Základy čištění dat (Data Cleaning)

Čištění dat je proces identifikace a opravy (nebo odstranění) chybných, poškozených, nesprávně formátovaných, duplicitních nebo neúplných dat v rámci datasetu. Cílem je vytvořit čistou, konzistentní sadu dat připravenou pro analýzu a trénování modelů.

1. Odstraňování duplicit

Duplicitní záznamy vznikají při spojování databází, chybami při sběru nebo opakovaným odesláním formulářů.

- **Problém:** Duplicity uměle zvyšují váhu určitých pozorování a mohou vést k přeučení modelu.
- **Řešení:** Identifikace unikátních identifikátorů (např. ID uživatele, časové razítko) a odstranění identických řádků.

2. Řešení chybějících hodnot (Missing Values)

Data často chybí (uživatel nevyplnil pole, senzor měl výpadek). Máme tři hlavní cesty:

- **Odstranění (Deletion):** Smazání celého řádku nebo sloupce. Vhodné jen pokud chybí velké procento dat.
- **Imputace (Imputation):** Nahrazení chybějící hodnoty:
  - **Numerická data:** Průměrem, mediánem nebo konstantou.
  - **Kategorická data:** Nečastější hodnotou (modusem) nebo kategorií „Neznámý“.
- **Predikce:** Použití jiného algoritmu (např. lineární regrese) k odhadu chybějící hodnoty.

3. Oprava nekonzistentních dat

Nekonzistence znemožňují správné seskupování a analýzu.

- **Typové chyby:** Čísla uložená jako text, různý formát data (12/01/2024 vs. 2024-01-12).
- **Překlepy a synonyma:** „USA“, „U.S.A.“, „United States“ – vše by mělo být sjednoceno na jeden tvar.
- **Měřítko:** Různé jednotky (metry vs. kilometry) v jednom sloupci.

4. Práce s odlehlými hodnotami (Outliers)

Odlehlé hodnoty mohou být buď chyby (výška 300 cm), nebo legitimní extrémy.

- **Postup:** Detekce pomocí statistických metod (IQR nebo Z-Score) a následné rozhodnutí, zda je smazat, transformovat nebo ponechat.

5. Strukturální chyby

Zahrnují nesprávné pojmenování sloupců, prázdné mezery na začátku/konci textu (trimming) nebo nesmyslné hodnoty (např. věk -5).

Průběh čištění dat (Workflow)

Krok	Akce
1. Průzkum (EDA)	Prohlédnutí statistik, grafů a nalezení zjevných chyb.
2. Filtrace	Odstranění nerelevantních sloupců a duplicit.
3. Oprava typu	Převod datových typů (string na float, object na datetime).
4. Handling Missing	Vyplnění nebo smazání prázdných hodnot.
5. Validace	Kontrola, zda data po vyčištění dávají smysl a splňují integritní omezení.

Ukázka v Pythonu (Pandas)

```
import pandas as pd

# Načtení dat
df = pd.read_csv('data.csv')

# Odstranění duplicit
df.drop_duplicates(inplace=True)

# Nahrazení chybějících hodnot mediánem
df['vek'] = df['vek'].fillna(df['vek'].median())

# Sjednocení textu na malá písmena a odstranění mezer
df['mesto'] = df['mesto'].str.lower().str.strip()

# Odstranění řádků s nesmyslnými hodnotami
df = df[df['vek'] > 0]
```

Související články:

- Detailně o odlehlých hodnotách
- Standardizace a normalizace
- Explorační analýza dat (EDA)

Tagy: ml preprocessing data_cleaning pandas data_science

From:
https://serviceit.cz/ - **IT ENCYKLOPEDIE**

Permanent link:
https://serviceit.cz/doku.php?id=it:ml:data_cleaning

Last update: **2026/01/02 12:48**