

Shlukování (Clustering)

Shlukování (nebo také clustering) je metoda **učení bez učitele**, která spočívá v rozdělování datových objektů do skupin (shluků) tak, aby si objekty uvnitř jedné skupiny byly co nejvíce podobné, zatímco objekty z různých skupin byly co nejvíce odlišné.

Na rozdíl od klasifikace zde nemáme předem definované třídy. Algoritmus hledá přirozené hranice v datech na základě jejich vlastností.

1. Základní principy měření podobnosti

Aby mohl algoritmus shlukovat, musí umět změřit „vzdálenost“ mezi dvěma body. Nejčastěji se používají:

- **Euklidovská vzdálenost:** Přímá vzdálenost mezi dvěma body v prostoru (vhodná pro numerická data).
- **Manhattanská vzdálenost:** Součet absolutních rozdílů souřadnic (cesta městem s pravoúhlými ulicemi).
- **Kosinová podobnost:** Měří úhel mezi dvěma vektory (často používaná u textových dat).

2. Klíčové algoritmy shlukování

A. K-Means (K-středy)

Nejrozšířenější algoritmus. Rozděluje data do předem určeného počtu shluků (**K**).

1. **Jak funguje:** Náhodně zvolí K středů (centroidů). Každý bod přiřadí k nejbližšímu středu. Poté středy přepočítá jako průměr přiřazených bodů a proces opakuje, dokud se středy nepřestanou hýbat.
2. **Výhody:** Velmi rychlý a efektivní.
3. **Nevýhody:** Nutnost předem znát počet shluků K, citlivost na odlehlé hodnoty.

B. DBSCAN (Hustotní shlukování)

Hledá oblasti s vysokou hustotou bodů, které jsou odděleny oblastmi s nízkou hustotou.

1. **Jak funguje:** Seskupuje body, které mají v určitém okolí dostatečný počet sousedů.
2. **Výhody:** Dokáže najít shluky libovolného tvaru (ne jen kruhové) a automaticky identifikuje šum (body, které nepatří nikam).
3. **Nevýhody:** Špatně funguje u dat s výrazně odlišnou hustotou v různých částech.

C. Hierarchické shlukování

Vytváří stromovou strukturu shluků (dendrogram).

1. **Aglomerativní:** Začne s každým bodem jako samostatným shlukem a postupně ty nejbližší spojuje.
2. **Divizivní:** Začne se všemi body v jednom shluku a postupně je dělí.
3. **Výhoda:** Názná vizualizace vztahů mezi daty.

3. Jak určit správný počet shluků (K)?

U algoritmů jako K-Means je určení správného počtu skupin největší výzvou. Používají se dvě hlavní metody:

1. ****Loketní metoda (Elbow Method):**** Graficky znázorníme součet čtverců vzdáleností bodů od středů. Tam, kde se křivka prudce ohne ("loket"), je ideální počet shluků.
2. ****Siluetový koeficient (Silhouette Score):**** Měří, jak dobře bod zapadá do svého shluku ve srovnání s ostatními shluky. Hodnota se pohybuje od -1 do 1 (čím vyšší, tím lépe).

4. Praktické využití v IT a byznysu

- **Marketing:** Segmentace zákazníků pro personalizované kampaně.
- **Zpracování obrazu:** Kvantování barev (snížení počtu barev v obrázku při zachování vzhledu).
- **Sociální sítě:** Identifikace komunit se společnými zájmy.
- **Biologie:** Seskupování genů nebo organismů s podobnými vlastnostmi.

5. Úskalí a omezení

- **Prokletí dimenzionality:** Ve velmi vysokých dimenzích (mnoho sloupců dat) se vzdálenosti mezi všemi body stávají téměř stejnými, což shlukování ztěžuje.
- **Škálování dat:** Protože většina algoritmů závisí na vzdálenosti, je nutné data předem normalizovat (např. aby výška v metrech neměla menší váhu než váha v gramech).

Související články:

- [Metriky vzdálenosti v ML](#)
- [Učení bez učitele](#)
- [Redukce dimenzionality \(PCA\)](#)

Tagy: ai ml clustering k-means dbscan unsupervised_learning

From:

<https://www.serviceit.cz/> - IT ENCYKLOPEDIE

Permanent link:

<https://www.serviceit.cz/doku.php?id=it:ml:clustering>

Last update: **2026/01/02 12:43**

