

Программная инженерия. Разработка ПО (Python для продвинутых специалистов. Машинное обучение)

Модуль: Предобработка данных и машинное обучение

Лекция 6: Масштабирование и нормализация данных

Дата: 02.06.2025

Содержание лекции

- Для чего нужна нормализация
- Виды нормализации



StandardScaler

Что делает:

Масштабирует данные так, чтобы **среднее = 0**, а **стандартное отклонение = 1**.

Формула:

$z = \frac{x - \mu}{\sigma}$ где μ — среднее значение признака,
а σ — стандартное отклонение.

Когда использовать:

- Когда данные имеют распределение, близкое к **нормальному (Gaussian)**.
- Когда важны отклонения от среднего (например, для линейной или логистической регрессии, PCA).

Уязвим к выбросам:  Да

Что делает:

Использует **медиану** и **интерквартильный размах (IQR)** вместо среднего и стандартного отклонения.

Формула:

$$x_{\text{scaled}} = \frac{x - \text{median}}{\text{IQR}} \text{ где}$$

$$\text{IQR} = Q3 - Q1$$

Когда использовать:

- Когда в данных есть **выбросы**.
- Когда данные асимметричны.

Уязвим к выбросам: **✗** Нет

MinMaxScaler

Что делает:

Масштабирует данные в диапазон от 0 до 1.

Формула:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Когда использовать:

- Когда признаки должны быть в **одном масштабе** (например, для k-NN, нейросетей).
- Когда нет сильных выбросов.

Уязвим к выбросам:  Да



Передовые
инженерные
школы



МИНОБРНАУКИ
РОССИИ



УНИВЕРСИТЕТ
ИННОПОЛИС



онлайн
университет

Спасибо за внимание

