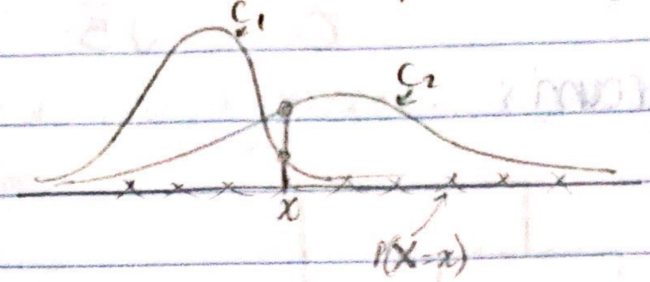
/26 Soft Clustering:

- Assign points to clusters with some probability

1) $N(\mu, \sigma) \to x | C_1$
2) $N(\mu_2, \sigma_2) \to x | C_2$



Mixture Models:

- x follows a mixture model with K mixture components if probability distribution of x is

$$P(X=x) = \sum_{j=1}^{K} P(C_j) P(X=x|C_j)$$

Normal / Gaussian

$$P(X=x) = P(C_1) P(X=x|C_1) + P(C_2) P(X=x|C_2)$$

$$P(C_1) + P(C_2) = 1$$

Gaussian Mixture Model

$$P(X=x) = \sum_{j=1}^{K} P(C_j) \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma_j}\right)^2}$$

$C_j \sim N(\mu_j, \sigma_j)$ $\quad N(\mu_j, \Sigma_j)$

Our data consists of $x_1, x_2, \ldots, x_n$ observations

$$P(x_1, x_2, \ldots, x_n) = P(x_1) P(x_2) \ldots P(x_n)$$

$$\prod_{i=1}^{n} \sum_{j=1}^{K} P(C_j) P(x_i = x | C_j)$$

Notice that a given cluster / component C_j is characterized by $P(C_j)$, μ_j, σ_j

$P(C_j)$: probability of being in the cluster

μ_j, σ_j: mean and standard deviation of the cluster

Define

$$\theta = \begin{Bmatrix} P(C_1), \dots, P(C_K) \\ \mu_1, \dots, \mu_K \\ \sigma_1, \dots, \sigma_K \end{Bmatrix}$$

* θ has all the parameters we want

* we want to maximize the likelihood $(\theta \mid x_1, \dots, x_n)$

Goal: Find $\theta^* = \operatorname{argmax} L(\theta \mid x_1, \dots, x_n)$

$$= \operatorname{argmax}_\theta \underbrace{\prod_{i=1}^{n} \sum_{j=1}^{K} P(C_j)\, P(x_i = x \mid C_j)}_{\text{(joint distribution)}}$$

* we want to maximize the probability of generating the data we saw

$\operatorname{argmax} \log (L(\theta))$

$\operatorname{argmax} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{K} P(C_j)\, P(x_i = x \mid C_j) \right) \leftarrow \ell(\theta)$

$\frac{d}{d\mu} \ell(\theta) = 0 \rightarrow$ find value of μ that makes this possible

$$\hat{\mu}_{j\,MLE} = \frac{\sum_{i=1}^{n} P(C_j \mid x_i)\, x_i}{\sum_{i=1}^{n} P(C_j \mid x_i)} \qquad \hat{\sigma}_{MLE} = \frac{\sum_{i=1}^{n} P(C_j \mid x_i)(x_i - \hat{\mu}_{MLE})^2}{\sum_{i=1}^{n} P(C_j \mid x_i)}$$

MLE ↓ maximum

$$\hat{P}(C_j) = \frac{1}{n} P(C_j | x_i)$$

$$P(C_j | x_i) = \frac{P(x_i | C_j) P(C_j)}{P(x_i)} = \frac{P(x_i | C_j) P(C_j)}{\sum_{j=1}^{K} P(C_j) P(x_i | C_j)}$$ ← Bayes' Rule

↓

probability that x_i was drawn from C_j

Expectation-Maximization Algorithm

1) Start with a random θ
2) Estimate $P(C_j | x_i)$
3) From 2) Compute estimates for μ_j's and σ_j's and $P(C_j)$'s
4) Repeat 2 and 3 until convergence
   (it will always converge)