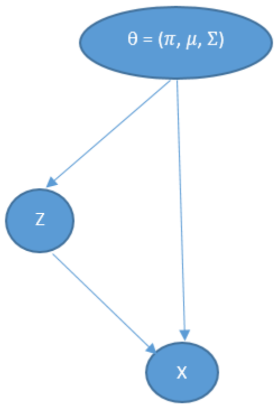
Gaussian-Mixture-Model Gaussian Mixture Model
This code implements the Gaussian Mixture Model.

- Problem: Given a number of documents, find a way to group them in to K topics (or clusters).
- Datasets: Processed 20-news-groups datasets, with each document contain docID and its processed words.
- Approach: We suppose document X is a multi-dimensional random variable generated from a process with parameter θ. We learn θ from observable samples of X.

Suppose there are K clusters. Assuming that each document X is a N-dimensional Gaussian random variable and X belongs to k^{th} cluster with probability π_k
The following process describes the generation of X:



Z is a discrete random variable indicating the cluster of document X.

$\pi = (\pi_1, \pi_2, ..., \pi_K)$ is a distribution over clusters.

$\mu = (\mu_1, \mu_2, ..., \mu_K)$ represents the centroids of each cluster. (or the mean parameters of each Gaussian)

$\Sigma = (\Sigma_1, \Sigma_2, ..., \Sigma_K)$ is the variance of each Gaussian.

Suppose there are m documents and X is a N-dimensional Gaussian random variable representing the document.

Generation process:

$$Z \sim Multinomial(\pi) \to k$$

$$X|Z, \theta \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Having observed many samples of X, we need to find $\theta = (\pi, \mu, \Sigma)$

To estimate θ, we will use Maximum Likelihood Estimation. Likelihood function:

$$P(X|\theta) = \prod_{i=1}^{m} P(X_i) = \prod_{i=1}^{m}(\sum_{k=1}^{K} P(Z_i = k)P(X_i|Z_i = k))$$
$$= \prod_{i=1}^{m}(\sum_{k=1}^{K} \pi_k P(X_i|Z_i = k))$$
$$= \prod_{i=1}^{m}(\sum_{k=1}^{K} \pi_k.\mathcal{N}(X_i|\mu_k, \Sigma_k))$$

We take the log of both side, we get the log likelihood function:

$$logP(X|\theta) = \sum_{i=1}^{m} log \sum_{k=1}^{K} \pi_k \mathcal{N}(X_i|\mu_k, \Sigma_k)$$

From here, we will lower the log likelihood to find a close optimization result.
Let $\gamma_{ik} = P(Z_i = k|X_i)$

$$\gamma_{ik} = \frac{P(Z_i = k, X_i)}{P(X_i)} = \frac{P(Z_i = k)P(X_i|Z_i = k)}{\sum_{k=1}^{K} P(Z_i = k)P(X_i|Z_i = k)} \propto \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

We have:

$$logP(X|\theta) = \sum_{i=1}^{m} log \sum_{k=1}^{K} \frac{\pi_k \mathcal{N}(X_i|\mu_k, \Sigma_k)}{q(Z_i = k)} q(Z_i = k) \geq \sum_{i=1}^{m} \sum_{k=1}^{K} q(Z_i = k) log \frac{\pi_k \mathcal{N}(X_i|\mu_k, \Sigma_k)}{q(Z_i = k)}$$

Let $\alpha_{ik} = q(Z_i = k)$, we have:

$$logP(x|\theta) \geq \sum_{i=1}^{m} \sum_{k=1}^{K} \alpha_{ik} log \frac{\pi_k \mathcal{N}(X_i|\mu_k, \Sigma_k)}{\alpha_{ik}}$$

$$\implies log(X|\theta) \geq \sum_{i=1}^{m} \sum_{k=1}^{K} \alpha_{ik}(log\pi_k + log\mathcal{N}(X_i|\mu_k, \Sigma_k) - log\alpha_{ik}) = LLB(X|\theta)$$

We need to find α, π, μ, Σ so that $LLB(X|\theta) \to$ max.

Optimization:

E-step: $\alpha_{ik} = \gamma ik$

$$LLB(X|\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} \gamma_{ik}(log\pi_k + log\mathcal{N}(X_i|\mu_k, \Sigma_k) - log\gamma_{ik})$$

M-step:

- $\pi_k = \frac{\sum_i \gamma_{ik}}{\sum_k \sum_i \gamma_{ik}} = \frac{\sum_i \gamma_{ik}}{m}$
- $\mu_k = \frac{\sum_i \gamma_{ik} x_i}{\sum_i \gamma_{ik}}$
- $\Sigma_k = \frac{\sum_i \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i \gamma_{ik}}$

Program the above step until convergence, we get π, μ, Σ