Clustering and Mixture Models

Wenting Tu

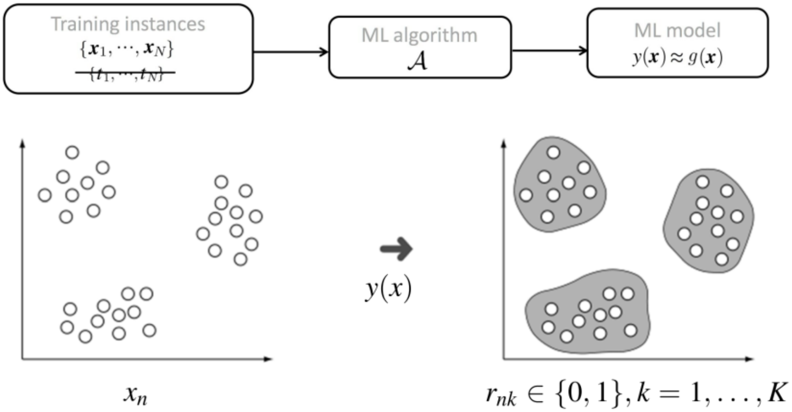
SHUFE, SIME

Machine Learning and Deep Lerning

Course No. 1638

Outline

Unsupervised Learning and Clustering



K-means

$$\boldsymbol{\mu}^*, \boldsymbol{r}^* = \arg\min \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|^2$$

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|^2$$

$$J = \sum_{k=1}^{K} r_{1k} \left\| \boldsymbol{x}_1 - \boldsymbol{\mu}_k \right\|^2 + \sum_{k=1}^{K} r_{2k} \left\| \boldsymbol{x}_2 - \boldsymbol{\mu}_k \right\|^2 + \cdots + \sum_{k=1}^{K} r_{Nk} \left\| \boldsymbol{x}_N - \boldsymbol{\mu}_k \right\|^2$$

$$\rightarrow r_{nk}^* = \begin{cases} 1 & \text{if } k = \arg\min_j \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_j \right\|^2 \\ 0 & \text{otherwise} \end{cases}$$

K-means

$$\boldsymbol{\mu}^*, \boldsymbol{r}^* = \arg\min \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|^2$$

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|^2$$

$$J = \sum_{n=1}^{N} r_{n1} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_1 \right\|^2 + \sum_{n=1}^{N} r_{n2} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_2 \right\|^2 + \cdots + \sum_{n=1}^{N} r_{nK} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_K \right\|^2$$

$$\to \boldsymbol{\mu}_k^* = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}}$$

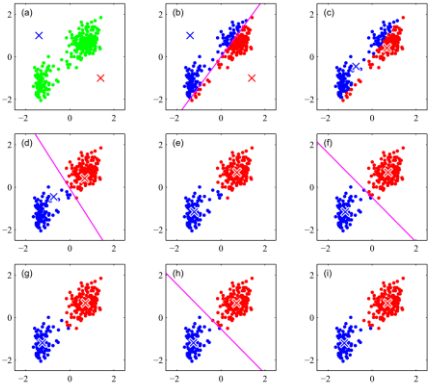
K-means

- choose some initial values for the $\boldsymbol{\mu}_k$
- repeated until convergence
- minimize J with respect to the r_{nk}

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_j \right\|^2 \\ 0 & \text{otherwise} \end{cases}$$

- minimize J with respect to the $\boldsymbol{\mu}_k$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}}$$



K-medoids

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}\left(\boldsymbol{x}_n, \boldsymbol{\mu}_k\right)$$

Others

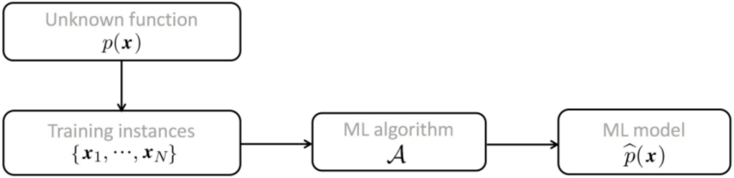
- Density-based Clustering
- Hierarchical Clustering

References:

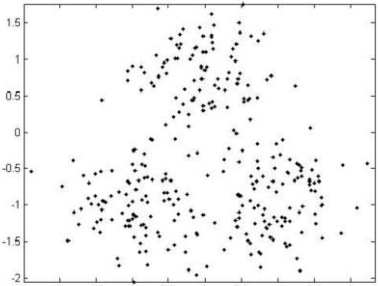
《机器学习》(chapters of *聚类*)

Outline

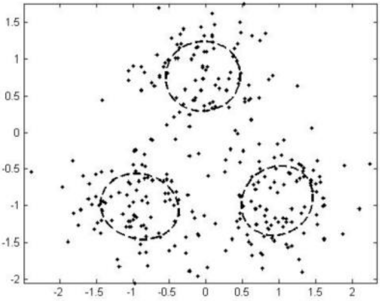
Unsupervised Learning

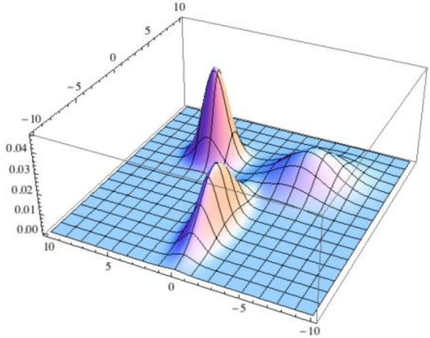


Mixtures of Gaussians

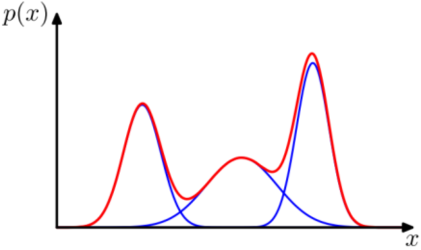


Mixtures of Gaussians





Mixtures of Gaussians



$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \sum_{k=1}^{K} \pi_k = 1 \quad 0 \leq \pi_k \leq 1$$

mixing coeficient   component of the mixture

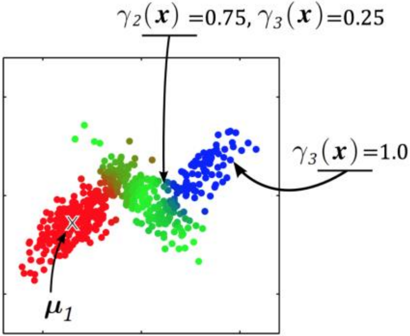
$$\boldsymbol{\pi} \equiv \{\pi_1, \ldots, \pi_K\} \quad \boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\} \quad \boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k\}$$

Mixtures of Gaussians

- Responsibilities and Soft Clustering

$$\gamma_k(\boldsymbol{x}) \equiv p(k|\boldsymbol{x})$$
$$= \frac{p(k)p(\boldsymbol{x}|k)}{\sum_l p(l)p(\boldsymbol{x}|l)}$$
$$= \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

(responsibilities)



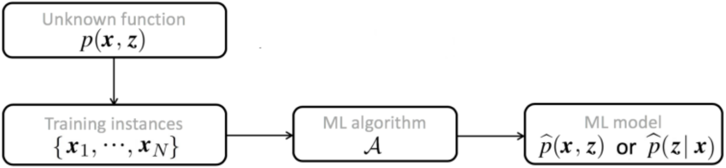
Maximum Likelihood Solution

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

Note: the maximum likelihood solution for the parameters no longer has a closed-form analytical solution.

Latent Variable

Introduce a K-dimensional binary random variable



$$z_k \in \{0, 1\}, \sum_k z_k = 1$$

$$p\left(\boldsymbol{x} \mid z_k = 1\right) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \quad p\left(z_k = 1\right) = \pi_k$$

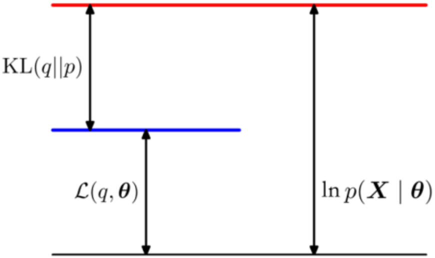
$$p(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)^{z_k}, \; p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

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

Expectation Maximization

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} - \sum_{Z} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

1st item: Evidence Lower Bound (ELBO), denoted as $\mathcal{L}(q, \boldsymbol{\theta})$
2nd item: Kullback-Leibler divergence, denoted as $\mathrm{KL}(q \parallel p)$



$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)} \ldots \longrightarrow \ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t+1)}\right) \geq \ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t)}\right)$$

Expectation Maximization

$$\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \ln p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}) + \ln p(\boldsymbol{X} \mid \boldsymbol{\theta})$$

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) - \ln p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$$

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \{\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) - \ln q(\mathbf{Z})\} - \{\ln p(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}) - \ln q(\mathbf{Z})\}$$

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} - \ln \left\{ \frac{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

Expectation Maximization

$$\ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t)}\right)$$

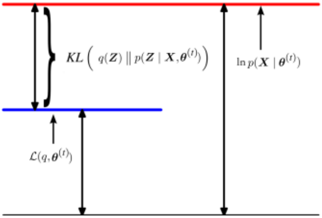
$$= \ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t+1)}\right)$$

Expectation Maximization

$$\ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t)}\right)$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p\left(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}\right)}{q(\mathbf{Z})} \right\} + \left(- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right)}{q(\mathbf{Z})} \right\} \right)$$

$$= \mathcal{L}\left(q, \boldsymbol{\theta}^{(t)}\right) + KL\left(q(\mathbf{Z}) \parallel p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right)\right)$$



$$= \ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t+1)}\right)$$

Expectation Maximization

$$\ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t)}\right)$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p\left(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}\right)}{q(\mathbf{Z})} \right\} + \left(- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right)}{q(\mathbf{Z})} \right\} \right)$$

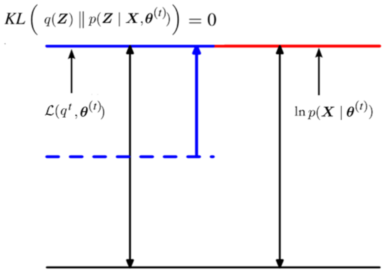
$$= \mathcal{L}\left(q, \boldsymbol{\theta}^{(t)}\right) + KL\left(q(\mathbf{Z}) \parallel p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right)\right)$$

$$q^t(\mathbf{Z}) = p\left(\mathbf{Z} \mid X, \theta^{(t)}\right) \rightarrow KL = 0$$

$$= \mathcal{L}\left(q^t, \boldsymbol{\theta}^{(t)}\right)$$

$$= \sum_{\mathbf{Z}} q^t(\mathbf{Z}) \ln \left\{ \frac{p\left(\boldsymbol{X}, Z \mid \boldsymbol{\theta}^{(t)}\right)}{q^t(\mathbf{Z})} \right\}$$

$$= \sum_{\mathbf{Z}} q^t(\mathbf{Z}) \ln p\left(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}\right) - \sum_{\mathbf{Z}} q^t(\mathbf{Z}) \ln q^t(\mathbf{Z})$$



$$= \ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t+1)}\right)$$

Expectation Maximization

$$\ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t)}\right)$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p\left(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}\right)}{q(\mathbf{Z})} \right\} + \left(- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right)}{q(\mathbf{Z})} \right\} \right)$$

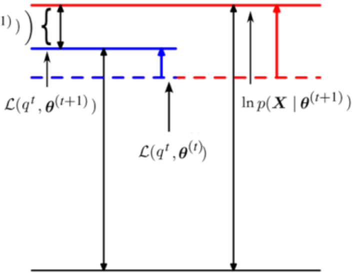
$$= \mathcal{L}\left(q, \boldsymbol{\theta}^{(t)}\right) + KL\left(q(\mathbf{Z}) \parallel p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right)\right)$$

$$q^t(\mathbf{Z}) = p\left(\mathbf{Z} \mid X, \theta^{(t)}\right) \rightarrow KL = 0$$

$$= \mathcal{L}\left(q^t, \boldsymbol{\theta}^{(t)}\right)$$

$$= \sum_{\mathbf{Z}} q^t(\mathbf{Z}) \ln \left\{ \frac{p\left(\boldsymbol{X}, Z \mid \boldsymbol{\theta}^{(t)}\right)}{q^t(\mathbf{Z})} \right\}$$

$$= \sum_{\mathbf{Z}} q^t(\mathbf{Z}) \ln p\left(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}\right) - \sum_{\mathbf{Z}} q^t(\mathbf{Z}) \ln q^t(\mathbf{Z})$$



$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(q^t, \boldsymbol{\theta}\right)$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} q^t(\mathbf{Z}) \ln p(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \rightarrow \mathcal{L}\left(q^t, \boldsymbol{\theta}^{(t)}\right) \leq \mathcal{L}\left(q^t, \boldsymbol{\theta}^{(t+1)}\right)$$

$$\leq \mathcal{L}\left(q^t, \boldsymbol{\theta}^{(t+1)}\right)$$

$$\leq \mathcal{L}\left(q^t, \boldsymbol{\theta}^{(t+1)}\right) + KL\left(q^t(\mathbf{Z}) \parallel p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t+1)}\right)\right)$$

$$= \ln p\left(\boldsymbol{X} \mid \boldsymbol{\theta}^{(t+1)}\right)$$

Expectation Maximization

- Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
- **E step** Evaluate $p(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$
- **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$

$$\boldsymbol{\theta}^{\text{new}} = \arg\max \mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)$$

where

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \sum_{\mathbf{Z}} p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}\right) \ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$$

- Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

and return to **E step**

EM for Mixtures of Gaussians

$$
\begin{aligned}
\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) &= \sum_{\mathbf{Z}} \ln p(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}) p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}\right) \\
&= \sum_{z_1} \cdots \sum_{z_N} \left\{ \sum_{n=1}^{N} \ln p\left(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}\right) \prod_{n=1}^{N} p\left(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}^{\text{old}}\right) \right\}
\end{aligned}
$$

EM for Mixtures of Gaussians

$$
\begin{aligned}
\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) &= \sum_{\mathbf{Z}} \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \\
&= \sum_{z_1} \cdots \sum_{z_N} \left\{ \sum_{n=1}^{N} \ln p\left(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}\right) \prod_{n=1}^{N} p\left(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}^{\text{old}}\right) \right\}
\end{aligned}
$$

$$
\frac{\partial \mathcal{Q}}{\partial \pi_k} = 0 \quad \text{s.t.} \ \sum_{k=1}^{K} \pi_k = 1
$$

$$
\pi_k^{\text{new}} = \frac{N_k}{N}
$$

$$
\gamma\left(z_{nk}\right) = \frac{\pi_k^{\text{old}} \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}}\right)}{\sum_{j=1}^{K} \pi_j^{\text{old}} \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}}\right)}
$$

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

EM for Mixtures of Gaussians

$$
\begin{aligned}
\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) &= \sum_{\mathbf{Z}} \ln p(\boldsymbol{X}, \mathbf{Z} \mid \boldsymbol{\theta}) p\left(\mathbf{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}\right) \\
&= \sum_{z_1} \cdots \sum_{z_N} \left\{ \sum_{n=1}^{N} \ln p\left(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}\right) \prod_{n=1}^{N} p\left(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}^{\text{old}}\right) \right\}
\end{aligned}
$$

$$
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\mu}_k} = 0 \quad \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\Sigma}_k} = 0
$$

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

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

EM for Mixtures of Gaussians

- Initialize $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$
- **E step**. $\gamma\left(z_{nk}\right) = \frac{\pi_k^{\text{old}} \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}}\right)}{\sum_{j=1}^{K} \pi_j^{\text{old}} \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}}\right)}$
- **M step**. $N_k = \sum_{n=1}^{N} \gamma\left(z_{nk}\right), \pi_k^{\text{new}} = \frac{N_k}{N}$

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

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

- Check for convergence

  If the convergence criterion is not satisfied, then let

$$\left(\pi_k^{\text{old}}, \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}}\right) \leftarrow \left(\pi_k^{\text{new}}, \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}\right)$$

  and return to **E step**

Outline

Semi-supervised Learning (SSL)