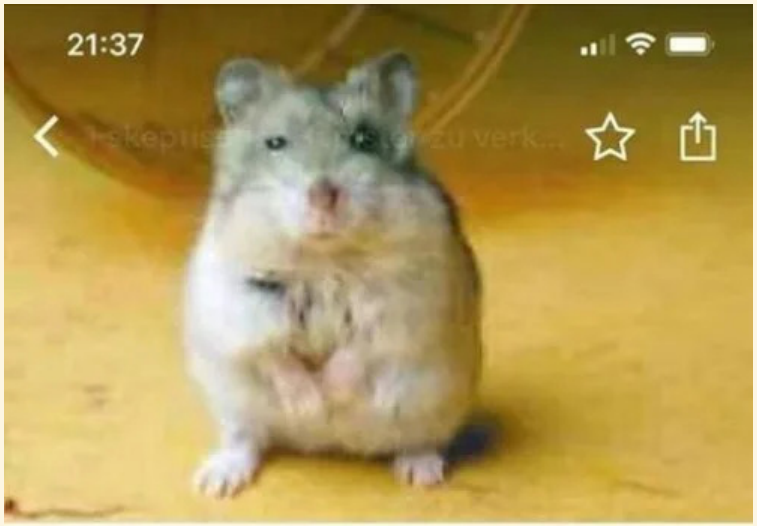
L1 Loss (LASSO regression):

$$Loss = Error(y, \hat{y}) + \lambda \sum |\beta_j| \tag{7}$$

L2 Loss (RIDGE regression):

$$Loss = Error(y, \hat{y}) + \lambda \sum \beta_j^2 \tag{8}$$

Sceptical hamster



1 skeptischer Hamster zu verkaufen

20 €

25899 Niebüll >

Art	Hamster

Er guckt einen skeptisch an, als würde man nichts richtig machen.
Es macht mich wahnsinnig, ich kann diesen vorwurfsvollen Blick nicht länger ertragen.
Sein Name ist Olaf.