

Strojno učenje – domaća zadaća 12

UNIZG FER, ak. god. 2016./2017.

Zadano: 20. 1. 2017.

Napomena: Zadatke možete rješavati samostalno ili u grupi. Ako zadatke rješavate u grupi, pobrinite se da svi članovi grupe pridonose rješenju i da ga naposlijetku svi razumiju. Po potrebi konzultirajte sve dostupne izvore informacija. Rješenja zadataka ponesite na iduće auditorne vježbe. Zabilježite sve nejasnoće i nedoumice, kako bismo ih prodiskutirali.

1. [Svrha: Razumjeti rad algoritma k-srednjih vrijednosti u smislu minimizacije kriterija pogreške. Razumjeti kako rad algoritma ovisi o broju grupa K i odabiru početnih središta.]

Algoritam k-srednjih vrijednosti minimizira kriterij pogreške $J(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K | \mathcal{D})$. Vrijednost tog kriterija ovisi o broju grupa K , koji je unaprijed postavljen, te o položajima središta, koja se mijenjaju kroz iteracije.

- (a) Nacrtajte skicu vrijednosti kriterija pogreške J kao funkcije broja grupa K . Koja je minimalna vrijednost funkcije J i zašto?
(b) Izaberite na skici iz zadatka (a) tri vrijednosti za K i skicirajte na jednom grafikonu vrijednost kriterija pogreške J kao funkcije broja iteracija (tri krivulje).
(c) Izaberite na skici iz zadatka (a) jednu vrijednost za K . Skicirajte na jednom grafikonu vrijednosti kriterija pogreške J kao funkcije broja iteracija, ali ovaj put uzbir stohastičnost uslijed slučajnog odabira početnih središta (nacrtajte nekoliko mogućih krivulja na istom grafikonu). Koje od tih krivulja su izglednije za algoritam k-means++?
2. [Svrha: Isprobati rad algoritma k-srednjih vrijednosti i k-medoida na konkretnom primjeru. Shvatiti da je složenost ovog drugog puno nepovoljnija.] Raspolažemo skupom neoznačenih primjera:

$$\mathcal{D} = \{a = (5, 2), b = (7, 1), c = (1, 4), d = (6, 2), e = (2, 8), f = (3, 6), g = (0, 4)\}.$$

- (a) Izvedite jedan korak algoritma k-srednjih vrijednosti uz $K = 3$. Za početna središta odaberite $\boldsymbol{\mu}_1 = b$, $\boldsymbol{\mu}_2 = c$ i $\boldsymbol{\mu}_3 = e$.
(b) Izvedite jedan korak algoritma k-medoida uz $K = 3$. Za početna središta odaberite primjere b , c i e .
(c) Usporedite računalnu složenost algoritma k-srednjih vrijednosti i k-medoida.
3. [Svrha: Razumjeti model miješane gustoće i razlog zašto maksimizacija log-izglednosti nije analitički rješiva. Razumjeti kako uvođenje latentnih varijabli rješava taj problem. Razumjeti, na općenitoj razini, E-korak i M-korak. Razumjeti rad algoritma kao maksimizacije log-izglednosti i razumjeti kako ishod

ovisi o broju grupa i početnoj inicijalizaciji.] Algoritam maksimizacije očekivanja (EM-algoritam), kada se koristi za grupiranje, zapravo je poopćenje algoritma k-srednjih vrijednosti.

- (a) Što je prednost, a što nedostatak, algoritma maksimizacije očekivanja u odnosu na algoritam k-srednjih vrijednosti?
 - (b) Napišite izraz za gustoću $p(\mathbf{x})$ za model miješane gustoće (bez latentnih varijabli) i izraz za pripadnu (nepotpunu) log-izglednost.
 - (c) Napišite izraz za mješavinu s latentnim varijablama i izvedite izraz za (potpunu) log-izglednost tog modela. Možemo li dalje raditi izravno s tom log-izglednošću? Zašto?
 - (d) Definirajte E-korak i M-korak algoritma maksimizacije očekivanja primjenjene na Gaussovou mješavinu.
 - (e) Skicirajte vrijednost log-izglednosti $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ modela Gaussove mješavine kao funkcije broja iteracija, i to za tri različite vrijednosti parametra K (broj grupa): $K = 1$, $K = 10$ i $K = 100$. Na istom grafikonu skicirajte krivulju za $K = 10$ kada se za inicijalizaciju središta koristi algoritam k-srednjih vrijednosti.
4. [Svrha: Isprobati rad algoritma hijerarhijskog aglomerativnog grupiranja (HAC) na konkretnom primjeru, za slučaj kada primjeri nisu vektori. Uočiti razliku između udaljenosti i sličnosti te razliku između jednostrukih i potpunih povezanih.] Jednako kao i algoritam k-medoida, algoritam hijerarhijskog algomerativnog grupiranja može se primijeniti u slučajevima kada primjeri nisu prikazani kao vektori značajki te kada umjesto mjere udaljenosti između vektora raspoložemo općenitjom mjerom sličnosti (ili različitosti). Neka je *sličnost* primjera iz \mathcal{D} definirana sljedećom matricom sličnosti:
- $$S = \begin{pmatrix} & a & b & c & d & e \\ a & 1.00 & 0.26 & 0.15 & 0.20 & 0.17 \\ b & 0.26 & 1.00 & 0.24 & 0.31 & 0.31 \\ c & 0.15 & 0.24 & 1.00 & 0.20 & 0.50 \\ d & 0.20 & 0.31 & 0.20 & 1.00 & 0.24 \\ e & 0.17 & 0.31 & 0.50 & 0.24 & 1.00 \end{pmatrix}$$
- (a) Izgradite dendrogram uporabom jednostrukog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presjekli taj dendrogram?
 - (b) Izgradite dendrogram uporabom potpunog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presljekli taj dendrogram?
5. [Svrha: Razumjeti kako se unutarnji kriterij algoritma grupiranja može (pokušati) upotrijebiti za provjeru grupiranja (odabir optimalnog broja grupa). Razumjeti da Akaikeov kriterij u stvari oponaša regulariziranu funkciju pogreške, koja pak aproksimira pogrešku generalizacije. Isprobati izračun Randovog indeksa na konkretnom primjeru.] Nedostatak svih algoritama grupiranja koje smo razmotrili jest što se broj grupa K mora zadati unaprijed. Osim u rijetkim slučajevima kada nam je taj broj unaprijed poznat, to predstavlja problem.
- (a) Skicirajte krivulju log-izglednosti kod EM-algoritma kao funkciju broja grupa K . Obrazložite izgled krivulje. Možete li temeljem ove krivulje odrediti optimalan broj grupa? Kako?

- (b) Optimizacija broja grupa K može se provesti nekim kriterijem koji kombinira funkciju pogreške (odnosno log-izglednost) i složenost modela. Takav kriterij odgovara strukturnome riziku modela, koji je minimalan za optimalan broj grupa. Jedan takav kriterij jest Akaikeov informacijski kriterij (AIC):

$$K^* = \operatorname{argmin}_K (-2 \ln \mathcal{L}(K) + 2q(K)) \quad (1)$$

gdje je $-\ln \mathcal{L}(K)$ negativna log-izglednost podataka za K grupa, a $q(K)$ je broj parametara modela s K grupa.

Prepostavite da podatci \mathcal{D} u stvarnosti dolaze iz $K = 5$ grupa. Podatke grupiramo dvjema varijantama EM-algoritma: standardni algoritam i preinačeni algoritam s dijeljenom kovarijacijskom matricom (zajednička kovarijacijska matrica procijenjena nad čitavim skupom primjera \mathcal{D} na početku izvođenja algoritma). Skicirajte za ta dva algoritma funkciju koju minimizira Akaikeov minimizacijski kriterij.

- (c) Kada su primjeri ili podskup primjera označeni, kvaliteta grupiranja (uključivo i broj grupa K) može se procijeniti Randovim indeksom. Randov indeks zapravo izračunava točnost s kojom ćemo par jednako označenih primjera smjestiti u istu grupu, odnosno par različito označenih primjera u različitu grupu. Izračunajte Randov indeks za sljedeću pariticiju označenih primjera:

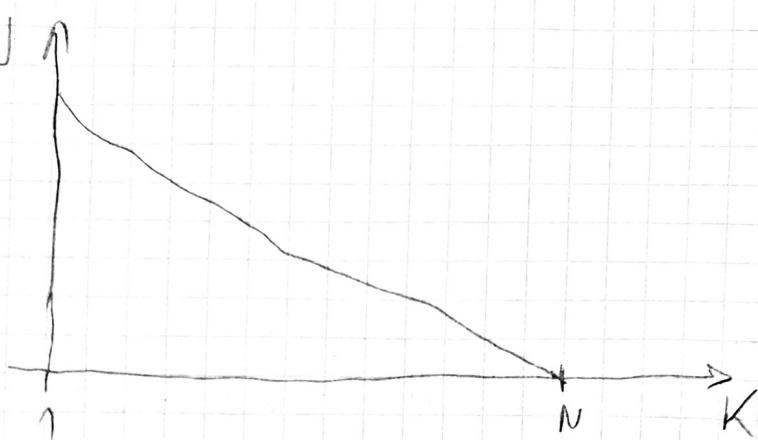
$$\{\{0, 0, 1, 2\}, \{1, 1\}, \{2, 2, 2, 1, 0\}\}.$$

- (d) Skicirajte vrijednost Randovog indeksa kao funkcije broja grupa K .
- (e) Randov indeks možemo koristiti samo ako su podatci označeni ili je podskup podataka označen. Međutim, čini se da to onda ujedno podrazumijeva da je unaprijed poznat broj grupa K . Imamo li koristi od Randovog indeksa čak i onda kada unaprijed znamo broj grupa? Možemo li ikako upotrijebiti Randov indeks, a da nam broj grupa nije unaprijed poznat?

12. DOMAĆA ZADAĆA

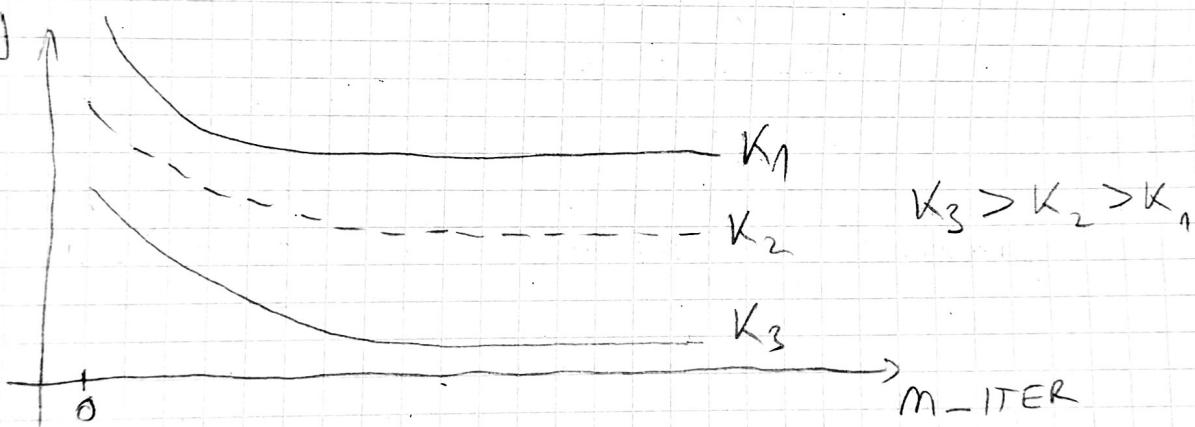
①

a) $J \uparrow$



MINIMALNA VRJEDNOST FUNKCIJE $J = \phi$, JER KADA $K=N$ (BROJ GRUPA = BROJ PRIMJERA), TADA SVAKI PRIMER PREDSTAVLJA JEDNU GRUPU U KOJOJ JE ON CENTROID PA NE POSTOJI POGREŠKA.

b) $J \uparrow$



OBJAŠNJENJE: FINALNA VRJEDNOST KRITERIJA J ĆE BITI VEĆA ZA MANJI BROJ GRUPA K ($K \uparrow$, SLOŽENOST \uparrow). SVE 3 VRJEDNOSTI NAKON NEKOG BROJA ITERACIJA POSTAJU KONSTANTNE JER ALGORITAM K -SREDnjIH VRJEDNOSTI KONVERGIRA.

c)



ZA k -MEANS++ JE NAJIZGLEDNIJA DODJA KRVULJA; IMA MALI POČETNI IZVOS KRITERIJA I BRZO KONVERGIRAJU.

(2)

$$\text{a) } \mu_1 = (7, 1), \mu_2 = (1, 4), \mu_3 = (2, 8)$$

1^o KORAK

$$x^{(1)} \rightarrow \|x_1 - \mu_1\| = \sqrt{(5-7)^2 + (2-1)^2} = 2.236$$

$$\|x_1 - \mu_2\| = 4.472$$

$$\|x_1 - \mu_3\| = 6.708$$

$$l = \arg \min_j \|x^{(1)} - \mu_j\| = 1 \rightarrow \underline{l_1^{(1)} = 1, l_2^{(1)} = 0, l_3^{(1)} = 0}$$

$$x^{(2)} \rightarrow x^{(2)} = c = \mu_1 \rightarrow \underline{l_1^{(2)} = 1, l_2^{(2)} = 0, l_3^{(2)} = 0}$$

$$x^{(3)} \rightarrow x^{(3)} = c = \mu_2 \rightarrow \underline{l_1^{(3)} = 0, l_2^{(3)} = 1, l_3^{(3)} = 0}$$

$$x^{(4)} \rightarrow \left. \begin{array}{l} \|x_4 - \mu_1\| = \sqrt{(6-7)^2 + (2-1)^2} = 1.4142 \\ \|x_4 - \mu_2\| = 5.3852 \\ \|x_4 - \mu_3\| = 7.211 \end{array} \right\} \arg \min = 1$$

$$\underline{l_1^{(4)} = 1, l_2^{(4)} = 0, l_3^{(4)} = 0}$$

$$x^{(5)} \rightarrow x^{(5)} = c = \mu_3 \rightarrow \underline{l_1^{(5)} = 0, l_2^{(5)} = 0, l_3^{(5)} = 1}$$

→

$$X^{(6)} \rightarrow \|X_6 - \mu_1\| = \sqrt{(3-7)^2 + (6-1)^2} = 6.403$$

$$\|X_6 - \mu_2\| = 2.8284$$

$$\|X_6 - \mu_3\| = 2.236$$

$$\underline{b_1^{(6)} = 0, b_2^{(6)} = 0, b_3^{(6)} = 1}$$

$$X^{(7)} \rightarrow \|X_7 - \mu_1\| = 7.616$$

$$\|X_7 - \mu_2\| = 1$$

$$\|X_7 - \mu_3\| = 4.472$$

$$\underline{b_1^{(7)} = 0, b_2^{(7)} = 1, b_3^{(7)} = 0}$$

NOVI CENTROIDI:

$$\mu_1 = \frac{\sum_{i=1}^N b_1^{(i)} X^{(i)}}{\sum_{i=1}^N b_i^{(i)}} = \frac{x_1 + x_2 + x_4}{3}$$

$$\mu_1 = \left(\frac{5+7+6}{3}, \frac{2+1+2}{3} \right) = (6, 1.667)_{//}$$

$$\mu_2 = \frac{x_3 + x_7}{2} = \left(\frac{1+0}{2}, \frac{4+4}{2} \right) = (2.5, 4)_{//}$$

$$\mu_3 = \frac{x_5 + x_6}{2} = \left(\frac{2+3}{2}, \frac{8+6}{2} \right) = (2.5, 7)_{//}$$

NE ULAZI U ISPIT! :

b) PRAĆUN VENTORA B JE ISTI KAO KOD
K - SREDJIH VRIJEDNOSTI.

U DRUGOM KORAKU PROTOTIPI GRUPA ODABIRU SE TAKO DA MINIMIZIRAJU KRITERIJ J. ZA SVAKU OD K GRUPI, SVAKI OD $(N-k)$ PRIMJERA KOJI TRENUVNO NISU ODABRANI KAO MEDOIDI ZANJENJUJE SE S TRENUVNO ODABRANIM

MEDOIDOM, IZRAČUNAVA SE BROJ MJEŠE NEMEDU
N-K PRIMJERA I PREDLOŽNOG MEDOIDA TE SE ODABIRE
ONAJ MEDOID ZA KOJI JE TA VRJEDNOST NAJMANJA.

c) UKUPNA VREMENSKA SLOŽENOST ALGORITMA
K-SREDNJIH VRJEDNOSTI JE $O(tNK)$, GDE JE T
BROJ ITERACIJA, A VREMENSKA SLOŽENOST ALGORITMA
K-MEDOIDA JE $O(tK(N-k)^2)$.

(2.)

a) EM-ALGORITAM JE MEKO GRUPIRANJE
(IMAMO PROBABILISTIČKI RELAZ, PRIMJER MOže
PRI PADATI U VIŠE GRUPA) ŠTO JE PREDNOST, ALI JE
ZNAČAJNO SLOŽENIJI U ODНОSU NA ALGORITAM SREDNJIH
VRJEDNOSTI.

b)

$$p(x) = \sum_{i=1}^K \pi_k p(x|z_k)$$

$$\ln L(\theta | D) = \ln \prod_{i=1}^N p(x^{(i)}) = \ln \prod_{i=1}^N \sum_{k=1}^K \pi_k p(x^{(i)}|z_k)$$

c)

$$q(x, z | \theta) = p(z) p(x | z, \theta) = \prod_{k=1}^K \pi_k^{z_k} \prod_{k=1}^K p(x | \theta_k)^{z_k} =$$

$\underbrace{\prod_{k=1}^K \pi_k^{z_k} p(x | \theta_k)^{z_k}}$, GDJE JE z_k INDIKATORSKA
VARIJABLA, $z_k = 1$ AKO JE PRIMJER GENERIRAN U
GRUPE θ_k .

$$\begin{aligned} \ln L(\theta | D, z) &= \ln \prod_{i=1}^N \prod_{k=1}^K \pi_k^{z_k^{(i)}} p(x^{(i)} | \theta_k)^{z_k^{(i)}} \\ &= \sum_{i=1}^N \sum_{k=1}^K z_k^{(i)} (\ln \pi_k + \ln p(x^{(i)} | \theta_k)) // \end{aligned}$$

NE MOŽEMO, JER NE ZNAMO VRIJEDNOSTI $z^{(i)}$ T.J.
NE ZNAMO U STARTU KOJOG GRUPI PRIMJER Pripada.

d)

$E = \text{KORAK} : (\text{KORAK PROCJENE})$

$$Q(\theta | \theta^{(t)}) = E_{z|D, \theta^{(t)}} [\ln \mathcal{L}(\theta | D, z)] =$$

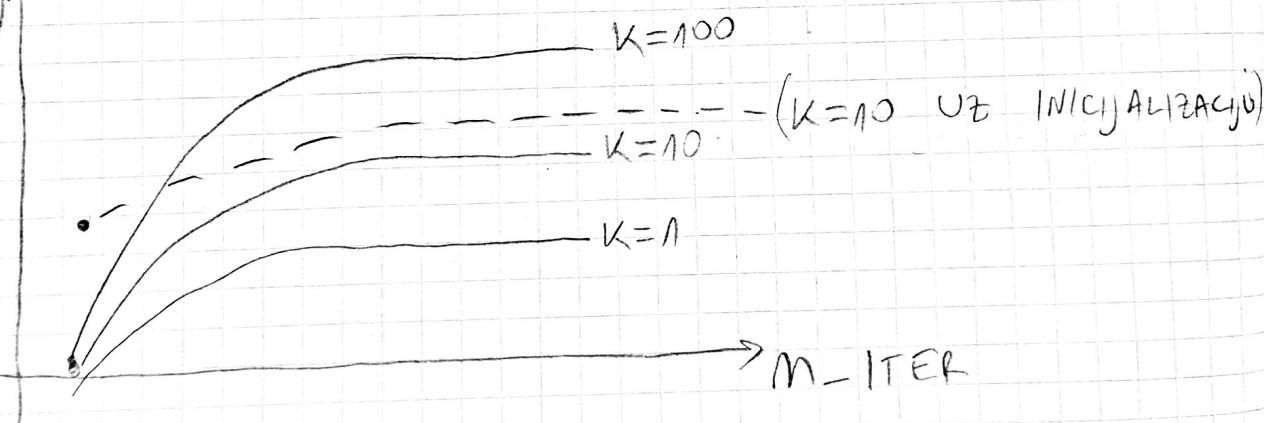
$$= E_{z|D, \theta^{(t)}} [\ln p(D, z | \theta)] = \sum_z p(z|D, \theta^{(t)}) \ln p(D, z | \theta),$$

→ OZNAKA ZA OCJEKIVANJE UZ POZNATE PARAMETRE $\theta^{(t)}$

M -KORAK: (KORAK MAKSIMIZACIJE)

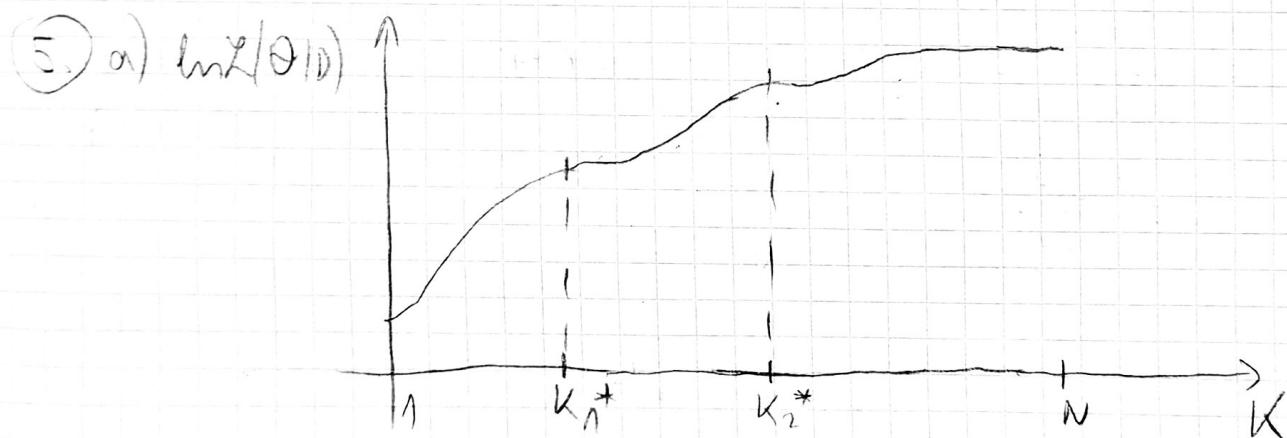
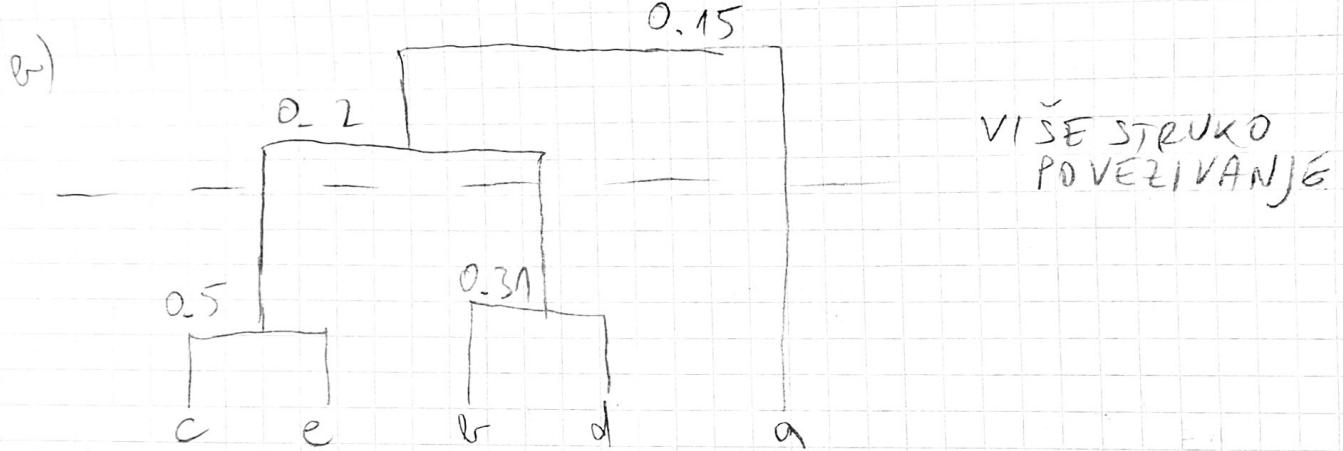
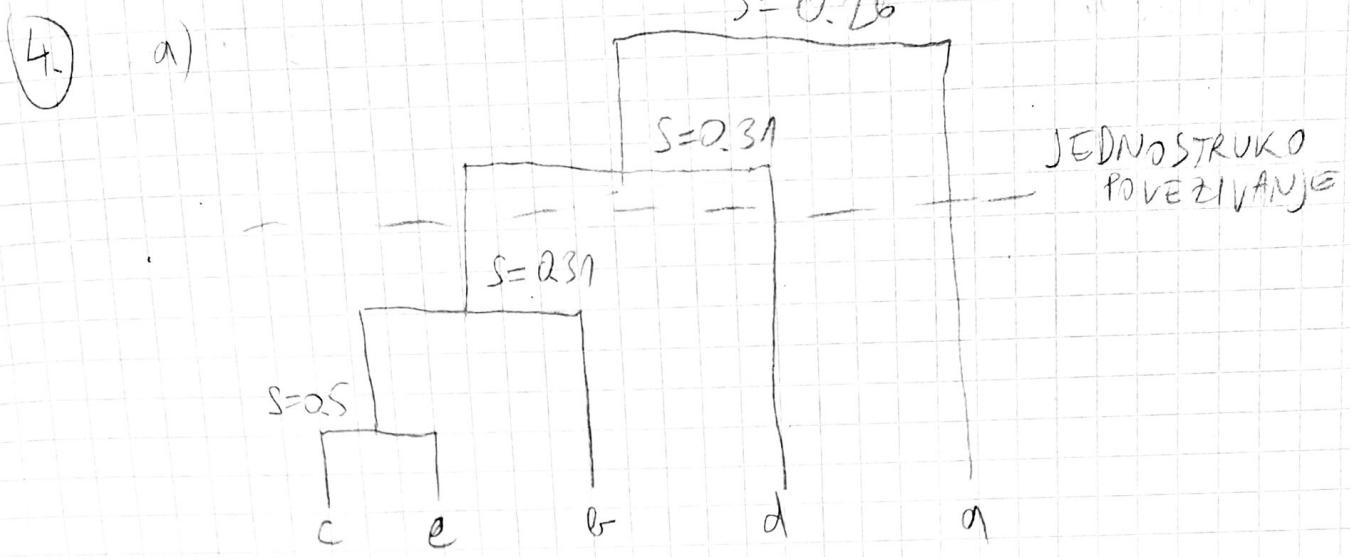
$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta | \theta^{(t)})$$

e) $\ln \mathcal{L}(\theta | D)$

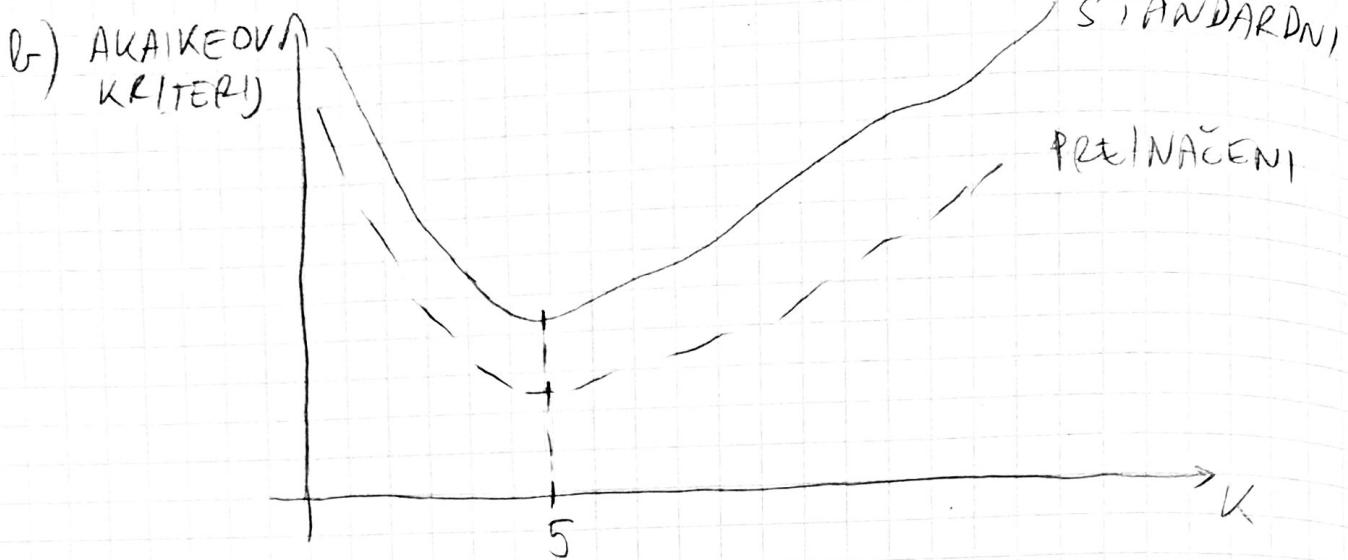


ZA $K=10$ UZ INICIJALIZACIJU SREDIŠTA ALGORITROM

k -SREDNJIH VRIJEDNOSTI LOG-IZGLEDNOST U
POČETNOM TRENUTKU (PRVO ITERACIJI) JE VEĆA.



LOG-IZGLEDNOST RASTE KAKO BROJ GRUPA K
RASTE JER TADA RASTE I SLOŽENOST MODELA.
ZA NEKE VRJEDNOSTI (K_1^*, K_2^*) PRIMIJETIT ĆEMO -
STAGNACIJU RASTA LOG IZGLEDNOSTI - MOŽEMO PREPOSTAVITI
DA SU TE VRJEDNOSTI BROJEVI STVARNIH GRUPA U NAŠEM
USTORU I TO SU KANDIDATI ZA OPTIMALAN K.



c)

$$\text{GRUPA } 0: \{0, 0, 1, 2\} \quad N = \binom{11}{2} = 55$$

$$\text{GRUPA } 1: \{1, 1\}$$

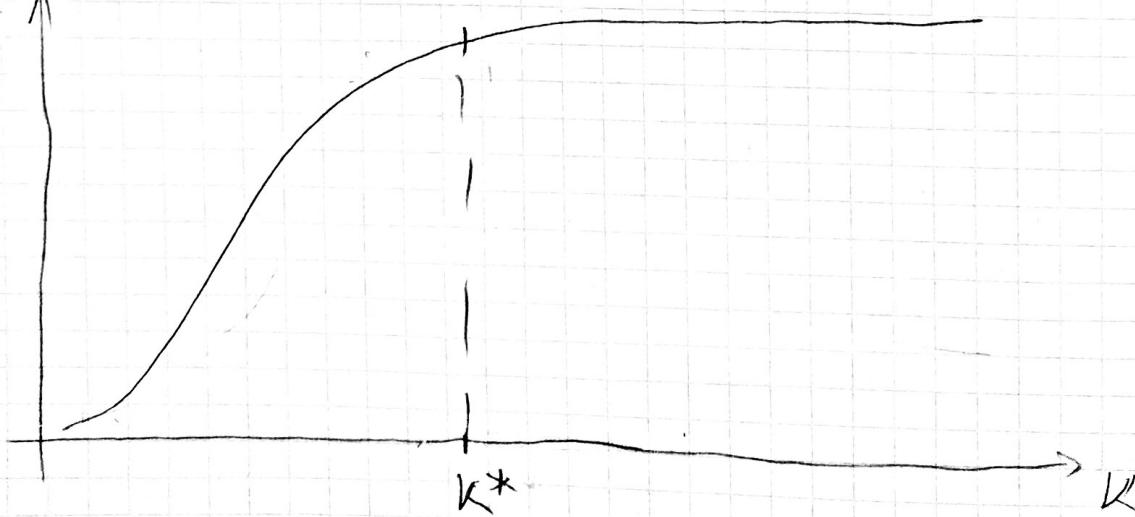
$$\text{GRUPA } 2: \{2, 2, 2, 1, 0\}$$

$$a = \binom{2}{2} + \binom{2}{2} + \binom{3}{2} + \binom{1}{2} + \binom{1}{2} + \binom{1}{2} + \binom{1}{2} = 5$$

$$b = 2 \cdot 2 + 1 \cdot 2 + 2 \cdot 3 + 2 \cdot 1 + 1 \cdot 3 + 1 + 1 + 1 + 2 \cdot 3 + 2 \cdot 1 = 28$$

$$R = \frac{a+b}{N} = \frac{33}{55} = 0.6 //$$

d) R ↑



e) IMA KORISTI, POMOĆU NJEGA MOŽEMO
USPOREĐIVATI REZULTATE DVAJU ALGORITAMA GRUPIRANJA
(NPR. HIJERARHIJSKOG I k -SREDJITIH VRJEDNOSTI).