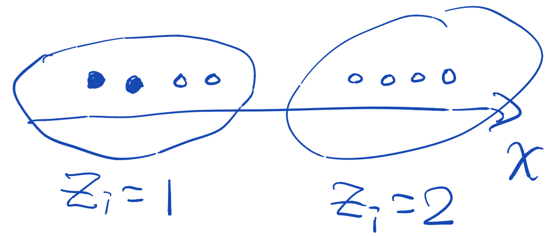
Expectation Maximization (EM)

Qiang Liu
UT Austin

Clustering

- **Inputs**: n objects (data points) $\{x_i\}_{i=1}^n$ and a number K of clusters.
- **Goal:** Group the points into several groups
  - Deciding a group ID, $z_i \in \{1, \ldots, K\}$, for each data point x_i.



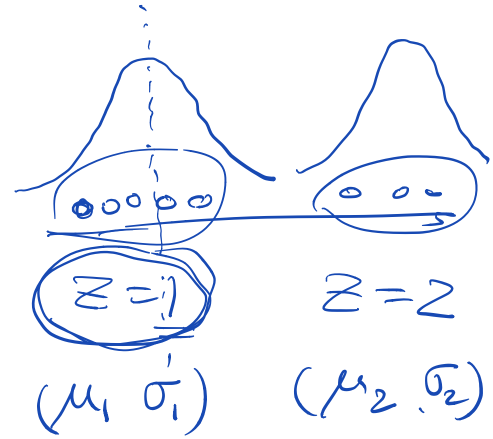
Probabilistic Modeling of Clustering

- **Probabilistic Approach:**
  - Assume a joint distribution $p(x, z \mid \theta)$ that generates data and labels (x, z).
  - Estimate parameter θ.
  - Infer the labels from the posterior distribution $p(z \mid x;\ \theta)$.

$$P(x, z) = P(z)\,P(x \mid z)$$

① $$P(z=k) = \pi_k \quad \left(\begin{array}{l} \pi_k \geq 0 \\ \sum_k \pi_k = 1 \end{array} \right)$$

② $$P(x \mid z=k) = \mathcal{N}(x \mid \mu_k\ \sigma_k^2)$$



$$P(x, z=k) = P(z=k)\,P(x \mid z=k)$$
$$= \pi_k\, \mathcal{N}(x \mid \mu_k, \sigma_k^2)$$

$$\theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}$$

$$P(x, z \mid \theta)$$

MLE with Complete Information

- If we observe both data and labels $\{x_i, z_i\}_{i=1}^n$ **(complete information)**, we can estimate the parameter θ by maximizing the **joint probability**:

$$\max_\theta \sum_{i=1}^n \log p(x_i,\ z_i \mid \theta)$$

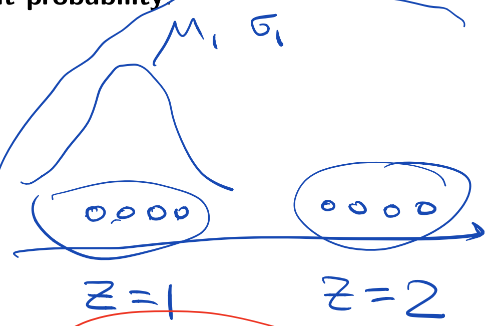
$$p(x_i, z_i = k \mid \theta) = \pi_k \mathcal{N}(x \mid \mu_k, \sigma_k^2)$$

$$\sum_{i=1}^n \sum_{k=1}^K \mathbb{I}(z_i = k) \log p(x_i, z_i = k \mid \theta)$$

$$= \sum_{i=1}^n \sum_{k=1}^K \mathbb{I}(z_i = k) \log\left(\pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)\right)$$

$$= \sum_{i=1}^n \sum_{k=1}^K \mathbb{I}(z_i = k) \left(\log \pi_k + \log \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)\right)$$

$$\pi_k = \frac{\sum_{i=1}^n \mathbb{I}(z_i = k)}{n}$$



$$\pi_1 = \frac{\#(z=1)}{n} = \frac{1}{2}$$

$$\mu_1 = \frac{\sum_{i=1}^n \mathbb{I}(z_i = 1)\, x_i}{\sum_{i=1}^n \mathbb{I}(z_i = 1)}$$

$$\sigma_1^2 = Var(\{x_i \mid z_i = 1\})$$

Clustering and Mixture Models

- If we only observe data $\{x_i\}_{i=1}^n$ (**incomplete information**), we shall estimate θ by maximizing the **marginal probability**:

$$\sum_{i=1}^n \log p(x_i \mid \theta) = \sum_{i=1}^n \log \sum_{z_i} p(x_i, z_i \mid \theta)$$

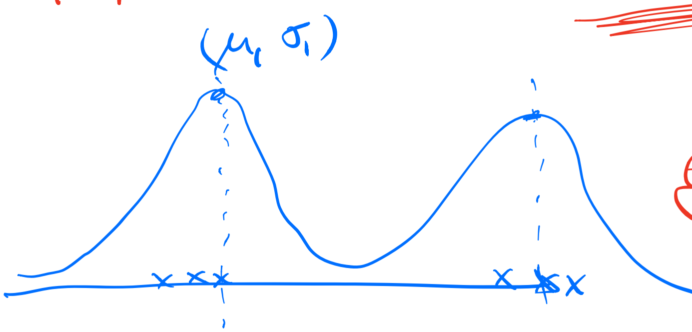
$$\sum_{i=1}^n \log P(x_i\ z_i \mid \theta)$$

$$P(x \mid \theta) = \sum_k P(x, z=k \mid \theta)$$

$$= \sum_k \pi_k \mathcal{N}(x \mid \mu_k, \sigma_k^2)$$

mixture of Gaussian distributions

$$\sum_{i=1}^n \log P(x_i \mid \theta) = \sum_{i=1}^n \log\Big(\sum_{k=1}^K \pi_k \mathcal{N}(x \mid \mu_k, \sigma_k^2)\Big)$$



$$\max_{\{\pi_k,\ \mu_k,\ \sigma_k\}_{k=1}^K}$$

$$P(x \mid \theta) = \frac{1}{2} \exp\Big(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\Big)\Big(\frac{1}{\sqrt{2\pi}\,\sigma_1}\Big)$$

$$+ \frac{1}{2} \exp\Big(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\Big)\Big(\frac{1}{\sqrt{2\pi}\,\sigma_2}\Big)$$

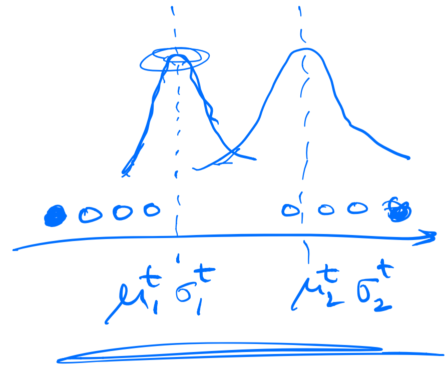
Expectation Maximization: Algorithm Procedure

Expectation Maximization (EM) is an iterative method for maximizing the marginal likelihood.

$$\max_{\theta} \sum_{i=1}^{n} \log p(x_i \mid \theta)$$

- □ Initialize parameter θ_0.
- □ For iteration t:
  - ○ Given θ_t, "impute" the missing labels by drawn samples from the posterior distribution

$$z_i \sim p(z \mid x_i;\ \theta_t).$$



  - ○ Update θ by maximizing the expected joint likelihood:

$$\theta^{t+1} = \arg\max_{\theta} \sum_{i=1}^{n} \mathbb{E}_{z_i \sim p(\cdot \mid x_i;\ \theta_t)}[\log p(x_i, z_i \mid \theta)].$$

$$\theta_t = [\pi_k^t, \mu_k^t, \sigma_k^t]_{k=1}^{K}$$

$$p(z_i = k \mid x_i, \theta_t) = \frac{P(x_i, z_i = k \mid \theta_t)}{\sum_{\ell} P(x_i, z_i = \ell \mid \theta_t)}$$

$$= \frac{\pi_k \mathcal{N}(x_i \mid \mu_k^t\ \sigma_k^t)}{\sum_{\ell} \pi_\ell \mathcal{N}(x_i \mid \mu_\ell^t\ \sigma_\ell^t)}$$

$$P(z_i = 1 \mid x_i\ \theta_t)$$

$$= \frac{\pi_1 \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}\right) / (\sqrt{2\pi}\ \sigma_1)}{\sum_{\ell} \pi_\ell \exp\left(-\frac{(x_i - \mu_\ell)^2}{2\sigma_\ell^2}\right) / (\sqrt{2\pi}\ \sigma_2)}$$

EM algorithm:

$$\theta^{t+1} = \arg\max_{\theta} \sum_{i=1}^{n} \mathbb{E}_{z_i \sim p(\cdot|x_i;\ \theta_t)}[\log p(x_i,\ z_i \mid \theta)]. \triangleq L(\theta)$$

Define $\boxed{\gamma_{ik}^{t} = P(z_i = k \mid x_i, \theta_t)}$

$$L(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik}^{t} \log P(x_i, z_i = k \mid \theta) .$$

$$\Rightarrow \quad \pi_k^{t+1} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{t}}{n}$$

$$\mu_k^{t+1} = \left(\sum_{i=1}^{n} \gamma_{ik}^{t} x_i\right) \Big/ \left(\sum_{i=1}^{n} \gamma_{ik}^{t}\right)$$

$$\sigma_k^{t+1} = \cdots$$