Strojno učenje

14. Algoritam maksimizacije očekivanja

prof. dr. sc. Bojana Dalbelo Bašić
doc. dr. sc. Jan Šnajder

Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Ak. god. 2012/13.

 v1.1

Danas...

1. Probabilističko grupiranje
2. Model Gaussove mješavine
3. Algoritam maksimizacije očekivanja

Danas. . .

1. Probabilističko grupiranje

2. Model Gaussove mješavine

3. Algoritam maksimizacije očekivanja

Probabilističko grupiranje

Prošli tjedan bavili smo se "čvrstim" grupiranjem (particijskim i hijerarhijskim).

k-means

HAC

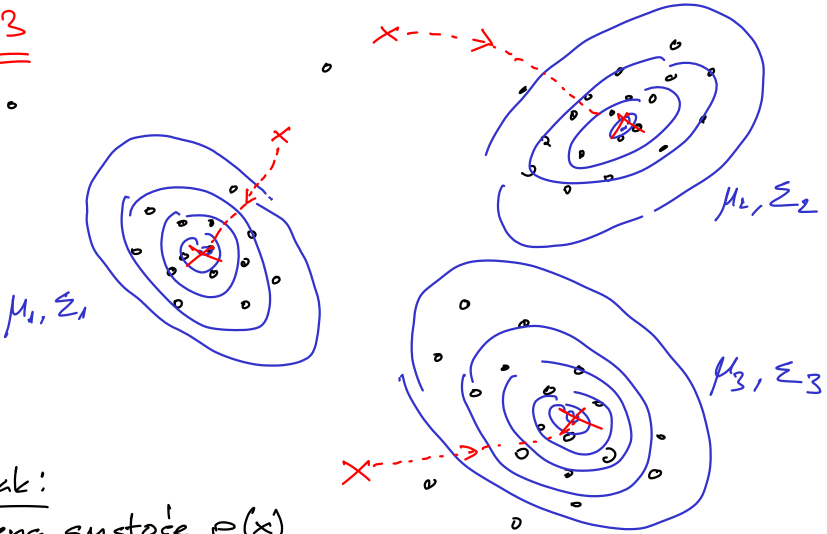
Danas razmatramo "meko" grupiranje: granice između grupa nisu čvrste (primjer može pripadati u više grupa).

Konkretno, razmotrit ćemo probabilističko grupiranje: svaki primjer pripada nekoj grupi s nekom vjerojatnošću.

Probabilističko grupiranje ostvarit ćemo primjenom algoritma maksimizacije očekivanja (EM-algoritam) na model Gaussove mješavine.

Zapravo, radi se o poopćenju algoritma k-srednjih vrijednosti.

$K=3$



zadatak:

procjena gustoće $p(x)$
s tri Gaussove komponente

Danas. . .

Model miješane gustoće

Generativni model: svaki primjer generiran je iz neke razdiobe.

Kod klasifikacije smo unaprijed znali koji primjer pripada kojoj klasi i zasebno smo modelirali izglednosti $p(\mathbf{x}|\mathcal{C}_j)$ za svaku klasu.

Kod grupiranja nemamo oznaka. Modeliramo miješanu gustoću kao linearnu kombinaciju K gustoća:

$$\boxed{p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}|\boldsymbol{\theta}_k)}$$

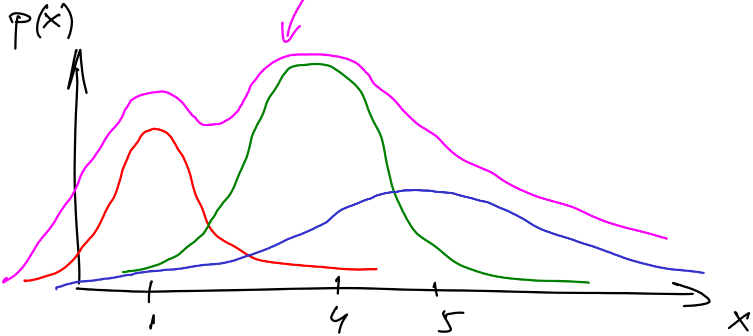
gustoća komponente

komponenta miješanke

Gaussova mješavina

Mješavina Gaussovih gustoća (engl. *mixture of Gaussians*): MoG

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



$K = 3$ $\mu_1 = 1$, $\mu_2 = 4$, $\mu_3 = 5$ $\pi_1 = \pi_2 = 0{,}2$ $\pi_2 = 0{,}6$

$\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 6$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \underbrace{P(\mathcal{G}_k) p(\mathbf{x}|\mathcal{G}_k)}$$

$$\sum_k P(x, G_k) = p(x)$$

Bayesovo pravilo:

$$P(\mathcal{G}_k|\mathbf{x}) = \frac{P(\mathcal{G}_k)p(\mathbf{x}|\mathcal{G}_k)}{p(\mathbf{x})} = \frac{P(\mathcal{G}_k)p(\mathbf{x}|\mathcal{G}_k)}{\sum_j P(\mathcal{G}_j)p(\mathbf{x}|\mathcal{G}_j)} = \frac{\pi_k p(\mathbf{x}|\boldsymbol{\theta}_k)}{\sum_j \pi_j p(\mathbf{x}|\boldsymbol{\theta}_j)} \equiv h_k$$

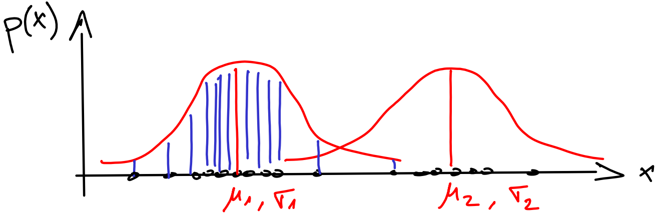
Odgovornost h_k – vjerojatnost da primjer $\mathbf{x}$ pripada grupi $\mathcal{G}_k$

$$h_k^{(i)} \longleftrightarrow b_k^{(i)}$$

$$\in [0, 1] \qquad \in \{0, 1\}$$

Metoda najveće izglednosti

Raspolažemo skupom neoznačenih primjera $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$



$K=2$

Pretpostavimo da je $\mathcal{D}$ generiran Gaussovom mješavinom uz fiksirani K:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \,\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Želimo naučiti parametre modela $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$

Metoda najveće izglednosti

Model:

$$\boxed{p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)}$$

Izglednost na $\mathcal{D}$:

problem!

$$\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) &= \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}) \\
&= \ln \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k\, p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k\, p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)
\end{aligned}$$

$\nabla_{\boldsymbol{\theta}} \ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = 0$ nema rješenja u z.f.

$\Rightarrow$ moramo koristiti iterativnu optimizaciju

- gradijentni spust
- EM algoritam
- MCMC metode

Metoda maksimizacije očekivanja

Prije korištenja iterativne optimizacije, model ćemo proširiti latentnim varijablama.

Latentna varijable opisuju vezu između primjera i grupa: koji primjer pripada kojoj grupi.

Izvana gledano, ne vidimo koji primjer pripada kojoj grupi, zato te varijable nazivamo latentnim (skrivenim).

Latentni modeli vrlo su popularni kod nenadziranog strojnog učenja.

Model s latentnim varijablama

Vektor indikatorskih varijabli:

$$\boxed{\mathbf{z} = (z_1, \ldots, z_k, \ldots, z_K)}$$

0-1 kodiranje

gdje $z_k = 1$ akko je primjer generiran iz grupe $\mathcal{G}_k$

Ako z tretiramo kao slučajnu varijablu

Apriorna vjerojatnost pojedine grupe:

$$\boxed{P(z_k = 1) = \pi_k}$$

Zajednička gustoća:

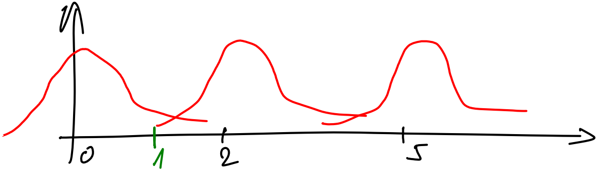
$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \underline{P(\mathbf{z})p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})} = \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k} = \boxed{\prod_{k=1}^{K} \pi_k^{z_k} p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k}}$$

Ovo je model s latentnim varijablama $\mathbf{z}$

$\neq 1$ za isti k

Model s latentnim varijablama – primjer

$n = 1, K = 3, \pi_k = 1/3, \mu_1 = 0, \mu_2 = 2, \mu_3 = 5, \sigma_k = 1$



$\mathbf{x} = 1, \mathbf{z} = (\underset{z_1}{0}, \underset{z_2}{1}, \underset{z_3}{0}), \underline{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})} = \prod_{k=1}^{K} \pi_k^{z_k} p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k} =$

$$\underbrace{\pi_1^{\overset{0}{\overset{\shortparallel}{z_1}}} p(x|\theta_1)^{\overset{0}{\overset{\shortparallel}{z_1}}}} \cdot \pi_2^{z_2} p(x|\theta_2)^{z_2} \cdot \underbrace{\pi_3^{\overset{0}{\overset{\shortparallel}{z_3}}} p(x|\theta_3)^{\overset{0}{\overset{\shortparallel}{z_3}}}}$$

$$1 \cdot \pi_2 \, p(x|\theta_2) \cdot 1$$

Model s latentnim varijablama

Početni model:

$$\boxed{p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)}$$

Miješana gustoća

Model s latentnim varijablama:

$$\boxed{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \pi_k^{z_k} p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k}}$$

Miješana gustoća
+ latentne var.

Umjesto sume imamo produkt, pa maksimizacija log-izglednosti ima rješenje u z.f.!

Log-izglednost prvog modela nazivamo nepotpuna log-izglednost, a drugog modela potpuna log-izglednost.

Potpuna log-izglednost

$[z^1\ z^2\ \ldots\ z^N]$

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

$= P(D, Z | \theta)$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} z_k^{(i)}\Big(\ln \pi_k + \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)\Big)$$

$\begin{cases} 1 & \text{ako je } x^i \text{ generiran od } k \\ 0 & \text{inače} \end{cases}$

Ovo bismo mogli analitički optimizirati, kada bismo znali vrijednosti varijabli $\mathbf{z}^{(i)}$ (koji primjer pripada kojoj grupi). Ali to ne znamo!

No, možemo izračunati očekivanje potpune log-izglednosti uz neke pretpostavljene vrijednosti za parametre π_k i $\boldsymbol{\theta}_k$.

$\Rightarrow$ algoritam maksimizacije očekivanja

Danas. . .

1. Probabilističko grupiranje
2. Model Gaussove mješavine
3. Algoritam maksimizacije očekivanja

Algoritam maksimizacije očekivanja

Pronalazi parametre $\boldsymbol{\theta}^*$ koji maksimiziraju očekivanje potpune log-izglednosti uz fiksirane parametre:

$$\boxed{\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} = \mathbb{E}_{\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}'}\Big[\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D},\mathbf{Z})\Big]}$$

Parametri $\boldsymbol{\theta}'$ su trenutna procjena parametara. Algoritam iterativno poboljšava tu procjenu.

Može se pokazati da parametri koji maksimiziraju $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D},\mathbf{Z})$ također maksimiziraju $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$, a to je upravo ono što tražimo.

Algoritam radi iterativno. U svakoj iteraciji radi se E-korak i M-korak.

Algoritam maksimizacije očekivanja

E-korak: korak procjene

Oznaka za očekivanje potpune log. izg. uz fiksirane parametre θ^t

$$\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \equiv \mathbb{E}_{\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}^{(t)}}\Big[\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D},\mathbf{Z})\Big] = \mathbb{E}_{\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}^{(t)}}\Big[\ln p(\mathcal{D},\mathbf{Z}|\boldsymbol{\theta})\Big]$$
$$= \sum_{\mathbf{Z}} P(\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}^{(t)}) \ln p(\mathcal{D},\mathbf{Z}|\boldsymbol{\theta})$$

$\Rightarrow P(\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}^{(t)})$ ćemo izračunati Bayesovim pravilom

M-korak: korak maksimizacije

funkcija od θ

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

$\Rightarrow$ provodimo analitički (moguće jer radimo s potpunom log-izglednošću)

E-korak

jedino z_k^i je slučajna varijabla!

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}^{(t)}}\Big[\ln\mathcal{L}(\boldsymbol{\theta}|\mathcal{D},\mathbf{Z})\Big]\\ &= \mathbb{E}_{\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}^{(t)}}\Big[\sum_{i=1}^{N}\sum_{k=1}^{K} z_k^{(i)}\big(\ln\pi_k + \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k^{(t)})\big)\Big]\\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \underbrace{\mathbb{E}[z_k^{(i)}|\mathcal{D},\boldsymbol{\theta}^{(t)}]}\big(\ln\pi_k + \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)\big)\end{aligned}$$

Očekivanje latentne varijable:

$$\begin{aligned}\mathbb{E}[z_k^{(i)}|\mathcal{D},\boldsymbol{\theta}^{(t)}] &= \mathbb{E}[z_k^{(i)}|\mathbf{x}^{(i)},\boldsymbol{\theta}^{(t)}] = P(z_k^{(i)}=1|\mathbf{x}^{(i)},\boldsymbol{\theta}^{(t)})\\ &= \frac{p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k^{(t)})\pi_k^t}{\sum_{j=1}^{K} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k^{(t)})\pi_j^t} = h_k^{(i)}\end{aligned}$$

$\leftarrow$ odgovornost

M-korak

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{N}\sum_{k=1}^{K} h_k^{(i)}\Big(\ln \pi_k + \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)\Big)\\
&= \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} h_k^{(i)} \ln \pi_k} + \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} h_k^{(i)} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)}
\end{aligned}
$$

Maksimizacija:

$$\boxed{\nabla_{\boldsymbol{\theta}}\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = 0}$$

λ - Lagrangeov multiplikator

$$\nabla_{\pi_k}\Big(\sum_{i=1}^{N}\sum_{k=1}^{K} h_k^{(i)} \ln \pi_k + \lambda\big(\sum_k \pi_k - 1\big)\Big) = 0$$

optimizacija uz ograničenje $\sum_k \pi_k = 1$

$$\nabla_{\boldsymbol{\theta}_k}\sum_{i=1}^{N} h_k^{(i)} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) = 0$$

M-korak

Za komponente mješavine:

$$\boxed{\pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} h_k^{(i)}}$$

→ zbroj vjerojatnosti da primjeri pripadaju grupi k

Za parametre Gaussovih komponenata:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_i h_k^{(i)}\mathbf{x}^{(i)}}{\sum_i h_k^{(i)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_i h_k^{(i)}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)})^{\mathrm{T}}}{\sum_i h_k^{(i)}}$$

meko grupiranje jer $h_k^i \in [0, 1]$

Algoritam maksimizacije očekivanja

EM-algoritam za Gaussovu mješavinu

inicijaliziraj parametre $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$

ponavljaj do konvergencije log-izglednosti ili parametara

E-korak:

Za svaki primjer $\mathbf{x}^{(i)} \in \mathcal{D}$ i svaku komponentu $k = 1, \ldots, K$:

$$h_k^{(i)} \leftarrow \frac{p(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^K p(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_j}$$

M-korak:

Za svaku komponentu $k = 1, \ldots, K$:

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_i h_k^{(i)} \mathbf{x}^{(i)}}{\sum_i h_k^{(i)}}, \quad \boldsymbol{\Sigma}_k \leftarrow \frac{\sum_i h_k^{(i)} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_i h_k^{(i)}}, \quad \pi_k \leftarrow \frac{1}{N} \sum_{i=1}^N h_k^{(i)}$$

Izračunaj trenutnu vrijednost log-izglednosti

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^N \ln \sum_{k=1}^K \pi_k \, p(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

raste do konvergencije

Veza s algoritmom k-srednjih vrijednosti

EM-algoritam je poopćenje algoritma k-srednjih vrijednosti!

Uz sljedeće pretpostavke dobivamo algoritam k-srednjih vrijednosti:

(1) odgovornosti $h_k^{(i)}$ se zaokružuju na 0 ili 1
$\Rightarrow$ čvrsto grupiranje

(2) Gaussove komponente imaju dijeljenu izotropnu kov. matricu, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$
$\Rightarrow p(\mathbf{x}|z_k, \boldsymbol{\theta}_k) \;=\; -\frac{1}{2\sigma^2}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2$

$$\Sigma = \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \ddots \end{bmatrix}$$

Ako (1) i (2), onda $h_k^{(i)} = b_k^{(i)}$

U tom slučaju vrijedi:

$$\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) \propto -J$$

kriterijska funkcija algoritma k-srednjih vrijednosti

Napomene

- EM-algoritam nužno konvergira, ali u lokalni optimum log-izglednosti. Rezultat vrlo ovisi o inicijalizaciji parametara.
- Poznato je da algoritam sporo konvergira. Radi ubrzanja, inicijalizacija središta $\boldsymbol{\mu}_k$ može se provesti algoritmom k-srednjih vrijednosti.
- Kao i kod svih algoritama grupiranja, broj grupa K je hiperparametar koji treba nekako unaprijed odrediti.
- Može se koristiti Akaikeov informacijski kriterij (AIC):

$$K^* = \underset{K}{\operatorname{argmin}} \big(-2 \ln \mathcal{L}(K) + 2q(K) \big)$$

$q(K)$

broj parametara za K grupa

- EM-algoritam je općenit algoritam za optimizaciju parametara modela s latentnih varijablama! Ovdje smo ga primijenili na grupiranje (na model s Gaussovim mješavinama).

Sažetak

- Kod **probabilističkog grupiranja** primjeri pripadaju grupama s određenom vjerojatnošću
- Probabilističko grupiranje možemo promatrati kao **optimizaciju log-izglednosti Gaussove mješavine**
- Taj problem je rješiv ako model proširimo **latentnim varijablama** i optimiziramo **potpunu log-izglednost**
- Optimizaciju potpune log-izglednosti provodimo **algoritmom maksimizacije očekivanja (EM-algoritam)**
- EM-algoritam je **poopćenje** algoritma k-srednjih vrijednosti
- Algoritam konvergira, ali ne nalazi nužno optimalno grupiranje
- Algoritam je općenito primjenjiv na optimizaciju parametara latentnih modela

Gotovo!