1 Několik myšlenek o shlukové analýze

Úloha: Objekty (vektory) $x_n,\ n = 1, ..., N$ chceme roztřídit do shluků $S_k,\ k = 1, ..., K$ tak, aby byly pohromadě "blízké" objekty.

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

2 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,

$$k(n) = \arg\min_j \left\| x_n - c_j \right\|^2$$

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:

- končí v **lokálním** extrému
- volba počátečních odhadů (např. některá z x_n, doporučují se daleko, ...)
- shluky potřebujeme neprázdné, raději zhruba stejně početné, ...
- záleží na metrice (lze napřed všechny souřadnice normalizovat, ani to nemusí být dobré)

3 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 $m > 1$ (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

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

Klesá kritérium

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

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

4 Odhad parametrů směsi normálních rozdělení

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

Směs má hustotu

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

věrohodnost je

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

a její logaritmus

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

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

5 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

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

(jmenovatel je normalizační faktor)

M. aktualizujeme váhu shluku

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

a jeho střed jako těžiště prvků vážených stupněm příslušnosti,

$$c_k := \frac{\sum\limits_n \alpha_{n,k} x_n}{\sum\limits_n \alpha_{n,k}} = \frac{\sum\limits_n \alpha_{n,k} x_n}{N\, q_k}$$

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.

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.