K:n keskiarvon ryvästysalgoritmin pysähtyminen

Väite. *Algoritmi K-means pysähtyy kustannusfunktion $wc(\mathcal{C})$ lokaaliin minimiin.*

Todistus.

Todistamme väitteen algoritmin versiolle, jossa ryppään edustajavektoriksi valitaan ryppään keskiarvo ja etäisyysfunktiona käytetään Euklidisen etäisyyden neliötä $d(x,z) = \sum_{j=1}^{d}(x_j - z_j)^2$.

Algoritmin pysähtymiseksi riittää osoittaa, että

1. Erilaisia ryvästyksiä on äärellinen määrä
2. Algoritmi vaihtaa joka iteraatiossa ryvästyksen toiseen tai pysähtyy
3. Algoritmi valitsee kunkin ryvästyksen korkeintaan kerran suorituksen aikana

Kohta (1): Erilaisia ryvästyksiä on $O(K^n)$ kappaletta (kukin esimerkki voidaan sijoittaa K ryppääseen toisistaan riippumatta). Kohta (2): Selvä pseudokoodin perusteella.

Kohdan (3) osoittamiseksi todetaan ensin, että ryvästyksen kustannus voi ainoastaan vähetä suorituksen aikana. Kustannusfunktion arvo muuttuu kahdessa kohdin: valittaessa kullekin esimerkille lähinnä oleva rypäs ja laskettaessa ryppäiden edustajavektorit uudelleen.

Tarkastellaan ensin ryppään edustajavektorin laskenta-askelta. Ryppään edustajavektorin valinta ei vaikuta muiden ryppäiden sisäiseen vaihteluun, joten voidaan keskittyä tarkastelemaan yhtä rypästä C, jonka edustajavektori r halutaan laskea.

Tarkastellaan ryppään sisäistä vaihtelua

$$wc(C) = \sum_{x \in C} \sum_{j=1}^{d} (x_j - r_j)^2$$

keskipisteen r funktiona. $wc(C)$ on on jatkuva usean muuttujan funktio, jonka määrittelyalue on R^d. Funktion minimi on lokaali ääriarvokohta, jossa funktion osittaisderivaatat keskipisteen r komponenttien r_j suhteen saavat arvon nolla

$$\frac{\partial wc(C)}{\partial r_j} = 0, j = 1, \ldots, d.$$

Lasketaan osittaisderivaatat komponenttien suhteen käyttäen tunnettuja deri-

vointisääntöjä

$$\begin{aligned}\frac{\partial wc(C)}{\partial r_j} &= \frac{\partial}{\partial r_j}\sum_{x\in C}\sum_{l=1}^{d}(x_l - r_l)^2 \\ &= \sum_{x\in C}\sum_{l=1}^{d}\frac{\partial}{\partial r_j}(x_l - r_l)^2 \\ &= \sum_{x\in C}\frac{\partial}{\partial r_j}(x_j - r_j)^2 \\ &= \sum_{x\in C} 2(x_j - r_j)\cdot(-1) = 0\end{aligned} \tag{1}$$

eli saadaan

$$-2(\sum_{x\in C} x_j - |C|r_j) = 0$$

josta ratkaistaan $r_j = \frac{1}{|C|}x_j$. Toistamalla sama kaikille $j = 1, \ldots, d$ saadaan parhaaksi edustajavektoriksi $r = \frac{1}{|C|}\sum_{x\in C} x$ eli ryppään pisteiden keskiarvo.

Osoitimme siis, että ryppään keskiarvon valinta edustajavektoriksi minimoi ryppään sisäisen vaihtelun. Ryppään edustajavektorin valinta voi siis ainoastaan vähentää ryvästyksen kustannusta.

Tarkastellaan seuraavaksi mielivaltaista esimerkkiä $x \in C_i$ ryvästyksessä $\mathcal{C}_{t_0}$, jonka algoritmi on valinnut askeleessa t_0 ja ryppäille on laskettu uudet keskipisteet $r_k, k = 1, \ldots, K$.

Jos ryvästyksessä $\mathcal{C}_{t_0+1}$ esimerkki x on ryppäässä $C_j, j \neq i$, jompikumpi seuraavista on voimassa

- $d(x, r_j) < d(x, r_i)$, jolloin siirtyminen pienentää ryppäiden yhteenlaskettua sisäistä vaihtelua $wc(C_i) + wc(C_j)$, tai
- $d(x, r_j) = d(x, r_i)$ ja $j = \mathbf{argmin}_l d(x, r_l)$, jolloin yhteenlaskettu vaihtelu pysyy ennallaan ja x siirtyy indeksiltään pienimpään lähimpään ryppääseen C_j. (Tässä oletetaan, että tasatilanteen sattuessa valitaan indeksiltään pienin ryväs)

Ryvästyksen kustannus voi siis ainastaan laskea siirtymisen seurauksena. Algoritmi ei siis missään vaiheessa valitse ryvästystä, jonka kustannus olisi suurempi kuin edellisen ryvästyksen.

Ylläolevasta seuraa, että jos esimerkki x palaa ryppääseen C_i ryvästyksessä $\mathcal{C}_t$ ryvästyksen kustannus väistämättä pienenee. Koska ryvästyksien kustannus ei ole voinut nousta välillä $[t_0, \ldots, t-1]$, seuraa että $wc(\mathcal{C}_t) < wc(\mathcal{C}_{t_0})$ eli $\mathcal{C}_{t_0} \neq \mathcal{C}_t$.

Algoritmi ei siis palaa kerran hylkäämäänsä ryvästykseen.

Se, että pysähtyminen on lokaali minimi seuraa siitä, että algoritmin suoritus jatkuu niin kauan kuin klusteroinnissa tapahtuu muutoksia ja siitä, että joka askeleessa valitaan lokaalisti paras muutos

- Paras edustajavektori, kun ryppään esimerkit on kiinnitetty
- Paras rypäs, kun edustajavektori on kiinnitetty