

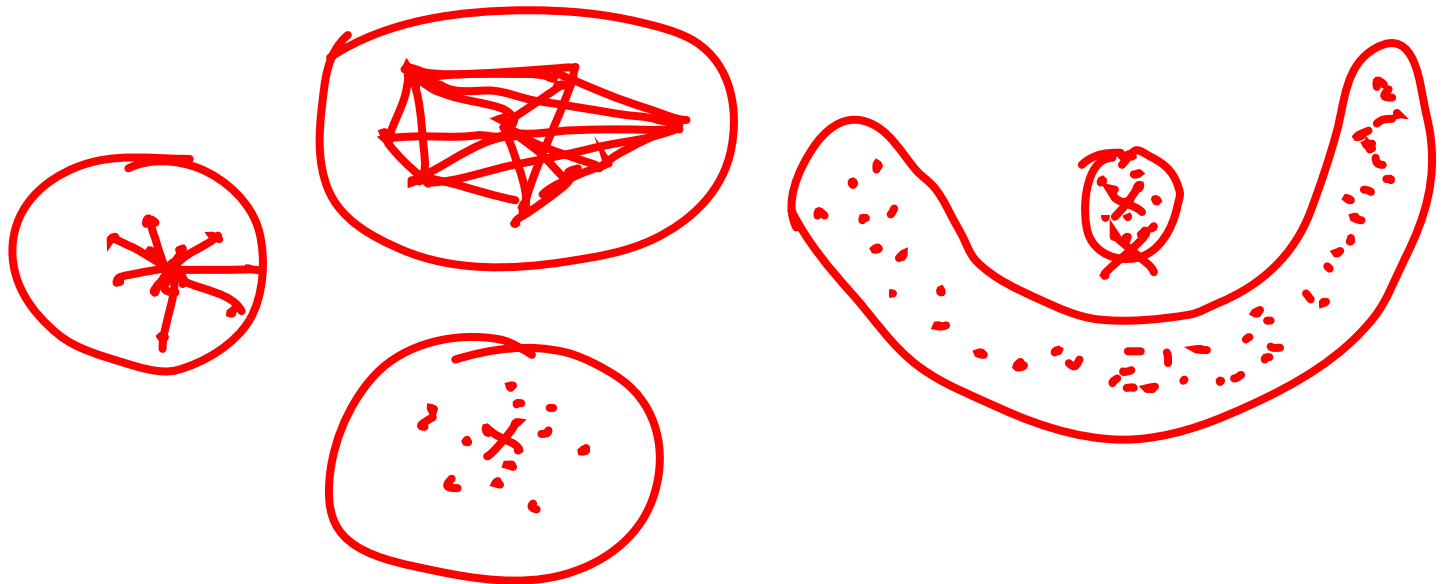
Shluková analýza

Úloha: Objekty (vektory) x_n , $n = 1, \dots, N$ chceme roztrdit do shluků S_k , $k = 1, \dots, K$ tak, aby byly pohromadě “blízke” objekty.

Možné kritérium:

$$\sum_k \sum_{x_n \in S_k} \sum_{x_j \in S_k} \|x_n - x_j\|^2.$$

Poznámka: Počet shluků považujeme za daný, ač jeho stanovení je velký problém.



1 Algoritmus k-means [MacQueen, 1967]

1. náhodně zvolíme středy shluků (centroids) c_k ,
2. každý objekt x_n přiřadíme ke shluku S_k , jehož střed je nejbližší,

$$k(n) \in \operatorname{argmin}_j \|x_n - c_j\|^2,$$

3. v každém shluku vypočteme nový střed jako těžiště prvků shluku,

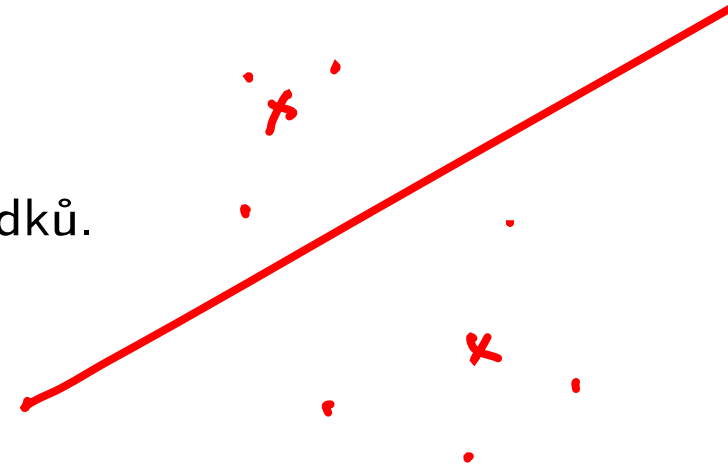
$$c_k := \frac{1}{|S_k|} \sum_{x_n \in S_k} x_n,$$

kde $|S_k|$ značí počet prvků shluku S_k ,

4. návrat na 2, pokud došlo k podstatné změně výsledků.

Klesá kritérium

$$\sum_k \sum_{x_n \in S_k} \|x_n - c_k\|^2$$



Problémy:

- Je-li nejbližších středů víc, vybereme z nich např. náhodně.
- Konverguje k **lokálnímu** extrému.
- Volba počátečních odhadů (např. některá z x_n , nesmí být stejná, doporučují se daleko...).
- Shluky potřebujeme neprázdné, raději zhruba stejně početné...
- Záleží na metrice (lze napřed všechny souřadnice normalizovat, ani to nemusí být dobré).

2 Algoritmus fuzzy c-means (FCM)

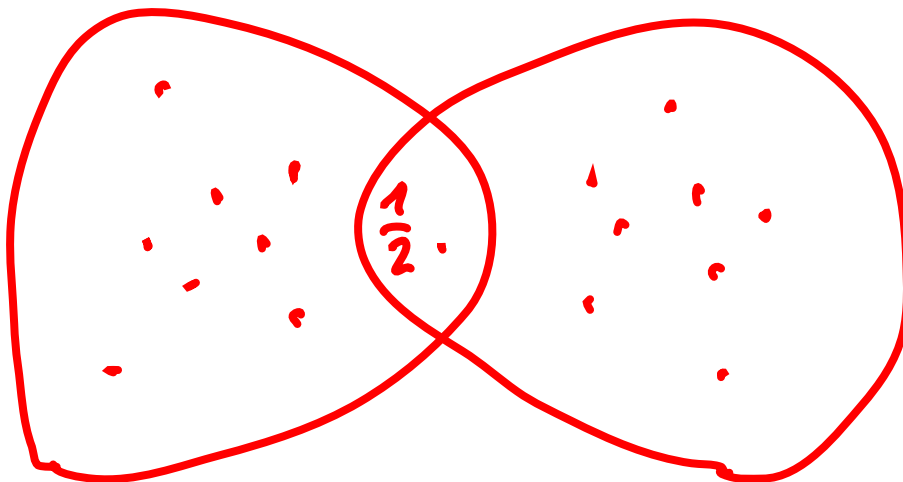
[Dunn 1973, vylepšil Bezdek 1981]

Shluky S_k jsou fuzzy množiny, charakterizovány maticí stupňů příslušnosti

$$\alpha_{n,k} = \mu_{S_k}(x_n),$$

které splňují

$$\sum_k \alpha_{n,k} = 1, \quad \sum_n \alpha_{n,k} > 0.$$



1. Náhodně zvolíme středy shluků c_k .

2. Stupeň příslušnosti objektu x_n ke shluku S_k stanovíme jako

$$\alpha_{n,k} := \frac{\frac{1}{\|x_n - c_k\|^{\frac{2}{m-1}}}}{\sum_j \frac{1}{\|x_n - c_j\|^{\frac{2}{m-1}}}},$$

kde $m > 1$ (např. 2) je parametr metody. (Jmenovatel je normalizační faktor; nutno ošetřit dělení nulou.)

3. V každém shluku vypočteme nový střed jako těžiště prvků shluku vážených m -tou mocninou jejich stupně příslušnosti,

$$c_k := \frac{\sum_n \alpha_{n,k}^m x_n}{\sum_n \alpha_{n,k}^m}.$$

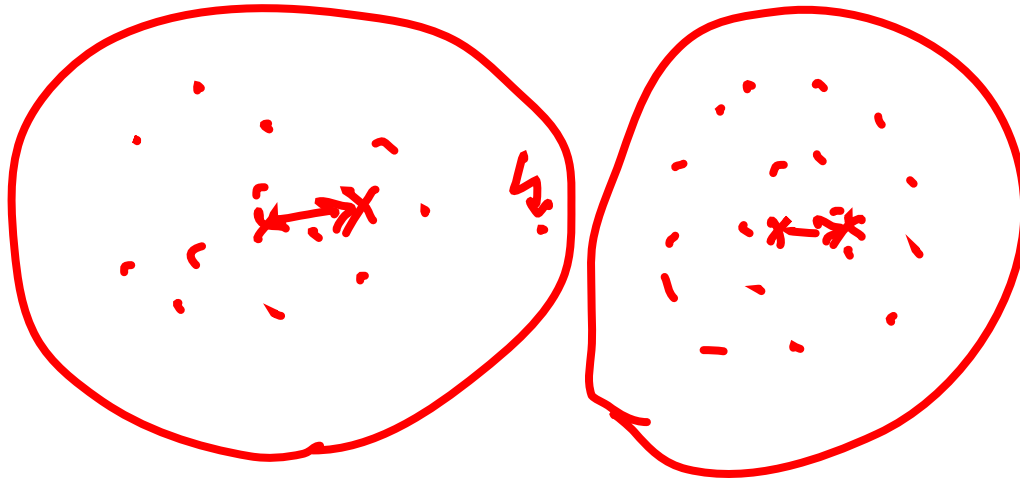
4. Návrat na 2, pokud došlo k podstatné změně výsledků.

Klesá kritérium

$$\sum_k \sum_{x_n \in S_k} \alpha_{n,k}^m \|x_n - c_k\|^2 .$$

Je-li to požadováno, konečný výsledek defuzzifikujeme (každý objekt zařadíme do toho shluku, k němuž má největší stupeň příslušnosti).

Problémy podobné, trochu menší; řada modifikací, např. vzdálenost vektorů (norma) může být obecnější.



3 Odhad parametrů směsi normálních rozdělání

Úloha: Na základě realizace náhodného výběru (x_1, \dots, x_N) (zde x_k jsou čísla) hledáme **maximálně věrohodný odhad** směsi normálních rozdělání se středními hodnotami c_k , $k = 1, \dots, K$, stejným rozptylem σ^2 a koeficienty směsi (váhami) q_k , $k = 1, \dots, K$.

Směs má hustotu

$$f(t) = \sum_k q_k f_{N(c_k, \sigma^2)}(t),$$

věrohodnost je

$$q_1, \dots, q_K, c_1, \dots, c_K, \sigma_1, \dots, \sigma_K$$

$$L(\underline{x}) = \prod_n \sum_k q_k f_{N(c_k, \sigma^2)}(x_n)$$

a její logaritmus

$$\ln(a+b) = \ln(a) \oplus \ln(b) \quad ?$$

$$\ell(\underline{x}) = \sum_n \ln \sum_k q_k f_{N(c_k, \sigma^2)}(x_n).$$

To se těžko řeší přímo, ale je zde iterační metoda:

4 EM algoritmus

EM (Expectation-Maximization) [Dempster, Laird, and Rubin 1977, M.I. Schlesinger 1968, US Army ~[1950]

Příslušnost x_n ke k -té složce směsi (shluku) popíšeme koeficientem $\alpha_{n,k} \in \langle 0, 1 \rangle$; dostaneme matici, jejíž koeficienty splňují

$$\sum_k \alpha_{n,k} = 1, \quad \sum_n \alpha_{n,k} > 0.$$

1. Náhodně zvolíme střední hodnoty shluků c_k .

E. Stanovíme koeficienty

$$\alpha_{n,k} := \frac{q_k f_{\mathbf{N}(c_k, \sigma^2)}(x_n)}{\sum_j q_j f_{\mathbf{N}(c_j, \sigma^2)}(x_n)}.$$

(Jmenovatel je normalizační faktor.)

M. Aktualizujeme váhu shluku

$$q_k := \frac{\sum_n \alpha_{n,k}}{\sum_j \sum_n \alpha_{n,j}} = \frac{1}{N} \sum_n \alpha_{n,k}$$

a jeho střed jako těžiště prvků vážených stupněm příslušnosti,

$$c_k := \frac{\sum_n \alpha_{n,k} x_n}{\sum_n \alpha_{n,k}} = \frac{\sum_n \alpha_{n,k} x_n}{N q_k}.$$

2. Opakujeme EM, pokud došlo k podstatné změně výsledků.

Věta: V průběhu EM algoritmu *věrohodnost neklesá*.

Toto je jen velmi speciální ukázka EM algoritmu; rozšíření na více dimenzí je snadné.

Lze jím hledat maximálně věrohodné odhady dalších parametrů rozdělení.

Ale **pozor:** Pro rozptyl $\sigma_k^2 \rightarrow 0$ dostaneme $\ell(\mathbf{x}) \rightarrow \infty$ a hledali bychom maximum, které neexistuje \implies rozptyl je nutno zdola omezit!

Použití pro parametry směsí rozdělení je typické, ne však jediné možné.



Opět jsou problémy s uvíznutím v lokálním extrému apod., nicméně se značně rozšiřují možnosti použití metody maximální věrohodnosti.

