Lec 24 EM algo (soft decision)

Recall: GMM & EM algo:

① Initialization:

Arbitary bump membership $\{B_j\}_{j=1}^{K}$

② Given $\{B_j\}$ estimate $\Omega = \{w_j, \mu_j, \Sigma_j\}_{j=1}^{K}$

$$w_j = \frac{N_j}{N} = \frac{|B_j|}{N}$$

$$\underline{\mu}_j = \frac{1}{N_j} \sum_{x_n \in B_j} \underline{x}_n$$

$$\Sigma_j = \frac{1}{N_j} \sum_{x_n \in B_j} (\underline{x}_n - \underline{\mu}_j)(\underline{x}_n - \underline{\mu}_j)^T$$

③ Given Ω, estimate $\{B_j\}$

$$j^* = \underset{j}{\operatorname{argmax}}\ N(\underline{x}_n ; \underline{\mu}_j ; \Sigma_j)\ w_j$$

Repeat step 2 and 3 until converge.

1. EM algo (soft decision)

(Assume a datapoint is divisible)

→ At each step, let γ_{nj} denote the fraction (prob.) of $\underline{x}_n$ that belongs to B_j.

$\therefore \sum_{j=1}^{K} \gamma_{nj} = 1$, $\gamma_{nj} \geq 0$ (γ_{nj} : No. of fraction go to bound j)

→ **subproblem 1 becomes:**

Given $\{\gamma_{nj}\}_{\substack{1 \leq n \leq N \\ 1 \leq j \leq K}}$, update Ω

$$N_j = \sum_{n=1}^{N} \gamma_{nj}$$

↑ Effective (expected) No. of data points in B_j.

$$w_j = \frac{N_j}{N}$$

$$\underline{\mu}_j = \frac{1}{N_j} \sum_{n=1}^{N} \gamma_{nj} \, \underline{x}_n$$

$$\Sigma_j = \frac{1}{N_j} \sum_{n=1}^{N} \gamma_{nj} \, (\underline{x}_n - \underline{\mu}_j)(\underline{x}_n - \underline{\mu}_j)^T$$

↳ **Subproblem 2:**

Given Ω, update $\{\gamma_{nj}\}$

$$\gamma_{nj} = Pr(j \mid \underline{x}_n)$$

$$= \frac{P(\underline{x}_n \mid j) \cdot Pr(j)}{\sum_{i=1}^{k} P(\underline{x}_n \mid i) \cdot Pr(i)}$$

$$= \frac{\mathcal{N}(\underline{x}_n ; \underline{\mu}_j , \Sigma_j)\, w_j}{\sum_{i=1}^{k} \mathcal{N}(\underline{x}_n ; \underline{\mu}_i , \Sigma_i)\, w_i}$$

↳ **Summary of soft-EM:**

① Initialization:

arbitary $\{\gamma_{nj}\}_{\substack{1 \le n \le N \\ 1 \le j \le k}}$ (e.g. k-means ⇒ binary γ_{nj})

① Given $\{\gamma_{nj}\}$, update Ω

② Given Ω, update $\{\gamma_{nj}\}$

Repeat step ① ② until convergence

$\begin{cases} \text{convergence gauranteed} \\ \text{hidden variables } \{\gamma_{nj}\} \text{ can be thrown out at the end} \end{cases}$

2. Relation to $E_{in}(\Omega)$:

(Hard decision case)

↳ Recall: $E_{in}(\Omega) = -\log \hat{P}_{\Omega}(\mathcal{D}) = -\log \prod_{n=1}^{N} \hat{P}(\underline{x}_n)$

$$= -\log \prod_{n=1}^{N} \underbrace{\sum_{j=1}^{K} w_j \, \mathcal{N}(\underline{x}_n ; \underline{\mu}_j , \Sigma_j)}_{\text{GMM}}$$

↳ In subproblem 1, $\{B_j\}$ is given

Let j_n be the bump that $\underline{x}_n$ belongs to.

$\hat{p}(\underline{x}_n \mid \underline{x}_n \in B_{j_n}) = \mathcal{N}(\underline{x}_n ; \underline{\mu}_{j_n}, \Sigma_{j_n})$

$\hat{p}(\underline{x}_n , \underline{x}_n \in B_{j_n}) = \hat{p}(\underline{x}_n \mid \underline{x}_n \in B_{j_n}) \Pr\{\underline{x}_n \in B_{j_n}\}$

$= w_{j_n} \, \mathcal{N}(\underline{x}_n ; \underline{\mu}_j , \Sigma_{j_n})$

$E_{in}^{\{B_j\}}(\Omega) = -\log \prod_{n=1}^{N} \hat{p}(\underline{x}_n , \underline{x}_n \in B_{j_n})$

(likelihood of $\mathcal{D}$ and $\{B_j\}$)

$= -\log \prod_{n=1}^{N} w_{j_n} \mathcal{N}(\underline{x}_n ; \underline{\mu}_j , \Sigma_{j_n})$

$= \sum_{n=1}^{N} -\log(w_{j_n}) + \sum_{n=1}^{N} -\log[\mathcal{N}(\underline{x}_n ; \underline{\mu}_j , \Sigma_{j_n})]$

$= \sum_{j=1}^{K} -N_j \log w_j + \sum_{j=1}^{K} \sum_{\underline{x}_n \in B_j} -\log[\mathcal{N}(\underline{x}_n ; \underline{\mu}_j , \Sigma_{j_n})]$

Seperately optimize those two parts:

① Lagrange $\Rightarrow$ $w_j^* = \frac{N_j}{N}$

② Gradient $= 0$ $\Rightarrow$ $\underline{\mu}_j = \frac{1}{N_j} \sum_{\underline{x}_n \in B_j} \underline{x}_n$ $\Rightarrow$ $\Sigma_j = \cdots$

(same as what we had)

↳ <u>In subproblem 2</u>

Given Ω, $\hat{j}^* = \underset{\{B_j\}}{\arg\max} \Pr\{j \mid \underline{x}_n\}$

$\equiv$ Minimize $E_{in}(\Omega)$