

Vytěžování dat, cvičení 4: Shlukování

Miroslav Čepek, Michael Anděl



Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

1. Doimplementujte K-Means algoritmus. Přiložená funkce v Matlabu implementuje část KMeans algoritmu.
2. Centroidy (reprezentanty) inicializujte náhodně a při každém spuštění jinak.
3. Shlukněte přiložená data vaším KMeans algoritmem. Zkuste různé počty reprezentantů $k = \{2, 3, \dots, 10\}$.
4. Spočítejte siluetu (funkce `silhouette`) pro všechny počty shluků k , vynesete průměrnou siluetu do grafu.
5. Silueta je jednou z heuristik na určení *přirozeného* počtu shluků. Stanovte tedy předpokládaný počet shluků v datech. Pro zajímavé počty shluků zobrazte siluety a interpretujte je.

- ▶ Vámi doplněný zdrojový kód.
- ▶ Protokol:
 1. Průměrné hodnoty siluety pro jednotlivé počty reprezentantů.
 2. Dále přiložte zajímavé grafy siluet a jejich interpretaci. Dodejte zdůvodnění přirozeného počtu shluků v datech. Volitelně, pokud vám přijde zajímavý, může zpráva také obsahovat 2D/3D bodový graf se zvýrazněnými shluky.