Gaussian Mixture Clustering

INSTRUCTOR: HONGJIE CHEN

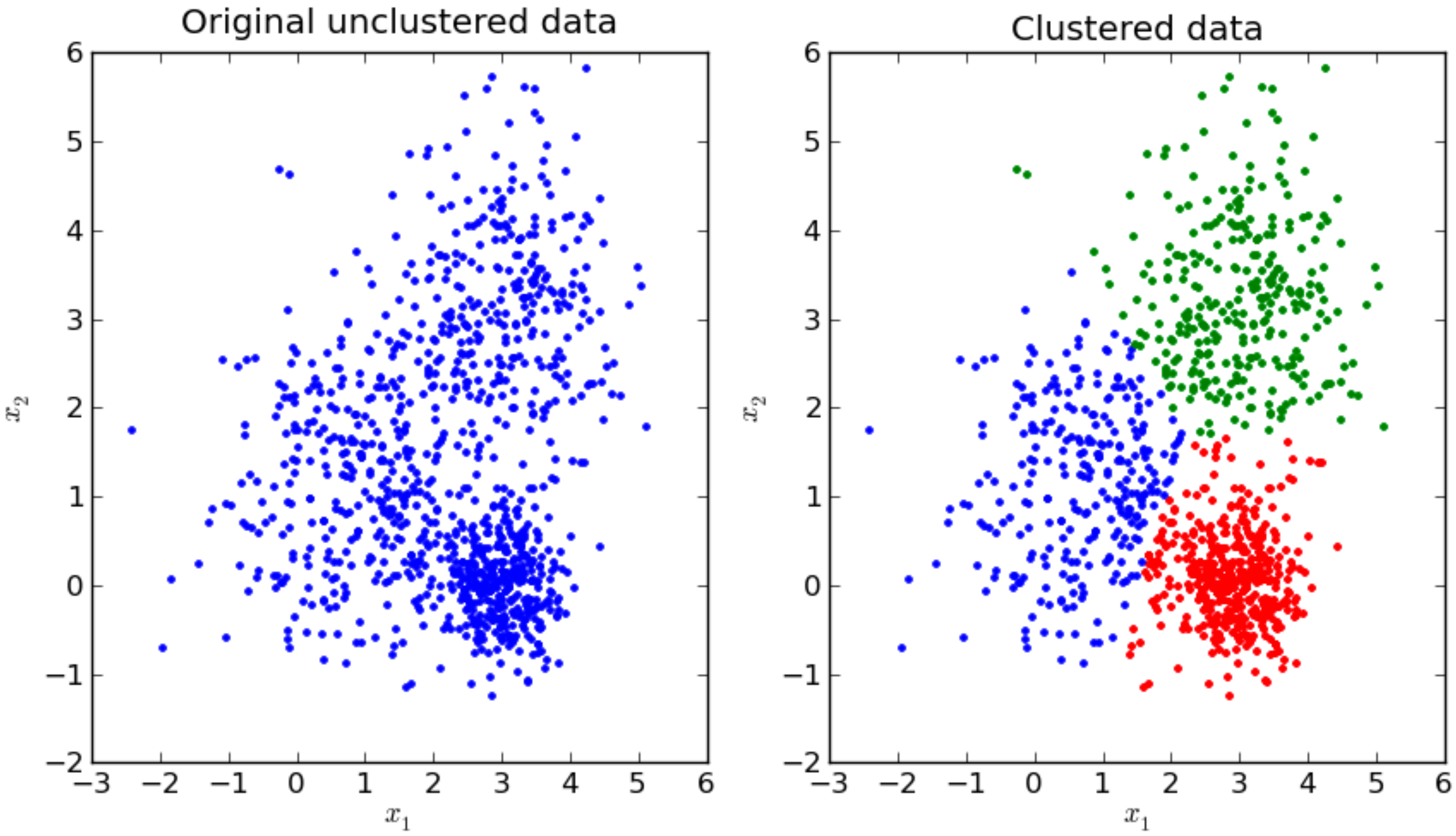
JUNE 16TH 2022

Unsupervised Learning

- Supervised Learning: training data have labels

- Unsupervised Learning: training data do not have labels
  - Clustering similar items
    - Products
    - Company profiles
    - Documents
    - News
    - Among many others…

Clustering Data



Mixture Distribution

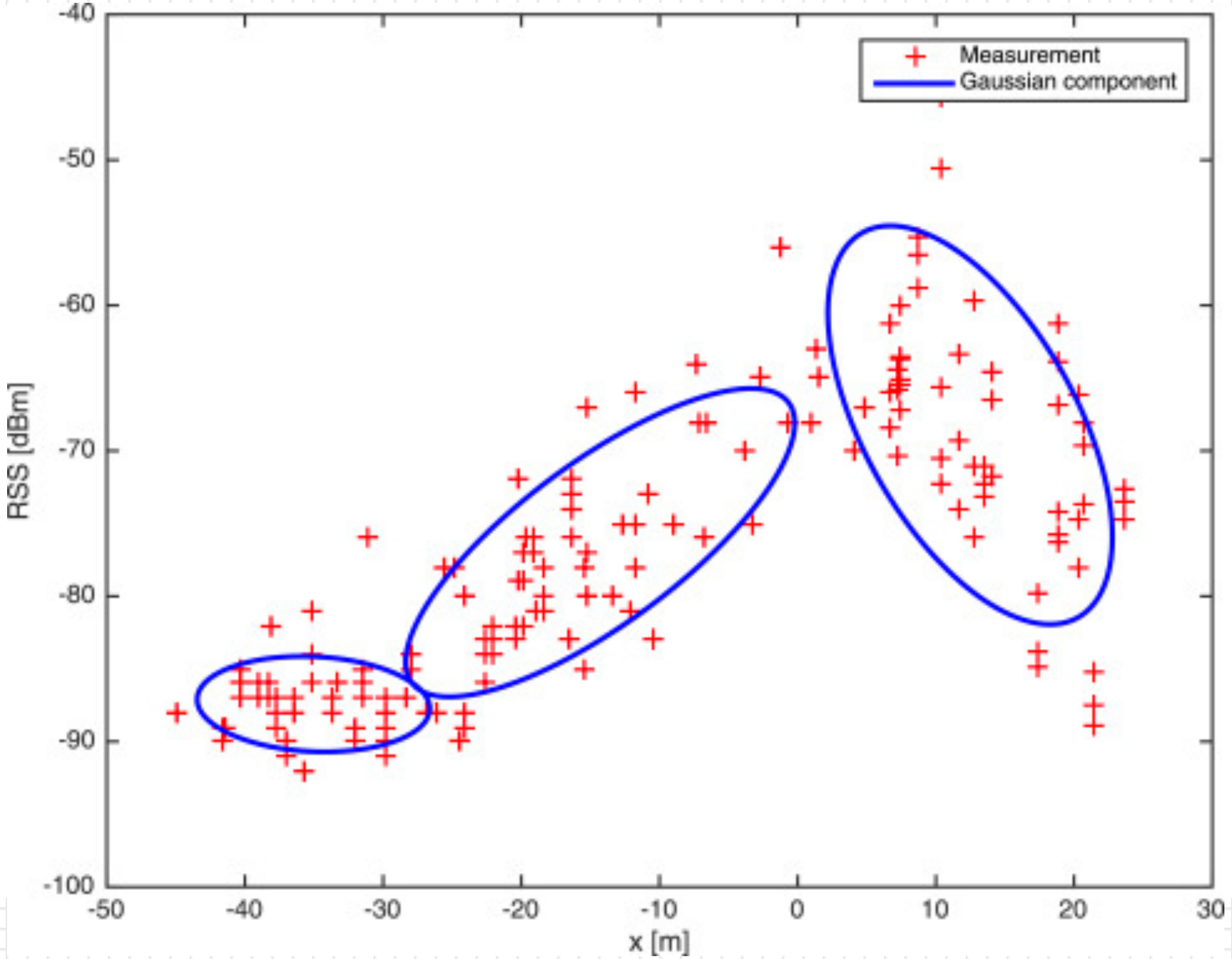
- Model joint distribution $P(X_1, \ldots, X_n)$ as mixture of multiple distributions

- Use discrete-valued R.V. Z to indicate which distribution is being used for each random draw

$$P(X_1, \ldots, X_n) = \sum_i P(Z = i) P(X_1, \ldots, X_n | Z)$$

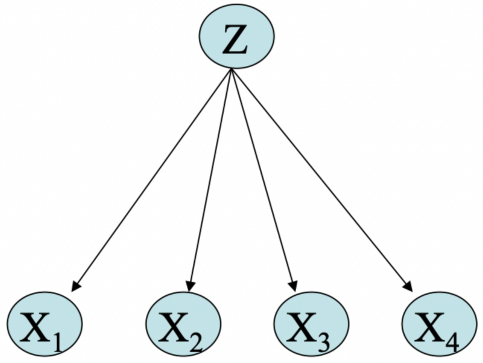
- *Gaussian* Mixtures
  - Assume each data point is generated by one of the Gaussian by
    - Randomly choose i(th) Gaussian i, based on $P(Z = i)$
    - Randomly generate a data point based on i(th) Gaussian

Gaussian Mixture Diagram



Gaussian Mixture Clustering with EM

- Assume n obversed R.V. $X = (X_1 \ldots X_n)$, and X_i are conditionally independent given Z
  - $$P(X \mid Z = j) = \prod_i N(X_i \mid \mu_{ji}, \sigma_{ji})$$



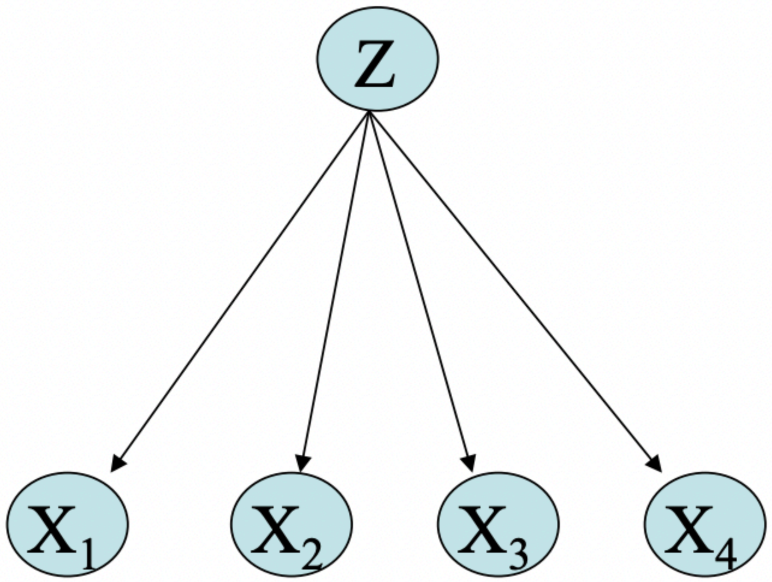
- Assume only two clusters, i.e., $j \in \{1,2\}$ and $\forall i, j,\ \sigma_{ji} = \sigma$
  - $$P(X) = \sum_{j=1}^{2} P(Z = j \mid \pi) \prod_i N(x_i \mid \mu_{ji}, \sigma)$$

π is the probability distribution of which Z is sampled from

- Assume σ known, $\pi_1 \ldots \pi_K, \mu_{1i} \ldots \mu_{Ki}$ unknown

EM Algorithm with Maths

- Iterate until convergence
  - E step: Calculate $P(Z|X, \theta)$
  - M step: $\theta \leftarrow \text{argmax}_{\theta'} Q(\theta'|\theta)$
  - $\theta = (\pi, \mu_{ji})$

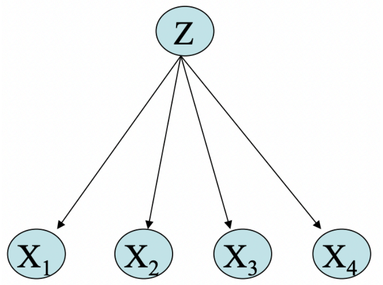


- Guaranteed to find local maximum
  - Each iteration increases $\mathbb{E}_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$

E step in Gaussian Mixture Clustering

Calculate $P(Z(n) \mid X(n), \theta)$ for each observed example $X(n)$,
$X(n) = (x_1(n), x_2(n), \ldots, x_T(n)), T = 4$

$$P(z(n) = k \mid x(n), \theta) = \frac{P(x(n) \mid z(n) = k, \theta)P(z(n) = k \mid \theta)}{\sum_{j=0}^{1} P(x(n) \mid z(n) = j, \theta)P(z(n) = j \mid \theta)}$$



$$= \frac{\prod_i P(x_i(n) \mid z(n) = k, \theta)P(z(n) = k \mid \theta)}{\sum_{j=0}^{1} \prod_i P(x_i(n) \mid z(n) = j, \theta)P(z(n) = j \mid \theta)}$$

$$= \frac{\prod_i N(x_i(n) \mid \mu_{k,i}, \sigma)(\pi^k(1-\pi)^{(1-k)})}{\sum_{j=0}^{1} \prod_i N(x_i(n) \mid \mu_{j,i}, \sigma)(\pi^j(1-\pi)^{(1-j)})}$$

M Step in Gaussian Mixture Clustering for π

- $Q(\theta'|\theta) = \mathbb{E}_{P(Z|X,\theta)}[\log P(X,Z|\theta')] = \mathbb{E}[\log P(X|Z,\theta') + \log P(Z|\theta')]$
- $\pi \leftarrow \text{argmax}_{\pi'} \mathbb{E}_{P(Z|X,\theta)}[\log P(Z|\pi')]$

$\theta = (\pi, \mu_{ji})$

Indepdendent of π', cause given Z

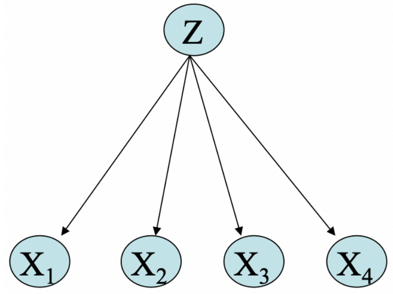
$$\mathbb{E}_{P(Z|X,\theta)}[\log P(Z|\pi')] = \mathbb{E}_{P(Z|X,\theta)}\left[\log(\pi'^{\sum_n z(n)}(1-\pi')^{\sum_n (1-z(n))})\right]$$

$$= \mathbb{E}_{P(Z|X,\theta)}\left[\sum_n z(n)\log \pi' + (\sum_n (1-z(n)))\log(1-\pi')\right]$$

$$= (\sum_n \mathbb{E}_{P(Z|X,\theta))}[z(n)])\log \pi' + (\sum_n \mathbb{E}_{P(Z|X,\theta))}[1-z(n)])\log(1-\pi')$$

- $$\frac{\partial \mathbb{E}_{P(Z|X,\theta)}[\log P(Z|\pi')]}{\partial \pi'} = (\sum_n \mathbb{E}_{P(Z|X,\theta))}[z(n)])\frac{1}{\pi'} + (\sum_n \mathbb{E}_{P(Z|X,\theta))}[1-z(n)])\frac{-1}{1-\pi'}$$

- $$\pi \leftarrow \frac{\sum_n \mathbb{E}[z(n)]}{\sum_n \mathbb{E}[z(n)] + \sum_n (1-\mathbb{E}[z(n)])} = \frac{1}{N}\sum_n \mathbb{E}[z(n)]$$



M Step in Gaussian Mixture Clustering for μ_{ji}

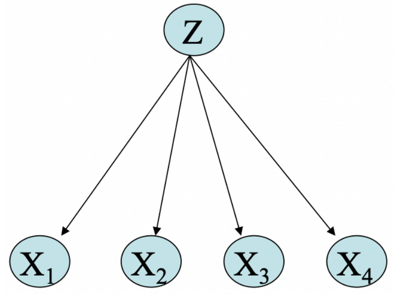
- $Q(\theta'|\theta) = \mathbb{E}_{P(Z|X,\theta)}[\log P(X,Z|\theta')] = \mathbb{E}[\log P(X|Z,\theta') + \log P(Z|\theta')]$
- $\mu_{ji} \leftarrow \text{argmax}_{\mu'_{ji}} \mathbb{E}_{P(Z|X,\theta)}[\log P(X|Z,\theta')]$

$\theta = (\pi, \mu_{ji})$

Indepdendent of μ'_{ji}

- Omitted Calculation…

- $$\mu_{ji} \leftarrow \frac{\sum_n P(z(n) = j \mid x(n), \theta) x_i(n)}{\sum_n P(z(n) = j \mid x(n), \theta)}$$
- Compare to MLE if Z is known $\mu_{ji} \leftarrow \frac{\sum_n \delta(z(n) = j) x_i(n)}{\sum_n \delta(z(n) = j)}$



VIRGINIA TECH

EM in Gaussian Mixture Clustering

- $Q(\theta' | \theta) = \mathbb{E}_{P(Z|X,\theta)}[\log P(X, Z | \theta')]$ where $\theta = (\pi, \mu_{ji})$
- E step: Calulcate the expectation for each unobserved variables

$$P(z(n) = k | x(n), \theta) = \frac{\prod_i N(x_i(n) | \mu_{k,i}, \sigma)(\pi^k(1-\pi)^{(1-k)})}{\sum_{j=0}^{1} \prod_i N(x_i(n) | \mu_{j,i}, \sigma)(\pi^j(1-\pi)^{(1-j)})}$$

- M step: Update $\theta \leftarrow$ argmax$_{\theta'} Q(\theta' | \theta)$

$$\pi \leftarrow \frac{\sum_n \mathbb{E}[z(n)]}{\sum_n \mathbb{E}[z(n)] + \sum_n (1 - \mathbb{E}[z(n)])} = \frac{1}{N} \sum_n \mathbb{E}[z(n)]$$

$$\mu_{ji} \leftarrow \frac{\sum_n P(z(n) = j | x(n), \theta) x_i(n)}{\sum_n P(z(n) = j | x(n), \theta)}$$

Demo: link