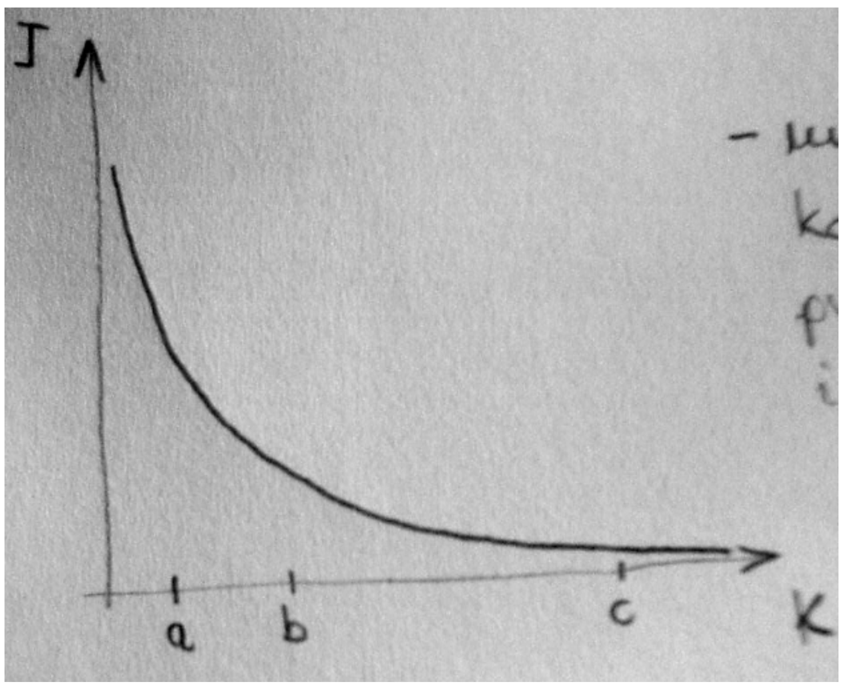
Rješenje zadatka 5.1 predmeta Strojno učenje

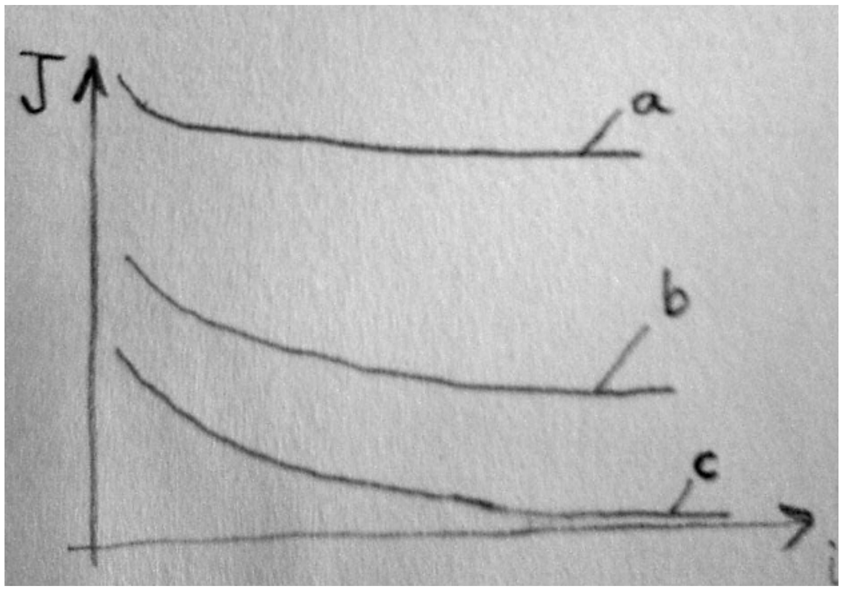
Siniša Biđin

5. veljače 2013.

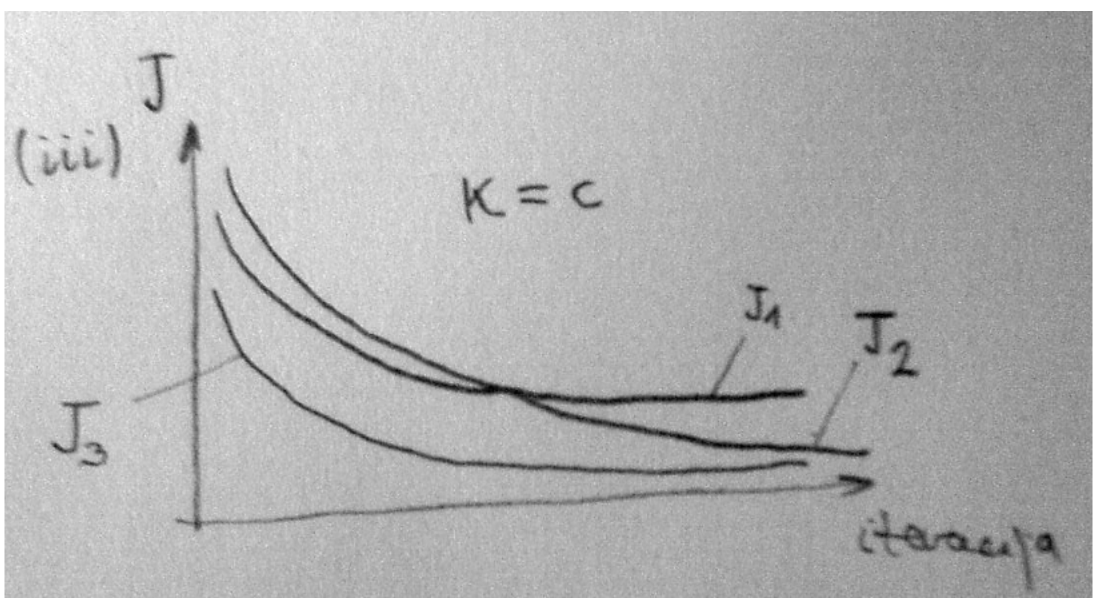
(a) (i) Minimalna vrijednost J je 0, za slučaj kada je broj grupa jednak broju primjera, a svaki je primjer ujedno i centroid.



(ii) Odabrane su tri vrijednosti K: a, b i c, označene na prethodnoj slici. Za svaki K skiciramo vrijednost kriterija pogreške u ovisnosti o broju iteracija algoritma.



(iii) Krivulja J_3 je najvjerojatnija, ukoliko koristimo algoritam k-*means*++.



(b) (i) Početni centroidi:

$$\mu_1 = (7,1)$$
$$\mu_2 = (1,4)$$
$$\mu_3 = (2,8)$$

Centroidima (grupama, $b^{(i)}$), pridružujemo najbliže primjere. Centroidu μ_1 pridružujemo primjere a, b, d; μ_2 pridružujemo g i c; μ_3 pridružujemo e i f. Zatim računamo nove centroide:

$$\mu_1 = \frac{a+b+d}{3} = \frac{(18,5)}{3} = (6, \frac{5}{3})$$
$$\mu_2 = \frac{c+g}{2} = \frac{(1,8)}{2} = (\frac{1}{2}, 4)$$
$$\mu_3 = \frac{e+f}{2} = \frac{(5,14)}{2} = (\frac{5}{2}, 7)$$

Time je gotov prvi korak algoritma. Koraci se ponavljaju sve do konvergencije vrijednosti centroida.

(ii) Početnim središtima

$$\mu_1 = b,\ \mu_2 = c,\ \mu_3 = e$$

pridružujemo najbliže primjere. Središtu μ_1 primjere a i d, središtu μ_2 primjer g, te središtu μ_3 f. Zatim, primjer d postaje nova vrijednost središta μ_1, jer vrijedi:

$$\nu(a,d) + \nu(b,d) < \nu(d,a) + \nu(b,a)$$
$$\nu(a,d) + \nu(b,d) < \nu(a,b) + \nu(d,b)$$

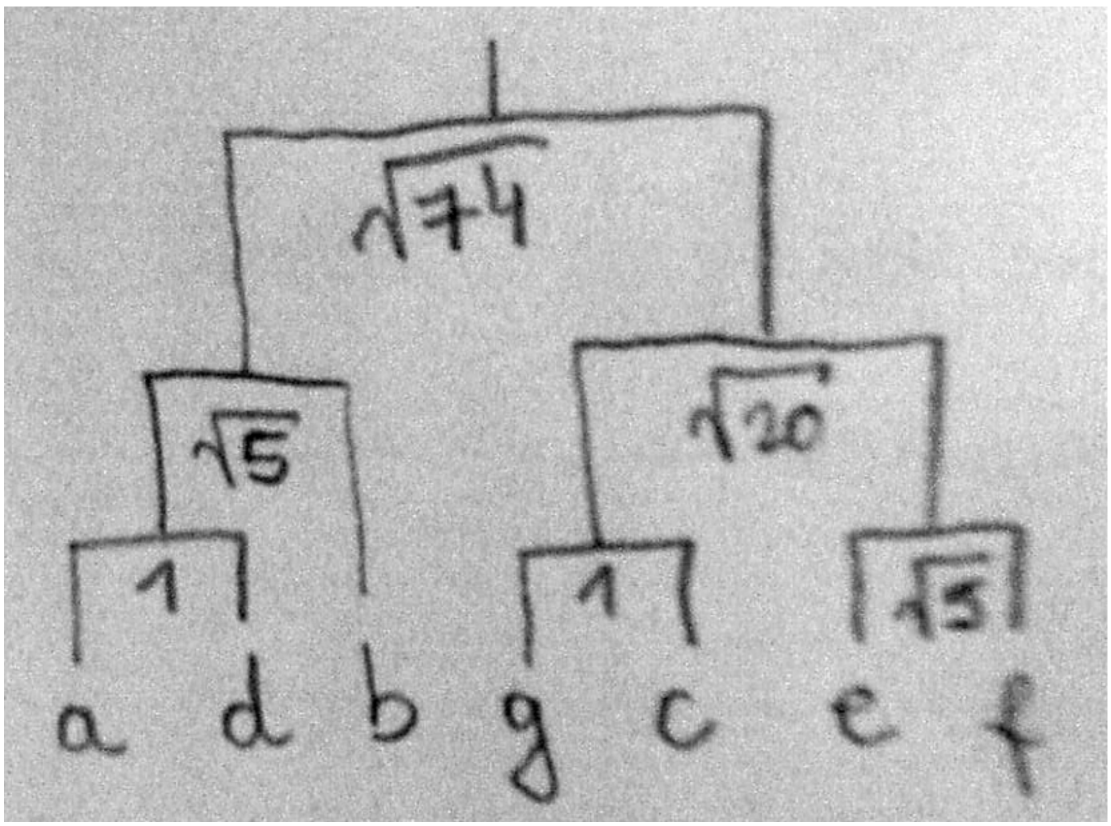
Ostala središta (μ_2 i μ_3) se ne mijenjaju, stoga vrijedi

$$\mu_1 = d,\ \mu_2 = c,\ \mu_3 = e.$$

Dok je vremenska složenost algoritma k-*means* $\mathcal{O}(TnNK)$, vremenska složenost k-medoida je $\mathcal{O}(TK(N-K)^2)$ i najveći je nedostatak algoritma. Do razlike u složenosti dolazi jer k-medoida u svakom koraku računa zbroj mjera ν između svakog para $(N-K)$ primjera.

(c)

$$
\begin{aligned}
d(a,d) &= 1 \\
d(g,c) &= 1 \\
d(ad,b) = d(a,b) &= \sqrt{5} \\
d(e,f) &= \sqrt{5} \\
d(cg,ef) = d(g,e) &= \sqrt{20} \\
d(abd,cegf) = d(b,e) &= \sqrt{74}
\end{aligned}
$$



(d) (i) Koristio bih algoritam k-medoida jer je informacija o sličnosti svih parova dana unaprijed putem rang-liste poželjnih ljudi. Za sličnost svakog para uzeo bih srednju vrijednost poželjnosti jednog člana za drugim i obrnuto. Na primjer, ako Siniša na svojoj rang-listi Vladimira navodi na prvom mjestu, a Vladimir Sinišu na petom, sličnost bi mogla biti $v(\text{Siniša}, \text{Vladimir}) = \frac{0+4}{2} = 2$. Mogli bismo i posebno penalizirati veće udaljenosti na rang-listama, na primjer kvadratnom udaljenošću.

(ii) Ne dobivamo informaciju o sličnosti, već sami moramo odrediti koji su gosti međusobno bliski ili ne. Na temelju danih odgovora na pitanja i važnosti pojedinih pitanja svakog bih gosta smjestio na točku u 10-dimenzijskom prostoru, pa zatim grupirao putem k-*means* algoritma. Grupe bi sadržavale goste s međusobno najsličnijim odgovorima.

(e) (i) Grupirao bih zajedno označene i neoznačene primjere te varirao K. Za svaki K, provjeravao bih nastalu grešku na označenim primjerima. Odabrao bih onaj K kod kojeg je greška u grupiranju na označenim primjerima najmanja.

(ii) Grafički bih prikazao ovisnost kriterijske funkcije o broju grupa i tražio "koljeno" krivulje. Za dovoljno velik K, kriterijska funkcija prestaje snažno padati jer algoritam kreće razdjeljivati prirodne grupe. Uzeo bih taj K kao najbolju procjenu broja prirodnih grupa.