Jak pracovat s odlehlými hodnotami (Outliers)

Odlehlá hodnota (Outlier) je datový bod, který se výrazně liší od ostatních pozorování v datové sadě. Může jít o chybu měření, chybu při zadávání dat, nebo o vzácný, ale reálný extrém (např. plat miliardáře v průzkumu mezd).

Práce s těmito hodnotami je klíčová, protože mnoho algoritmů strojového učení je na ně extrémně citlivých.

1. Jak detekovat odlehlé hodnoty?

Než začneme hodnoty odstraňovat, musíme je identifikovat. Existují k tomu grafické i statistické metody:

A. Grafické metody

- **Boxplot (Krabicový graf):** Nejjednodušší nástroj. Body zobrazené vně „vousů“ (whiskers) jsou považovány za kandidáty na odlehlé hodnoty.
- **Scatter plot (Bodový graf):** Pomáhá vidět body, které nezapadají do trendu mezi dvěma proměnnými.

B. Statistické metody

- **Z-Score (Z-skóre):** Určuje, o kolik směrodatných odchylek je bod vzdálen od průměru. Pravidlem bývá, že body s $|Z| > 3$ jsou odlehlé.
- **Interquartile Range (IQR):** Rozdíl mezi 3. kvartilem ($Q3$) a 1. kvartilem ($Q1$).
    - Spodní hranice = $Q1 - 1.5 \cdot IQR$
    - Horní hranice = $Q3 + 1.5 \cdot IQR$

2. Strategie řešení: Co s nimi dělat?

Jakmile odlehlé hodnoty najdeme, máme čtyři hlavní možnosti, jak s nimi naložit:

1. Odstranění (Trimming / Dropping)

Pokud víme, že jde o chybu (např. věk 200 let), záznam jednoduše vymažeme.

- **Riziko:** Můžeme ztratit důležité informace, pokud je dat málo.

2. Transformace

Použití matematických funkcí, které „přitáhnou“ extrémní hodnoty blíže ke středu.

- **Logaritmická transformace ($log(x)$):** Velmi účinná u dat s kladným sešikmením (např. příjmy).
- **Odmocnina ($\sqrt{x}$):** Jemnější forma redukce rozptylu.

3. Imputace (Nahrazení)

Nahrazení odlehlé hodnoty mediánem nebo průměrem celé sady.

- **Winsorizace:** Extrémní hodnoty se nesmažou, ale „zastropují“ na určitém percentilu (např. vše nad 95. percentilem bude mít hodnotu 95. percentilu).

4. Ponechání a výběr robustního algoritmu

Někdy jsou odlehlé hodnoty to nejdůležitější (např. detekce podvodů). V takovém případě zvolíme algoritmy, které s nimi umí pracovat:

- **Rozhodovací stromy / Random Forest:** Jsou velmi odolné vůči outlierům.
- **Robustní regrese:** Místo MSE minimalizuje metriku méně citlivou na extrémy (např. MAE).

3. Příklad v Pythonu (Metoda IQR)

```
import pandas as pd

# Výpočet mezikvartilového rozpětí
Q1 = df['sloupec'].quantile(0.25)
Q3 = df['sloupec'].quantile(0.75)
IQR = Q3 - Q1

# Definice hranic
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filtrování outlierů
df_clean = df[(df['sloupec'] >= lower_bound) & (df['sloupec'] <=
upper_bound)]
```

4. Shrnutí: Kdy mazat a kdy ne?

Situace	Doporučený postup
Zjevná technická chyba (mínusová cena)	Smazat nebo opravit.

Situace	Doporučený postup
Přirozený extrém (extrémně bohatý klient)	Ponechat, ale použít robustní algoritmus nebo transformaci.
Hodnota je předmětem zkoumání (podvody)	**Vždy ponechat**, zde jsou outlieři cílem analýzy.

Související články:

- Standardizace a normalizace
- Základy čištění dat
- Regresní analýza

Tagy: ml statistics outliers data_cleaning preprocessing