General EM

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 9

Gaussian mixture model (GMM)

- Let $\mathbf{z}$ be a random variable that denotes the clustering.
  - $\mathbf{z}$ is one-hot and $z_k = 1$ implies choosing the k-th cluster.

$\Updownarrow$ $z = e_k$

- The marginal distribution over $\mathbf{z}$ is given by

$$p(z_k = 1) = \pi_k$$

where the parameters satisfy

$$0 \leqslant \pi_k \leqslant 1$$

$$\sum_{k=1}^{K} \pi_k = 1$$

Gaussian mixture model (GMM)

- Similar to multi-class classification, we can write

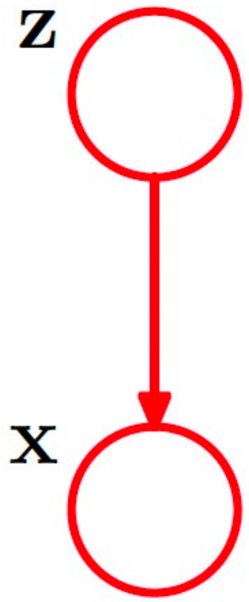
$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- In a Gaussian mixture model (GMM), the conditional distribution $p(\mathbf{x}|\mathbf{z})$ satisfies

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- That is, each cluster is a Gaussian. We can write

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$



Gaussian mixture model (GMM)

- Therefore,

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

- By Bayes' Theorem,

$p(z=e_k|x)$

$\|$

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

"responsibility" that z_k takes in explaining $\mathbf{x}$

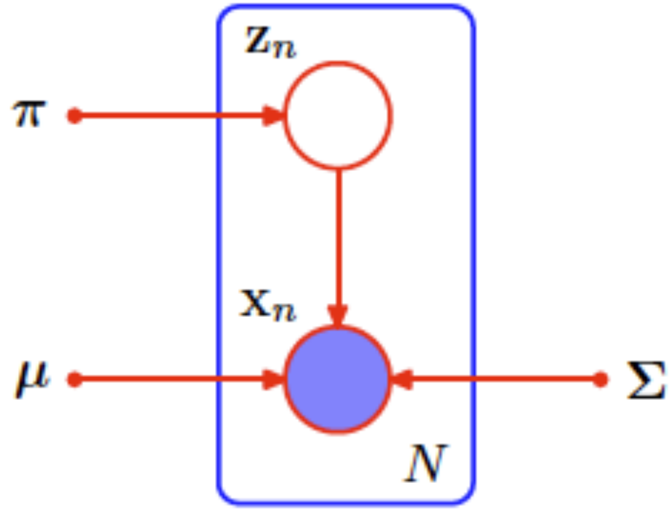
MLE for GMM

- Question: if we are given a sample $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^{N}$, how do we derive the MLE of the underlying GMM?

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$$\|$$

$$\ln \prod_{n=1}^{N} p(x_n|\pi, \mu, \Sigma)$$



$$\ln p(X \mid \pi, \mu, \Sigma) = \underbrace{\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \, \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right)}_{f\left(\underbrace{\{\pi_k\}_{k=1}^{K}}_{\pi},\ \underbrace{\{\mu_k\}_{k=1}^{K}}_{\mu},\ \underbrace{\{\Sigma_k\}_{k=1}^{K}}_{\Sigma}\right)}$$

$$\frac{\partial f}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k \frac{\partial}{\partial \mu_k} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

$$\frac{\partial}{\partial \mu_k} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

$$= \frac{\partial}{\partial \mu_k} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right)$$

$$= \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \cdot \frac{\partial}{\partial \mu_k} \left(-\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right)$$

$$= \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \ \Sigma_k^{-1} (x_n - \mu_k)$$

Therefore,

$$\frac{\partial f}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k \, \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} \Sigma_k^{-1} (x_n - \mu_k)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \ \Sigma_k^{-1} (x_n - \mu_k)$$

Setting $\frac{\partial f}{\partial \mu_k} = 0$ yields

$$\sum_{n=1}^{N} \gamma(z_{nk}) \cancel{\Sigma_k^{-1}} (x_n - \mu_k) = 0$$

That is, $\left[\sum_{n=1}^{N} \gamma(z_{nk})\right] \mu_k = \sum_{n=1}^{N} \gamma(z_{nk}) \, x_n$

Denote $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$. We have

$$\boxed{\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \, x_n}$$

Next,

$$\frac{\partial f}{\partial \Sigma_k} = \sum_{n=1}^{N} \frac{\pi_k \frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

$$\frac{\partial}{\partial \Sigma_k} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

$$= \frac{\partial}{\partial \Sigma_k} \left[\frac{1}{(2\pi)^{\frac{D}{2}}} \cdot \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right) \right]$$

$$= \frac{1}{(2\pi)^{\frac{D}{2}}} \left[\left(\frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \right) \cdot \exp(\cdots) + \left(-\frac{1}{2}\right) \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \cdot \exp(\cdots) \left(\frac{\partial}{\partial \Sigma_k} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right) \right]$$

$$\frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}}$$

$$= -\frac{1}{2|\Sigma_k|^{\frac{3}{2}}} \frac{\partial}{\partial \Sigma_k} |\Sigma_k|$$

$$= -\frac{1}{2|\Sigma_k|^{\frac{3}{2}}} |\Sigma_k| \left(\Sigma_k^{-1}\right)^T$$

(by HW1, Problem #4; or [Bi:] Appendix C Eq (C.22))

$$= -\frac{1}{2|\Sigma_k|^{\frac{1}{2}}} \Sigma_k^{-1}$$

$$\frac{\partial}{\partial \Sigma_k} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)$$

$$= \frac{\partial}{\partial \Sigma_k} \operatorname{tr} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)$$

$$= \frac{\partial}{\partial \Sigma_k} \operatorname{tr} \left(\Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \right)$$

$$= - \Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1}$$

(by e.g. 3 in the previous lecture

$$\frac{\partial}{\partial A} \operatorname{tr}(A^{-1} B) = -(A^{-1} B A^{-1})^T$$
)

Therefore,

$$\frac{\partial}{\partial \Sigma_k} N(x_n | \mu_k, \Sigma_k)$$

$$= \frac{1}{(2\pi)^{\frac{D}{2}}} \Big[\Big(- \frac{1}{2|\Sigma_k|^{\frac{1}{2}}} \Sigma_k^{-1} \Big) \cdot \exp(\cdots) +$$

$$\left(-\frac{1}{2}\right) \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp(\cdots) \Big(- \Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \Big) \Big]$$

$$= - \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{2|\Sigma_k|^{\frac{1}{2}}} \exp(\cdots) \Sigma_k^{-1} \cdot$$

$$\left(I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right)$$

$$= -\frac{1}{2} N(x_n | \mu_k, \Sigma_k) \Sigma_k^{-1} \left(I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right)$$

and

$$\frac{\partial f}{\partial \Sigma_k} = \sum_{n=1}^{N} \frac{\pi_k N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)} \left(-\frac{1}{2}\right) \cdot$$

$$\Sigma_k^{-1} \left(I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \left(-\frac{1}{2}\right) \Sigma_k^{-1} \left(I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right)$$

Setting this to zero yields

$$\sum_{n=1}^{N} \gamma(z_{nk}) \cancel{\left(-\frac{1}{2}\right)} \cancel{\Sigma_k^{-1}} \left(I - (x_n - \mu_k)(x_n - \mu_k)^T \Sigma_k^{-1} \right) = 0$$

Right-multiplying with Σ_k, we have

$$\left(\sum_{n=1}^{N} \gamma(z_{nk}) \right) \Sigma_k = \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T$$

$$\boxed{\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T .}$$

Next, for π, we need to

$$\max_{\pi} \underbrace{\ln p(X \mid \pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)}_{g(\pi)}$$

Setting

$$\frac{\partial g}{\partial \pi_k} = \sum_{n=1}^{N} \frac{N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)} + \lambda = 0$$

yields

$$\sum_{n=1}^{N} \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)} = -\pi_k \lambda$$

Summing over k, we have

$$\sum_{n=1}^{N} 1 = -\lambda$$

Therefore, $\lambda = -N$

Hence, $\boxed{\pi_k = \dfrac{\sum_{n=1}^{N} \gamma(z_{nk})}{-\lambda} = \dfrac{N_k}{N}}$

MLE for GMM

- Note that it is not a closed-form solution that

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\pi_k = \frac{N_k}{N}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})}$$

MLE for GMM

- Note that it is not a closed-form solution that

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\pi_k = \frac{N_k}{N}$$

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\underbrace{\frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})}$$

EM for GMM

(E-step): expectation

$$\underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}$$

for fixed parameters, find the responsibilities $\gamma(z_{nk})$

~~for fixed $\mu_1, \cdots, \mu_K$~~, ~~find $\{r_{nk}\}$ that minimize J~~

(M-step): maximization

for fixed responsibilities, find the corresponding $\left\{ \begin{array}{l} \boldsymbol{\mu}_k \\ \boldsymbol{\Sigma}_k \\ \pi_k \end{array} \right.$

~~for fixed $\{r_{nk}\}$~~, ~~find $\mu_1, \cdots, \mu_K$ that minimize J~~

general EM

complete dataset with latent variables

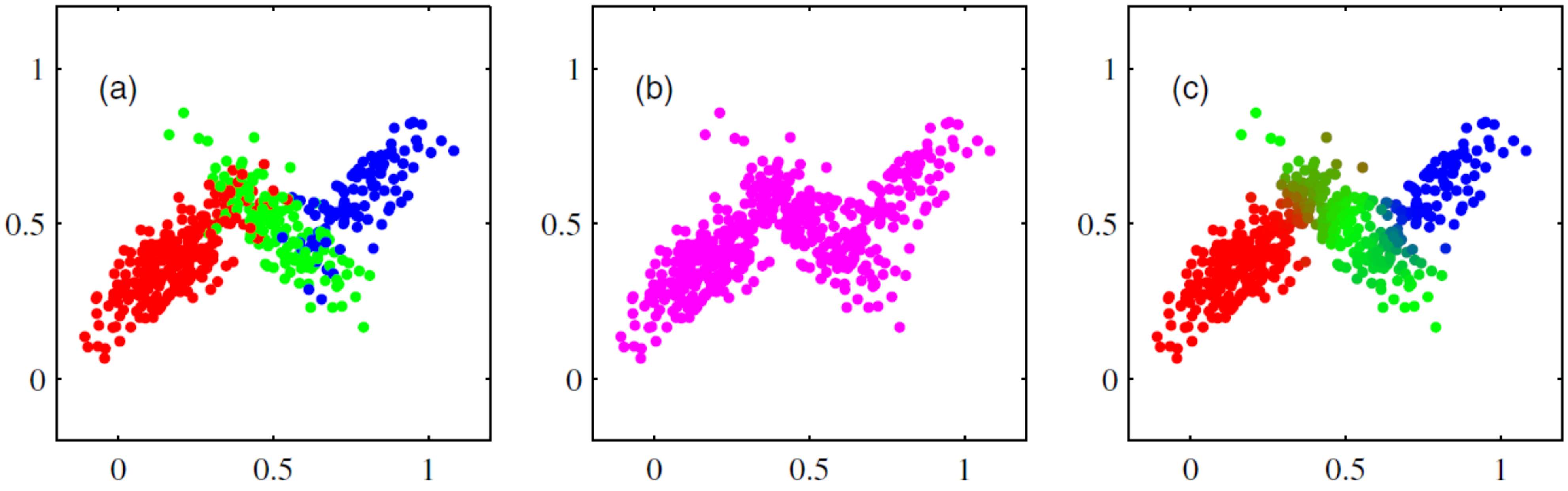
- Suppose $\mathbf{X}$ is the data matrix, and $\mathbf{Z}$ the corresponding latent variables (assumed to be discrete). Then

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

- $\{\mathbf{X}, \mathbf{Z}\}$ is called the complete data set; $\mathbf{X}$ is incomplete
- In practice, we are not given the complete data set; the only way we estimate $\mathbf{Z}$ is by the posterior

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

complete dataset with latent variables



complete data set with both $\{\mathbf{X}, \mathbf{Z}\}$

incomplete data set with only $\mathbf{X}$

$\mathbf{X}$ with the posterior $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$

Image taken from [Bi]

general EM

(E-step): expectation

1. for fixed parameters, find $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
2. calculate the expectation $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$

(M-step): maximization

solve for $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$

Questions?

Reference

- *K-means:*
  - *[Al] Ch.7.3*
  - *[HaTF] Ch.13.2.1*
  - *[Bi] Ch.9.1*
- *EM:*
  - *[Al] Ch.7.2, 7.4*
  - *[HaTF] Ch.13.2.3*
  - *[Bi] Ch.9.2-9.4*