TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Shlukování 2

František Kynych
31. 10. 2024 | MVD

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Část I.:
Rozšíření K-Means algoritmu

K-Means inicializace

- Standardně v prvním kroku vybráno K náhodných centroidů
  - Algoritmus je potřeba provést několikrát
- Další možnosti
  - K-Means++ (2007)
    - První centroid je vybrán náhodně
    - Další centroid je nejdále od prvního centroidu, …
  - Využití řazení
    - Podle nějaké metriky seřadíme body
      - Vzdálenost od centra, hustota, …
    - Na základě seřazení vybereme centroidy
      - Prvních K bodů (+ zákaz bodů bližších než ε k již existujícím centroidům)
      - Každý bod na indexu $\frac{N}{K}$

K-Medoids clustering

- Průměrný bod je nahrazen za medoid
- Medoid
  - Centrální bod v clusteru
  - Vždy je reprezentován reálným objektem z dat

- Nejpoužívanější K-Medoids metoda – Partitioning Around Medoids (PAM, vznik 1990)

1. Vybereme K medoidů
2. Jednotlivé body přiřadíme do clusteru s nebližšími medoidy
3. Prohledáváme body v každém clusteru
   - Pokud nějaký bod v clusteru snižuje průměrnou vzdálenost, tak se stane medoidem
4. Pokud se změnil alespoň jeden medoid, tak dále iterujeme od bodu 2.

K-Medoids clustering vylepšení

- Výpočetní náročnost PAM - $O(K(n-K)2)$
  - Problém s velkým množstvím dat
- **CLARA (1990)**
  - Řeší výpočetní náročnost PAM
  - Neprohledáváme všechny body v clusteru, ale pouze S bodů
  - $O(KS^2 + K(n-K))$
  - Špatné výsledky pokud je jeden nebo více počátečních medoidů daleko od nejlepších možných medoidů
- **CLARANS (1994)**
  - Vylepšené hledání nových medoidů

K-Medians

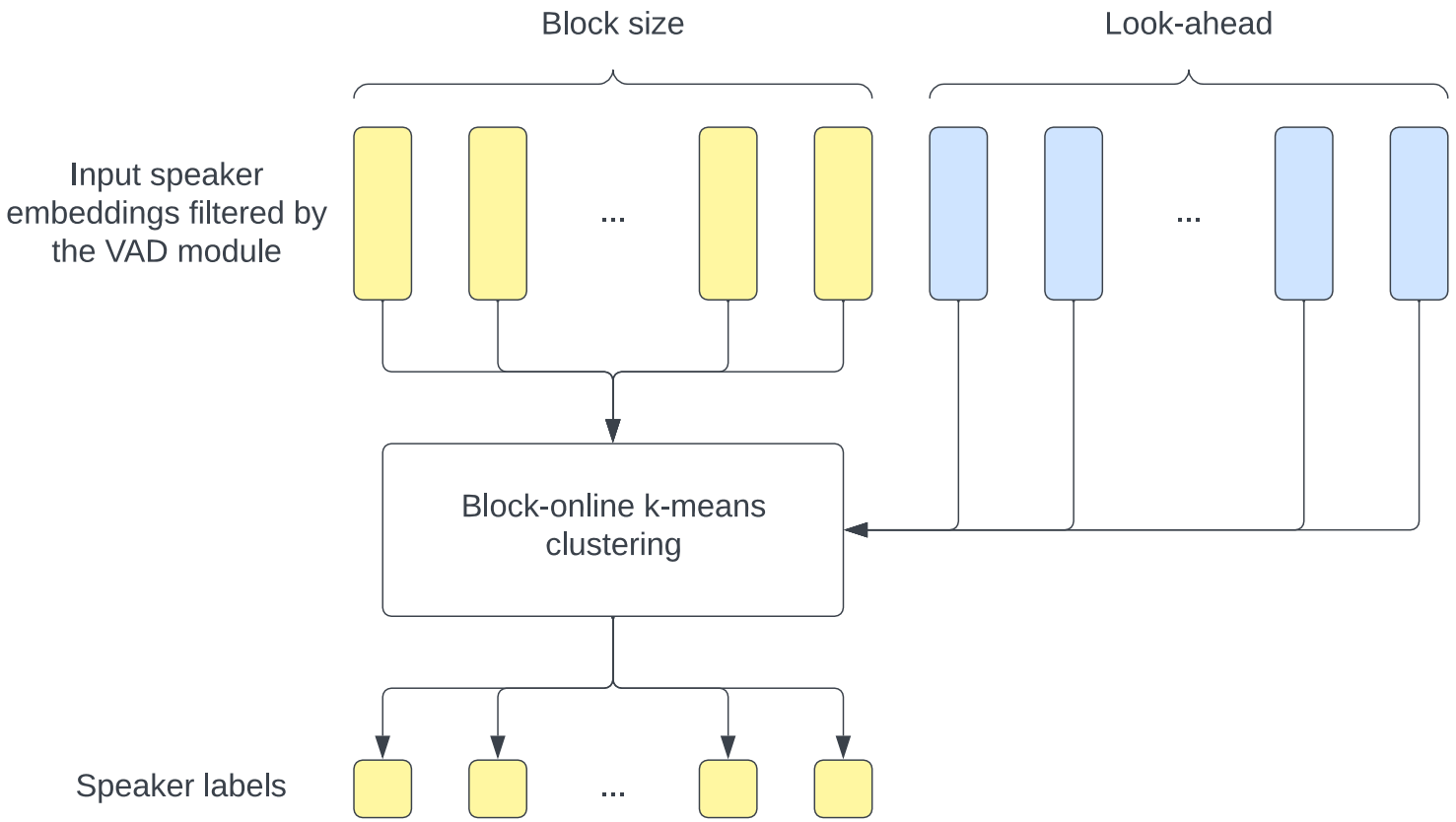
- **K-Medians**
  - Medián je méně náchylný na outliery
  - Mediány jsou použity jako centroidy (+ použití L1 normy jako vzdálenostní metriky)
  - Stejné postup jako u K-Means, pouze nový medián je vypočten jako medián jednotlivých příznaků (každé dimenze)

Sequential K-Means

- Podobný jako standardní K-Means algoritmus
- S přicházejícími body postupně aktualizujeme polohu centroidů

- Existuje i možnost, kdy předem nelze odhadnout počet shluků
  - Vytvoření prvního shluku s počátečními daty
  - Pro každý nový bod je potřeba se rozhodnout, zda bude zařazen do již existujícího shluku nebo bude vytvořen nový
    - Řešeno na základě vzdálenostní metriky

Block-online K-Means s look-ahead



Část II.:
Shlukování založené na hustotě

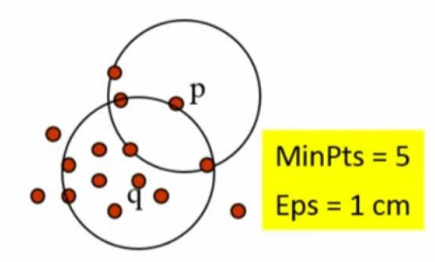
Shlukování založené na hustotě

- Shluky jsou založeny na základě hustoty dat
  - Lokální kritérium

- Výhody
  - Nalezne shluky libovolného tvaru
  - Nevadí šum v datech
  - Stačí projít jednou

DBSCAN

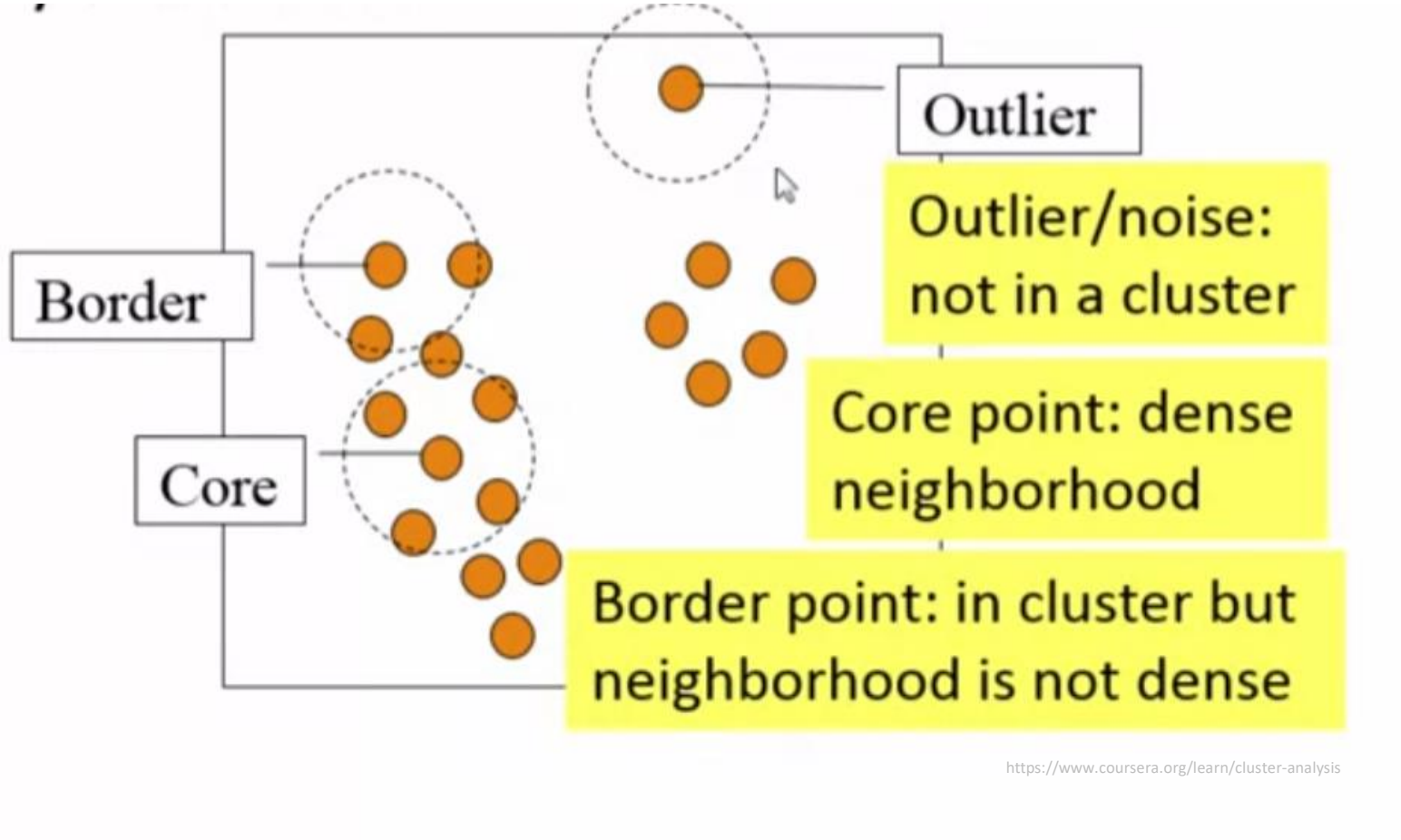
- Density-Based Spatial Clustering Algorithm
- Shluk je definován jako maximální množina hustě spojených bodů



https://www.coursera.org/learn/cluster-analysis

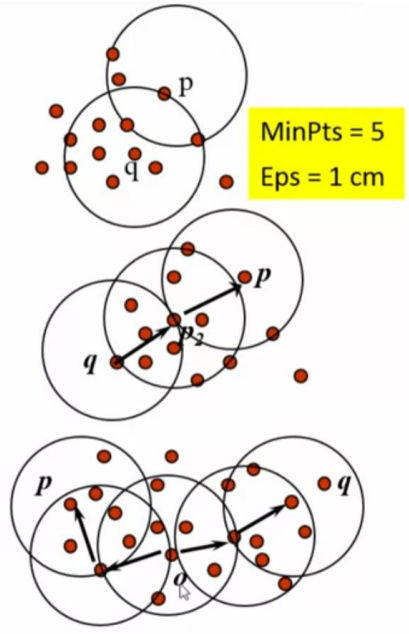
- Dva parametry
  - **Epsilon (Eps)**
    - Radius okolo bodu
  - **MinPts**
    - Minimální počet bodů v okolí Eps
    - Eps okolí bodu q: $N_{Eps}(q) = \{p \; patří \; do \; D \mid d(p,q) \leq Eps\}$

DBSCAN


Outlier
Border
Core
Outlier/noise:
not in a cluster
Core point: dense
neighborhood
Border point: in cluster but
neighborhood is not dense
https://www.coursera.org/learn/cluster-analysis

DBSCAN

- Přímá dosažitelnost
  - Bod p je přímo dosažitelný z bodu q, pokud:
    - p patří do $N_{Eps}(q)$ (bod p patří do okolí bodu q)
    - $|N_{Eps}(q)| \geq MinPts$ (bod q je core bod)
- Nepřímá dosažitelnost
  - Bod p je nepřímo dosažitelný z bodu q, pokud existuje řada bodů $p_1, \ldots, p_n$ $(p_1 = q, p_n = p)$ taková, že p_{i+1} je přímo dosažitelný z p_i
- Propojenost
  - Body p a q jsou propojené, pokud existuje bod o, ze kterého jsou oba body (p a q) nepřímo dosažitelné



https://www.coursera.org/learn/cluster-analysis

DBSCAN

Algoritmus:

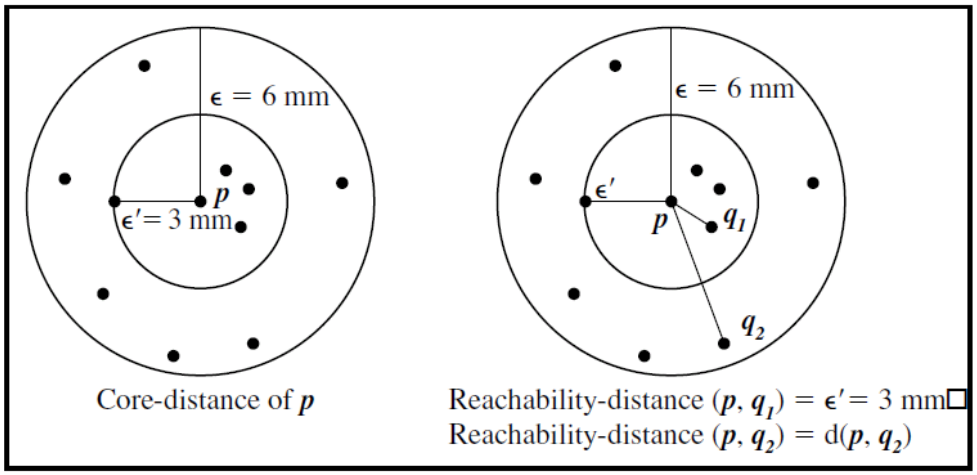
1. Vybrat počáteční bod *p*
2. Získat všechny nepřímo dosažitelné body
   a) Pokud je bod *p* core bod -> cluster hotový
   b) Pokud je *p* border bod, tak z něj ostatní body nemohou být nepřímo dosažitelné -> vybrat další počáteční bod z databáze
3. Pokračujeme, dokud jsme neprošli všechny body

OPTICS

- Ordering Points To Identify Clustering Structure
- Vytvořeno (téměř) stejnými autory jako DBSCAN
  - DBSCAN je citlivý na nastavení správných parametrů

- Parametr maximální Epsilon poskytujeme pouze pokud chceme zrychlit výpočet
- Vytvoření grafu dosažitelnosti, ze kterého lze extrahovat shluky

OPTICS



https://www.coursera.org/learn/cluster-analysis

- **Core distance** – minimální hodnota ε, se kterou bude v okolí *MinPts* bodů (aby se stal core bodem)
- **Reachability-distance** – minimální radius, díky kterému je p nepřímo dosažitelný z bodu q

$$\max(core\_distance(q), d(q,p))$$

  - Pokud je q core bod, jinak nedefinováno

OPTICS

- Graf dosažitelnosti
  - Čím hlubší je oblast, tím hustší je shluk
- Body patřící do shluku mají nízkou vzdálenost dosažitelnosti k jejich nejbližším sousedům, proto vytvoří údolí

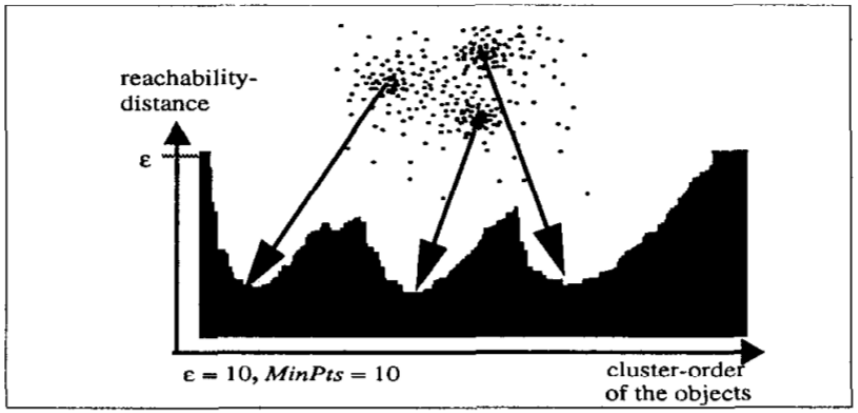


Figure 9. Illustration of the cluster-ordering

Figure 12. Reachability-plots for a data set with hierarchical clusters of different sizes, densities and shapes

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.407.5572&rep=rep1&type=pdf

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Část III.:
Spektrální (grafové) shlukování

Spektrální shlukování

- Založeno na teorii grafů
- Využívá spektrální informace matice podobnosti dat
- Transformuje data do nového prostoru pomocí vlastních vektorů
  - Umožňuje aplikaci K-means algoritmu

Spektrální shlukování

Vytvoření grafu z dat

- **Uzly** reprezentují jednotlivé datové body
- **Hrany** reprezentují podobnost mezi datovými body
- **Matice podobnosti** (W) vyjadřuje podobnost mezi body *i* a *j*

Vytvoření matice W

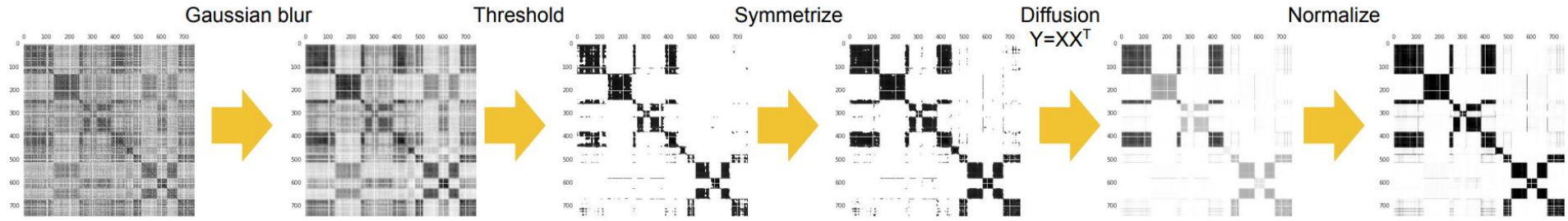
$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

U některých aplikací se používá např. kosinová podobnost

$$W_{ij} = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|}$$

Spektrální shlukování

Upravení grafu



https://github.com/wq2012/SpectralCluster

Spektrální shlukování

Matice stupňů (Degree matrix)

- Diagonální matice D, kde každý prvek na diagonále představuje stupeň uzlu i v grafu
- Stupeň Uzlu: Součet všech podobností daného uzlu s ostatními uzly

$$D_{ii} = \sum_j W_{ij}$$

- Vlastnosti
  - Všechny ne-diagonální prvky jsou nulové
  - Všechny prvky D_{ii} jsou pozitivní

Spektrální shlukování

Laplacián grafu

- Matice L definovaná jako rozdíl mezi maticí stupňů D a maticí podobnosti W

$$L = D - W$$

- Normalizace

$$L_{sym} = D^{-1/2} L D^{-1/2}$$

- Zachycuje strukturu grafu a vztahy mezi uzly
- Používá se při spektrální analýze k identifikaci shluků pomocí vlastní vektorů

- Vlastní vektory Laplaciánu poskytují nové reprezentace dat, které lépe odhalují skryté struktury a shluky

Spektrální shlukování

- Získání vlastních hodnot λ a vlastních vektorů v Laplaciánu

- Transformace do spektrálního prostoru
  - Vybereme k vlastních vektorů odpovídajících nejmenším vlastním hodnotám $\lambda_1, \lambda_2, \ldots, \lambda_k$ do matice V_k

$$Y = X \cdot V_k$$

- Každý datový bod je nyní reprezentován k-dimenzionálním vektorem v spektrálním prostoru

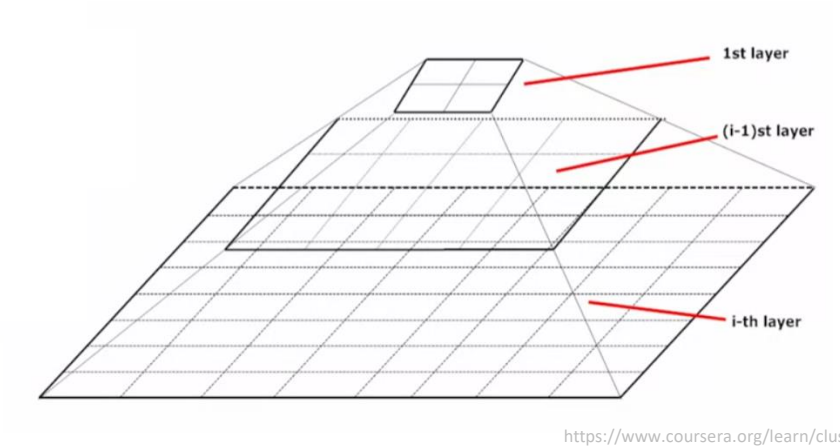
Spektrální shlukování

- Shlukování
  - Aplikace K-means algoritmu na transformovaná data
- Výhody
  - Shluky, které jsou nelineárně separovatelné v původním prostoru, mohou být lineárně separovatelné ve spektrálním prostoru
  - Možnost využití různých metrik podobnosti podle povahy dat
  - Efektivní pro datové struktury, kde tradiční metody selhávají
- Nevýhody
  - Vysoká výpočetní náročnost
  - Je nutné znát předem hodnotu k (jako u K-means)
    - Řeší další vylepšení (eigengap)

Část IV.:
Shlukování založené na mřížce

STING

- Statistical Information Grid
- Vytváříme mřížkovou strukturu
  - Pro každou buňku jsou spočítány statistiky
    - Počet bodů, průměr, min, max, typ rozdělení
  - Postupně vytváříme další vrstvy



https://www.coursera.org/learn/cluster-analysis

CLIQUE

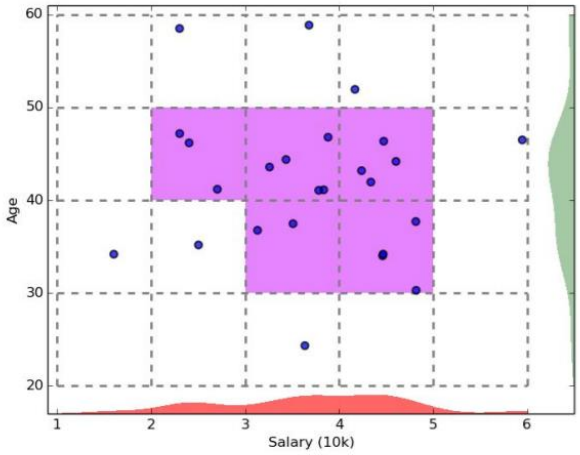
- Clustering in Quest

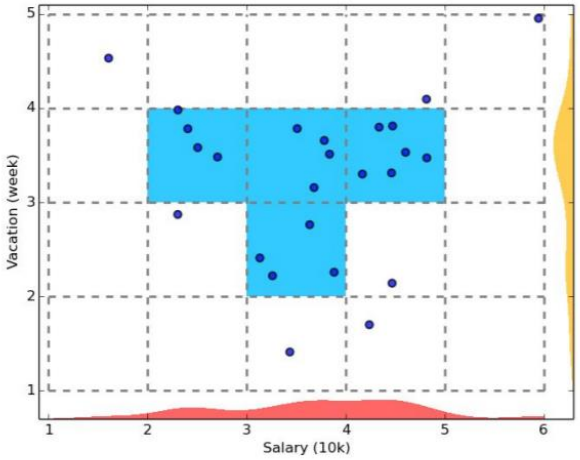
- Založený na mřížce
  - Rozděluje prostor mřížkou a počítá počet bodů v buňce
- Založený na hustotě
  - Shluk je vytvořen z husté množiny sousedících bodů
  - Za podmínky, že je počet bodů v buňce větší, než vstupní parametr modelu

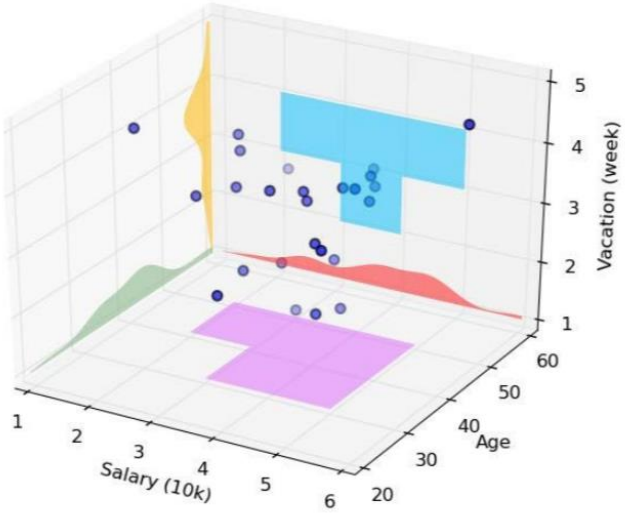
- Automaticky nalezne podprostor, který umožňuje lepší shlukování než původní prostor dat – založeno na Apriori principu

CLIQUE

- Bottom-up přístup







https://list01.biologie.ens.fr/wws/d_read/machine_learning/SubspaceClustering/CLIQUE_algorithm_grid-based_subspace_clustering.pdf

Část III.:
Vyhodnocení shlukování

Měření kvality shlukování

- Externí (Supervised)
  - Porovnání s předem daným požadovaným výsledkem např. na základě vzdálenosti
- Interní (Unsupervised)
  - Vyhodnocení správnosti shlukování na základě toho, jak dobře jsou shluky separované a kompaktní
  - Silhouette koeficient
- Relativní
  - Porovnání výsledků algoritmu s různým nastavením vstupních parametrů

Měření kvality shlukování

- Purity (čistota)
  - Měří zastoupení dominantních členů jednotlivých tříd ve shlucích (hodnota 0 až 1)
  - Pro shluk i:

$$purity_i = \frac{1}{n_i} \max_j \{nij\}$$

  - Pro celý výsledek

$$purity = \sum_{i=1}^{r} \frac{n_i}{n} purityi = \frac{1}{n} \sum_{i=1}^{r} \max_j \{nij\}$$

  - Např.: $purity_1 = \frac{30}{50}$; $purity_2 = \frac{20}{25}$; $purity_3 = \frac{25}{25}$;

$$purity = \frac{30 + 20 + 25}{100} = 0.75$$

  - Problém – výpočet čistoty je stejný u obou tabulek

C\T	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

C\T	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

https://www.coursera.org/learn/cluster-analysis

Měření kvality shlukování

- Maximum matching
  - Třída může patřit pouze jednomu shluku
  - Váha $w(e_{ij}) = n_{ij}; w(M) = \sum_{e \in M} w(e)$
  - Maximum weight matching

$$match = armax_M \{\frac{w(M)}{n}\}$$

  - Např.: zelená -> $match = purity = 0.75$
    oranžová ->
    1. Možnost -> $\frac{w_1(M)}{n} = \frac{30 + 5 + 25}{100} = 0.6$
    2. Možnost -> $\frac{w_2(M)}{n} = \frac{20 + 20 + 25}{100} = \mathbf{0.65}$

$$\boldsymbol{match = 0.65}$$

C\T	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100

C\T	T_1	T_2	T_3	Sum
C_1	0	30	20	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	50	25	100

https://www.coursera.org/learn/cluster-analysis

Užitečná literatura / kurzy

- How much can k-means be improved by using better initialization and repeats?
  - Článek k porovnání různých k-means přístupů (2019)