Mesterséges intelligencia II.

Felügyelet nélküli tanulás

A $K-means$ klaszterező

Adott $x_1, x_2, \ldots, x_n \in X$ n-darab minta. A feladat ezen minták k db csoportba, klaszterbe sorolása. A klaszterezés a klaszterközéppontok meghatározásával történik és minden pont a hozzá legközelebbi középponthoz tartozó klaszterbe kerül. Ehhez definiálnunk kell egy távolságfüggvényt a minták felett.

Jelölések:
$\mu_\mathbf{i}$: az i-edik klaszter középpontja.
$\mathbf{r_{ij}} = 1$, ha az i-edik klaszterbe soroljuk a j-edik mintát; 0, egyébként.

Ezek alapján fölírhatunk egy hibafüggvényt:

$$Err = \sum_i \sum_j r_{ij} ||\mu_i - x_j||^2 \rightarrow min$$

Amylet minimalizálni szeretnénk, azaz az i-edik klaszterbe tartozó minták távolságnégyzetösszege az i-edik középponttól minimális legyen. Az optimum ott van, ahol az $Err\mu$ szerinti deriváltja 0, azaz ha a klaszterközéppontok egybeesnek a klaszterekhez tartozó pontokal.

Algoritmus:

1. inicializáljuk a μ_j klaszzerközéppontokat (pl. egyenletes eloszlással a térben)
2. legyen $r_{ij} = 1$ ha az x_i a μ_j-hez van legközelebb, egyébként 0. (minták középponthoz rendelése)
3. legyen $\mu_j = \frac{\sum_i r_{ij} x_i}{\sum_i r_{ij}}$ minden $j = 1, 2, \ldots, k$-ra. (középpontok újradefiniálása)
4. ismételjük a 2. ponttól, amíg változik valamelyik klaszterközéppont.

Problémák a $k-means$-el:

1. ha egy csoportra két közepet inicializálunk (túl nagy k válazstása)
2. ha két csoport közé inicializálunk egy közepet (túl kicsi k választása)

Példa applet: http://www.rob.cs.tu-bs.de/content/04-teaching/06-interactive/Kmeans/Kmeans.html
Következmény: helyesen kell megválasztani a közepeket... $(X - means)$

Az $X - means$ klaszterező

Az $X - means$ algoritmus abban különbözik a $K - means$-től, hogy nem egy konkrét k értéket definiálunk, hanem egy alsó és felső korlátot. Az algoritmus a minimális k érték beállításával futtatja a $K - means$ klaszterezőt, majd növeli a k érétkét, amíg nem romlik a négyzetes hiba.

Pontnövelési stratégiák:

Egy középpont kettéosztása: valamilyen heurisztika alapján válasszunk egy középpontot és azt osszuk ketté.

A középpontok felének a kettéosztása: valamilyen heurisztika alapján válasszuk ki a középpontok felét, majd azokat osszuk ketté.

(Egy lehetséges heurisztika a megfelelő középpont választására: válasszuk azt ahol a középponthoz tartozó négyzetes eltérés maximális.)

A középpontok számának dinamikus növelése: Egy harmadik lehetséges módszert D. Pelleg és A. Moore írt le egy 2000-es cikkükben. Az alapötlet, hogy minden közepet osszunk szét két részre, majd futtasuk a $k - means$ algoritmust $k = 2$-re lokálisan. Ezután meg kell vizsgálin, melyik kettéosztás vezet a helyes eredményre. Ezt a vizsgálatot egy úgynevezett BIC (Bayesian Information Criteria) érték alapján végzik. Kiszámolják az egy középhez és a kettéosztott középhez tartozó BIC értéket, majd azt a modellt választják, amelynél ez az érték nagyobb.