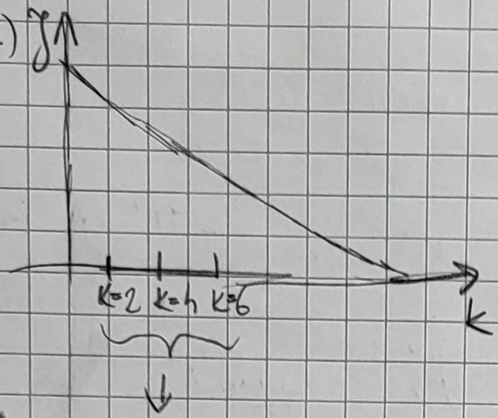
0036333595 Naomi Kombol

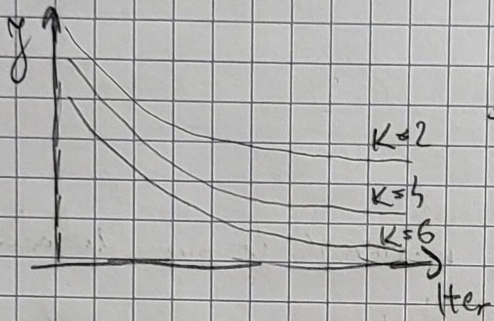
V19 - 1.1

a)



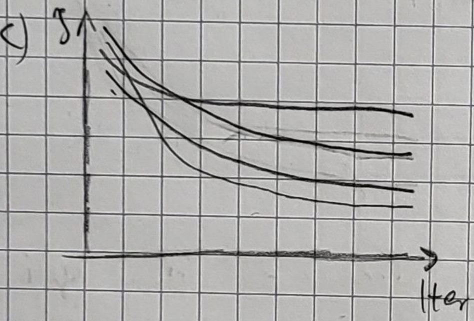
Minimalna vrijednost je 0 što postižemo kad K = br. primjera (svaki primjer postaje centroid svoje grupe pa je euklidska udaljenost od centroida 0)

b)



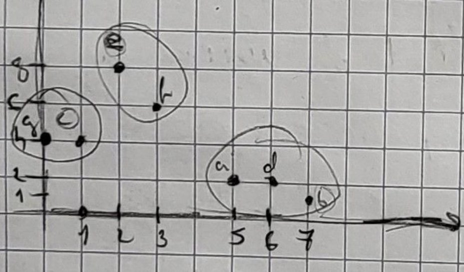
$\Rightarrow J_{K=2} > J_{K=4} > J_{K=6}$

c)



Za kmeans ++, zbog načina odabira centroida, vjerojatnije su krivulje koje imaju niže J kad konvergiraju te one koje konvergiraju brže.

V19 - 1.2



a)

	μ_1	μ_2	μ_3
	b	c	e
a	$\sqrt{5}$	$\sqrt{20}$	$\sqrt{45}$
d	$\sqrt{2}$	$\sqrt{29}$	$\sqrt{52}$
h	$\sqrt{41}$	$\sqrt{29}$	$\sqrt{5}$
g	$\sqrt{58}$	$\sqrt{1}$	$\sqrt{20}$

$\Downarrow$

$\mu_1 = (6, 1.66)$ (aritmetička sred. od a, d, b)
$\mu_2 = (0.5, 4)$ (-||- c, g)
$\mu_3 = (2.5, 7)$ (-||- e, h)

b)

K_1	K_2	K_3
a	c	e
d	g	h
b		

Matrica udaljenosti (euklidskih jer ne koje [illegible] nisu) [illegible] Σ udaljenosti u klasi

	a	b	c	d	e	f	g	Σ
~~a~~	~~0~~	~~5~~	~~20~~	~~1~~	~~45~~	~~20~~	~~29~~	
b	5	0	45	2	74	41	58	7
~~c~~	~~20~~	~~45~~	~~0~~	~~29~~	~~17~~	~~8~~	~~1~~	
d	1	2	29	0	52	25	40	3
~~e~~	~~15~~	~~74~~	~~17~~	~~52~~	~~0~~	~~5~~	~~20~~	
h	20	41	8	25	5	0	13	5
g	29	58	1	40	20	13	0	1

novi medoidi su
d za K_1
h za K_3
g za K_2

0036533595 Naomi Kombol

V19 - 1.2 NASTAVAK

c) K-Means - $O(TnNK)$ br. značajki

K-Medoid - $O(TK(N-K)^2)$ T je br. iteracija

O(K-Medoid) je kvadratne složenosti, dok je O(K-Means) linearne u broju primjera

b) PROS: podatci ne moraju biti vektori
moguće su različite mjere udaljenosti / različitosti

CONS: velika složenost

V20 - 1.1

a) EM koristi meko grupiranje → probabilistički izlaz te primjer može pripadati više grupa → moguće računati potpunu log-izglednost.

Nedostatak je značajno veća složenost.

b) $p(x) = \sum_{k=1}^{K} p(x, y=k) = \sum_{k=1}^{K} P(y=k)\, p(x|y=k) = \sum_{k=1}^{K} \pi_k\, p(x|\theta_k)$

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

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

$\ln(\mathcal{L}(\theta|\mathcal{D}, Z)) = \ln \prod_{i=1}^{N} p(x^i, z^i|\theta) = \ln \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_k^i}\, p(x, \theta_k)\, z_k^i$

$= \sum_{i=1}^{N} \sum_{k=1}^{K} z_k^i \left(\ln \pi_k + \ln p(x^i|\theta_k)\right)$

Kad bi znali z^i-eve, mogli bi analitički maksimizirati log izglednost. Kako ih ne znamo, moramo iterirati π_k i θ_k za računanje očekivanja izglednosti.

d) E: računanje potpune log izglednosti uz trenutne parametre u iteraciji t

$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|\mathcal{D}, \theta^t}\left[\sum_{i=1}^{N} \sum_{k=1}^{K} z_k^i \left(\ln \pi_k + \ln p(x^i|\theta_k)\right)\right]$

$= \sum_{i=1}^{N} \sum_{k=1}^{K} \underbrace{\mathbb{E}[z_k^i|\mathcal{D}, \theta^t]}_{h_k^i} \left(\ln(\pi_k) + \ln p(x^i|\theta_k)\right)$

0036533595 Wladimir Kombol

V20 - 1.1 NASTAVAK

d) M-korak - izračun parametara za iteraciju t+1 koji maksimiziraju očekivanje.

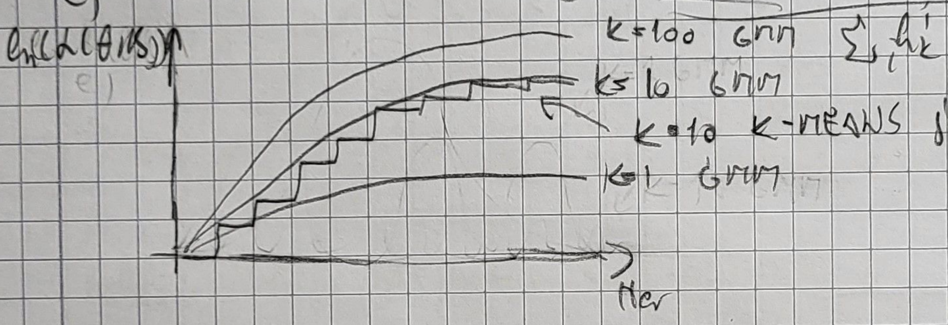
$$\nabla_\theta Q(\theta, \theta^t) = 0$$

$$\nabla_{\pi_k}\left(\sum_{i=1}^{N} \sum_{k=1}^{K} h_k^i \ln \pi_k + \lambda\left(\sum_k \pi_k - 1\right)\right) = 0 \Rightarrow \pi_k^{t+1} = \frac{1}{N}\sum_{i=1}^{N} h_k^i$$

$$\nabla_{\theta_k} \sum_{i=1}^{N} h_k^i \ln p(x^i|\theta_k) = 0 \Rightarrow \mu_k^{t+1} = \frac{\sum_i h_k^i x^{(i)}}{\sum_i h_k^i}$$

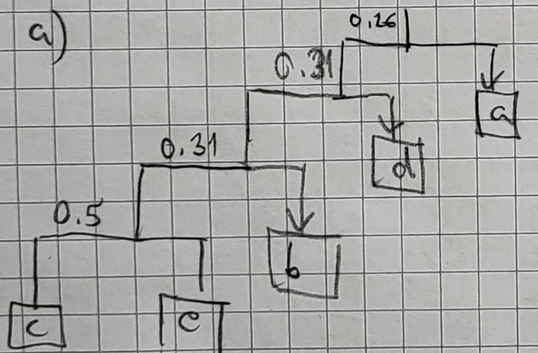
$$\Sigma_k^{t+1} = \frac{\sum_i h_k^i (x^i - \mu^{t+1})(x^i - \mu_k^{t+1})^T}{\sum_i h_k^i}$$

e)



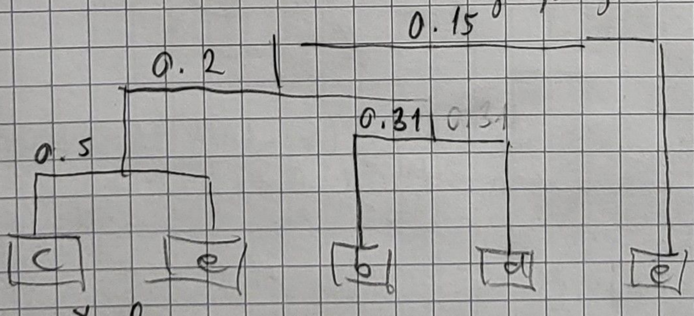
V20 - 1.2

a)



Ako gledamo da želimo naći gdje su primjeri prirodno grupirani i grane dugačke, rekli bi da siječemo nakon grupacije "c" i "e".

b)



I ovdje je najdulja grana između "c" i "e" pa opet siječemo na istom.

V21 - 1.1

y	h
1	1
0	2
2	2
1	2
1	1
0	0
1	1
2	1
0	1
2	0
2	1

TRUE

PRED	0	1	2
0	1	0	1
1	1	3	2
2	1	1	1

Micro

	1	0
1	5	6
0	6	16

0

	1	0
1	1	1
0	2	7

1

	1	0
1	3	3
0	1	3

2

	1	0
1	1	2
0	3	5

Micro: Acc = 21/33

Pr = 5/11

R = 5/11

$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{5}{11}$

Macro: Acc = $\frac{8/11 + 7/11 + 6/11}{3}$ = 21/33

$Pr = \frac{4}{9}$

$R = \frac{4}{9}$ $\Rightarrow F_1 = \frac{4}{9}$