20. Grupiranje II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.1

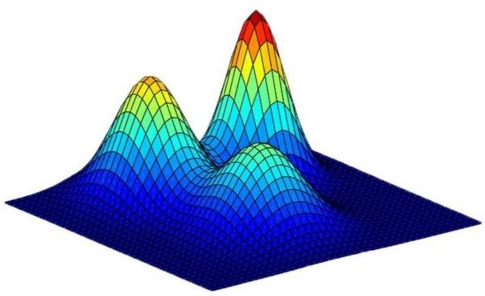
1 Model Gaussove mješavine

- **Model Gaussove mješavine (GMM)** $\Rightarrow$ probabilističko meko particijsko grupiranje
- Poopćenje algoritma K-sredina: umjesto $b_k^{(i)} \in \{0,1\}$ imamo $h_k^{(i)} \in [0,1]$
- $h_k^{(i)}$ je **odgovornost** – vjerojatnost da je primjer $\mathbf{x}^{(i)}$ generirala grupa k
- GMM je poseban slučaj **modela miješane gustoće** (*mixture model*)
- Model miješane gustoće je linearna kombinacija K **komponenti**:

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

  gdje su π_k **koeficijenti mješavine**, a $p(\mathbf{x}|\boldsymbol{\theta}_k)$ **gustoće komponenti**

- Primjer: model bivarijatne Gaussove mješavine s $K = 3$ grupe:



- Odgovornost možemo izračunati Bayesovim pravilom:

$$h_k^{(i)} = P(y = k|\mathbf{x}^{(i)}) = \frac{P(y = k)p(\mathbf{x}^{(i)}|y = k)}{\sum_j P(y = j)p(\mathbf{x}^{(i)}|y = j)} = \frac{\pi_k p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)}{\sum_j \pi_j p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_j)}$$

- Parametri modela su $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\theta}_k\}_{k=1}^{K}$, $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Parametre možemo (pokušati) procijeniti metodom MLE

- Log-izglednost parametara modela (tzv. **nepotpuna izglednost**):

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

$\Rightarrow$ ne faktorizira se po komponentama $\Rightarrow$ maksimizacija nema analitičko rješenje

2 Algoritam maksimizacije očekivanja

- Proširenje modela miješane gustoće **latentnim varijablama** (varijable koje ne opažamo)
- Latentna kategorička varijabla $\mathbf{z}^{(i)}$ definira koja je grupa generirala primjer $\mathbf{x}^{(i)}$:

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

- Distribucija kategoričke varijable $\mathbf{z}^{(i)}$:

$$P(\mathbf{z}^{(i)} = k) = \prod_{k=1}^{K} \pi_k^{z_k^{(i)}}$$

- Zajednička gustoća varijabli $\mathbf{x}$ i $\mathbf{z}$:

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

$\Rightarrow$ model s **latentnim varijablama** $\mathbf{z}$

- Log-izglednost parametara modela (tzv. **potpuna izglednost**):

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

$\Rightarrow$ ako su $\mathbf{z}^{(i)}$ poznate, maksimizacija ove log-izglednosti ima analitičko rješenje

- $\mathbf{z}^{(i)}$ su nepoznate, no možemo izračunati **očekivanje** izglednosti uz fiksirane π_k i $\boldsymbol{\theta}_k$
- Može se pokazati: povećanje očekivanja od $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}, \mathbf{Z}) \Rightarrow$ povećanje $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$
- **Algoritam maksimizacije očekivanja (EM-algoritam)**: iterativna optimizacija $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$
- Dva koraka algoritma: E-korak (*expectation*) i M-korak (*maximization*)
- **E-korak**: Izračun očekivanja potpune izglednosti uz fiksirane parametre u iteraciji t:

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

- **M-korak**: Izračun parametara za iteraciju $(t+1)$ koji maksimiziraju očekivanje:

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

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

$$\nabla_{\boldsymbol{\theta}_k} \sum_{i=1}^{N} h_k^{(i)} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_i h_k^{(i)}\mathbf{x}^{(i)}}{\sum_i h_k^{(i)}}$$

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

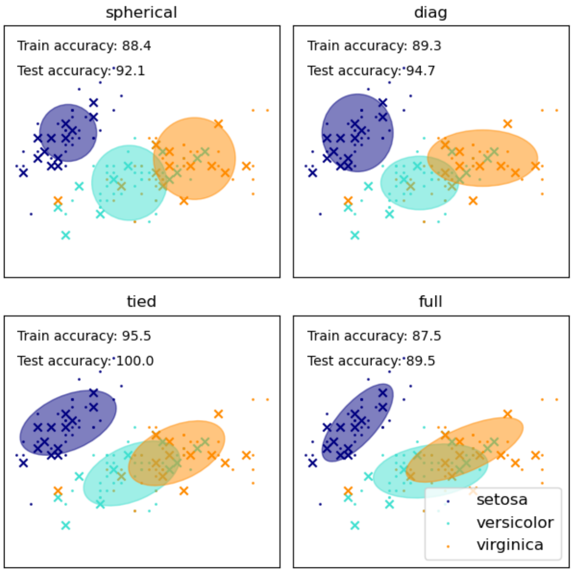
Algoritam GMM (model GMM + EM-algoritam)

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 $k = 1, \dots, K$:
  $h_k^{(i)} \leftarrow \frac{p(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^{K} p(\mathbf{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j}$
  M-korak:
  Za svaku komponentu $k = 1, \dots, K$:
  $\boldsymbol{\mu}_k \leftarrow \frac{\sum_i h_k^{(i)}\mathbf{x}^{(i)}}{\sum_i h_k^{(i)}}, \quad \boldsymbol{\Sigma}_k \leftarrow \frac{\sum_i h_k^{(i)}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_i h_k^{(i)}}, \quad \pi_k \leftarrow \frac{1}{N}\sum_{i=1}^{N} h_k^{(i)}$
  Izračunaj trenutnu vrijednost log-izglednosti
  $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k\, p(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

- EM-algoritam konvergira, ali ne nužno u globalni optimum log-izglednosti
- Akaikeov informacijski kriterij (AIC) za odabir optimalnog broja grupa:

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

- Moguća pojednostavljenja: dijeljena matrica, dijagonalna ili izotropna matrica Σ



3 Hijerarhijsko grupiranje

- Hijerarhijsko grupiranje producira **dendrogram** – stablasti prikaz hijerarhije grupa
- Provodi se na temelju mjere udaljenosti ili mjere sličnosti/različitosti
- Može biti **aglomerativno** (bottom-up) ili **divizivno** (top-down)
- **Hijerarhijsko aglomertivno grupiranje (HAC)**: iterativno stapa najbliže parove grupa
- **Povezivanje** (*linkage*) – način izračuna udaljenosti između dvije grupe:
  - **Jednostruko povezivanje** (*single linkage*)

$$d_{min}(\mathcal{G}_i, \mathcal{G}_j) = \min_{\mathbf{x}\in\mathcal{G}_i, \mathbf{x}'\in\mathcal{G}_j} d(\mathbf{x}, \mathbf{x}')$$

  - **Potpuno povezivanje** (*complete linkage*)

$$d_{max}(\mathcal{G}_i, \mathcal{G}_j) = \max_{\mathbf{x}\in\mathcal{G}_i, \mathbf{x}'\in\mathcal{G}_j} d(\mathbf{x}, \mathbf{x}')$$

  - **Prosječno povezivanje** (*average linkage*)

$$d_{avg}(\mathcal{G}_i, \mathcal{G}_j) = \frac{1}{N_i N_j} \sum_{\mathbf{x}\in\mathcal{G}_i} \sum_{\mathbf{x}'\in\mathcal{G}_j} d(\mathbf{x}, \mathbf{x}')$$

  - **Povezivanje centroida** (*centroid linkage*)

$$d_{cent}(\mathcal{G}_i, \mathcal{G}_j) = \left\| \frac{1}{N_i} \sum_{\mathbf{x}\in\mathcal{G}_i} \mathbf{x} - \frac{1}{N_j} \sum_{\mathbf{x}\in\mathcal{G}_j} \mathbf{x} \right\|$$

Algoritam hijerarhijskog aglomerativnog grupiranja (HAC)

inicijaliziraj K, $k \leftarrow N$, $\mathcal{G}_i \leftarrow \{\mathbf{x}^{(i)}\}$ za $i = 1, \dots, N$
ponavljaj
  $k \leftarrow k - 1$
  $(\mathcal{G}_i, \mathcal{G}_j) \leftarrow \underset{\mathcal{G}_a, \mathcal{G}_b}{\operatorname{argmin}} \, d(\mathcal{G}_a, \mathcal{G}_b)$
  $\mathcal{G}_i \leftarrow \mathcal{G}_i \cup \mathcal{G}_j$
dok je $k > K$

- Prostorna složenost: matrica udaljenosti za $\binom{N}{2}$ parova primjera $\Rightarrow$ $\mathcal{O}(N^2)$
- Vremenska složenost: općenito $\mathcal{O}(N^3)$, $\mathcal{O}(N^2 \log N)$ s prioritetnom listom