Vytěžování dat, cvičení 5: Shlukování

Miroslav Čepek

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Zadání domácího úkolu

1. Doimplementujte K-Means algoritmus. Přiložená funkce v Matlabu implementuje část KMeans algoritmu (nalezení nejbližších reprezentantů (centroidů)). Vaším úkolem je doplnit přesun reprezentantů do středu nových shluků a určit, zda je možné ukončit algoritmus nebo má smysl pokračovat další iterací.
2. Centroidy (reprezentanty) inicializujte náhodně a při každém spuštění jinak.
3. Shlukněte přiložená data vaším KMeans algoritmem. Zkuste různé počty reprezentantů (2,3, . . . ,10). Spočítejte průměrnou siluetu pro všechny počty shluků a určete pro který počet reprezentantů vyjde průměrná silueta nejlépe. Pro zajímavé počty reprezentantů zobrazte grafy siluet.

Zadání domácího úkolu (2)

1. Pro nejlepší počet reprezentantů, který vám vyšel v minulém bodě, (alespoň 5x) spusťte algorimus KMeans s různými náhodnými počátečními pozicemi reprezentantů.
2. Shlukněte data pomocí hierarchického shlukování. Vytvořte stejný počet shluků, který vám vyšel nejlépe v algoritmu KMeans. Do zprávy vložte dendrogram, graf siluety a průměrnou siluetu. Krátce okomentujte rozdíly mezi výsledky hierarchického shlukování a KMeans algoritmu.

Obsah zprávy

1. Vámi doplňený zdrojový kód. !!A jeho stručný popis!!
2. Průměrné hodnoty siluety pro počty reprezentantů: 2, 3, 4, . . ., 10.
3. Dále přiložte zajímavé grafy siluet. Volitelně, pokud vám přijde zajímavý, může zpráva také obsahovat 2D/3D bodový graf se zvýrazněnými shluky.
4. Hodnoty průměrných siluet a výsledných souřadnic reprezentantů pro různé náhodné počáteční pozice reprezentantů. Pro počet reprezentantů, který vám vyšel nejlepší, v minulém bodě.
5. Dendrogram, který vám vyšel z hierarchického shlukování. A průměrná silueta a graf siluety stejný pro počet shluků, jako vám vyšel nejlepší v algoritmu KMeans.

Užitečné funkce

- silhouette
- kmeans
- linkage
- pdist
- cluster
- cophenet
- scatter