Primena algoritma maksimizacije očekivanja u klasterovanju

MILICA KOJIČIĆ
ANA ĐORĐEVIĆ

Klasterovanje k-sredina

- Određivanje k-centara tako da se minimizuje deformisanost kvadratne greške

$$\text{DISTORTION}(Data, Centers) = \frac{1}{n} \sum_{\text{all points } DataPoint \text{ in } Data} d(DataPoint, Centers)^2$$

- Centar gravitacije - tačka čija je i-ta koordinata prosečna vrednost i-tih koordinata svih tačaka skupa

Lojdov algoritam

- Najčešće korišćena heuristika klasterovanja k-sredina
- Proizvoljno se bira k tačaka iz skupa podataka
- Iterativno ponavljanje 2 koraka:
- **Korak od centara ka novim klaterima** – kada su centri odabrani, pridružiti svaku tačku klasteru koji je određen najbližim centrom
- **Korak od klastera ka novim centrima** – kada su tačke dodeljene, izabrati centar gravitacije kao novi centar klastera

EM i Lojdov algoritam

- Svaka od n tačaka dimenzije m pripada nekom od k klastera i tu pripadnost definišemo n-dimenzionalnim vektorom
- $HiddenVector = (HiddenVector_1, ..., HiddenVector_n)$
- Svaka koordinata skrivenog vektora uzima vrednost od 0 do 1
- Predstavljamo centre kao k tačaka m-dimenzionalnog prostora $(\theta_1, \theta_2, ..., \theta_k)$
- i nazivamo ih parametrima
- Skriveni vektor i parametri su nepoznati, Lojdov algoritam počinje slučajnim izborom parametara

EM I Lojdov algoritam

- Dva osnovna koraka:
- **Korak od centara ka novim klasterima**: (Data, ?, Parameters) --> HiddenVector
- **Korak od klastera ka novim centrima:** (Data, HiddenVector, ?) **-->** Parameters
- Različiti kriterijumi zaustavljanja algoritma

Primena maksimizacije očekivanja na Lojdov algoritam

- Algoritam započinje slučajnim izborom centara
- **Korak od centara ka novim klasterima (E - step):**
- Na osnovu k-centara $(x_1, \ldots, x_k)$ i n tačaka podataka, konstruišemo k x n matricu pri čemu vrednost $HiddenMatrix_{i,j}$ predstavlja odgovornost da tačka j pripada klasteru i. Ova odgovornost se računa u skladu sa Njutnovim zakonom gravitacije:

$$HiddenMatrix_{i,j} = \frac{1/d(Data_j, x_i)^2}{\sum_{\text{all centers } x_i} 1/d(Data_j, x_i)^2}$$

Primena maksimizacije očekivanja na Lojdov algoritam

- U praksi se češće koristi particiona funkcija:

$$HiddenMatrix_{i,j} = \frac{e^{-\beta \cdot d(Data_j, x_i)}}{\sum_{\text{all centers } x_i} e^{-\beta \cdot d(Data_j, x_i)}}$$

- gde β predstavlja parametar mekoće.

Primena maksimizacije očekivanja na Lojdov algoritam

- **Korak od klastera ka novim centrima (M - step):**
- Ukoliko sa $HiddenMatrix_i$ označimo i-ti red skrivene matrice, tada vršimo ažuriranje centra x_i Za centar x_i njegovu j-tu koordinatu računamo kao:
- $$x_{i,j} = \frac{HiddenMatrix_i \cdot Data^j}{HiddenMatrix_i \cdot \vec{1}}$$
- Ažurirani centar se naziva težinski centar gravitacije

Primene u bioinformatici

- **Klasterovanje genskih ekspresija u cilju detektovanja kancera debelog creva**
- Kao relevantan podataka uzima se prosečan intenzitet ekspresije u tumornom i normalnom tkivu, čija je vrednost računski određena na osnovu svih intenziteta ekspresija
- **Klasterovanje gena kvasca**
- Istražuje se na kojim genima kvasca dolazi do promene kako bi proces fermentacije bio moguć