K-Means Clustering and Gaussian Mixture Model

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Weekly Objectives

- Understand the clustering task and the K-means algorithm
  - Know what the unsupervised learning is
  - Understand the K-means iterative process
  - Know the limitation of the K-means algorithm
- Understand the Gaussian mixture model
  - Know the multinomial distribution and the multivariate Gaussian distribution
  - Know why mixture models are useful
  - Understand how the parameter updates are derived from the Gaussian mixture model
- Understand the EM algorithm
  - Know the fundamentals of the EM algorithm
  - Know how to derive the EM updates of a model

Expectation of GMM

- Similar problem of K-means algorithm
  - Two interacting parameters
  - As before, we apply the expectation and the maximization algorithm
    - Expectation: the assignment between the clusters and the data points
    - Maximization: the update of the parameters
- Expectation step
  - Assign a data point to a nearest cluster → the assignment probability
    - Given the parameters and the data point, calculate the likelihood
  - $\gamma(z_{nk}) \equiv p(z_k = 1|x_n) = \frac{P(z_k=1)P(x|z_k = 1)}{\sum_{j=1}^{K} P(z_j=1)P(x|z_j = 1)} = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$
    - Here, x, π, μ, Σ are given, calculate $\gamma(z_{nk})$
    - $\gamma(z_{nk})$ are used to calculate π, μ, Σ
    - The new $\gamma(z_{nk})$ motivates the update of the old parameters

Maximization of GMM

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

$$\mathbf{ln}N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{C}$$

$$\mathbf{ln}N(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) + \boldsymbol{C}$$

$$\frac{d}{d\mu}\ln N(X|\mu,\Sigma) = 0 \rightarrow -\frac{1}{2}\times 2\times -1\times\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}}) = 0 \rightarrow \widehat{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n}{N}$$

$$\frac{d}{d\boldsymbol{\Sigma}^{-1}}\ln N(X|\mu,\Sigma) = 0 \rightarrow \widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})^{\boldsymbol{T}}$$

$$\gamma(z_{nk}) = \frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)}$$

- Maximization step
  - Update the parameters given $\gamma(z_{nk})$
  - Parameters to update: π,μ,Σ
    - $\ln P(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)\}$
    - Typical methods
      - Derivative → set the equation to zero when the function is smooth
      - Lagrange method when there is a constraint. Which parameter has the constraint?
  - $\frac{d}{d\mu_k}\ln P(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\widehat{\mu_k})=0$

$$\rightarrow \sum_{n=1}^{N}\gamma(z_{nk})\,(\boldsymbol{x}_n-\widehat{\mu_k})=0 \rightarrow \widehat{\mu_k} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\,\boldsymbol{x}_n}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

  - $\frac{d}{d\Sigma_k}\ln P(X|\pi,\mu,\Sigma) = 0$

$$\rightarrow \Sigma_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\,(\boldsymbol{x}_n-\widehat{\mu_k})(\boldsymbol{x}_n-\widehat{\mu_k})^{\boldsymbol{T}}}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

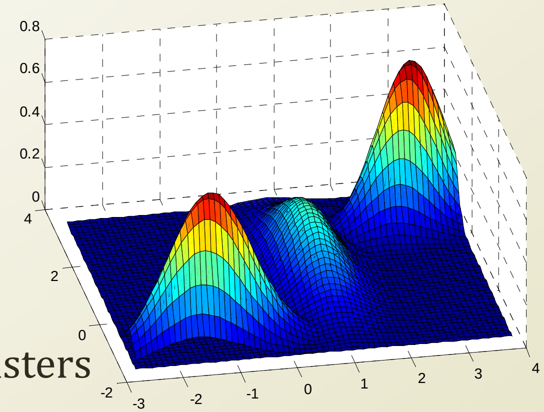
  - $\frac{d}{d\pi_k}\ln P(X|\pi,\mu,\Sigma) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) = 0$

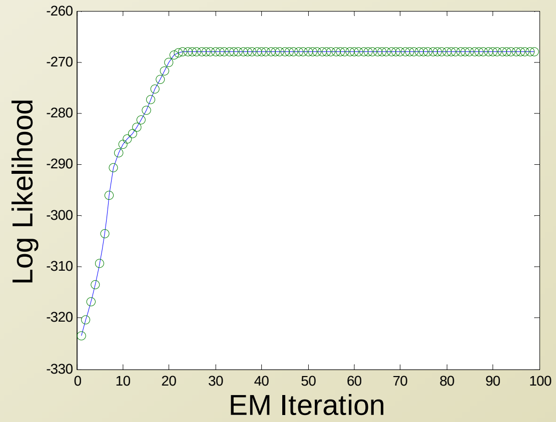
$$\rightarrow \sum_{n=1}^{N}\frac{N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)} + \lambda = 0 \rightarrow \sum_{k=1}^{K}\left\{\sum_{n=1}^{N}\frac{\pi_k N(x|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x|\mu_j,\Sigma_j)} + \pi_k\lambda\right\} = 0$$

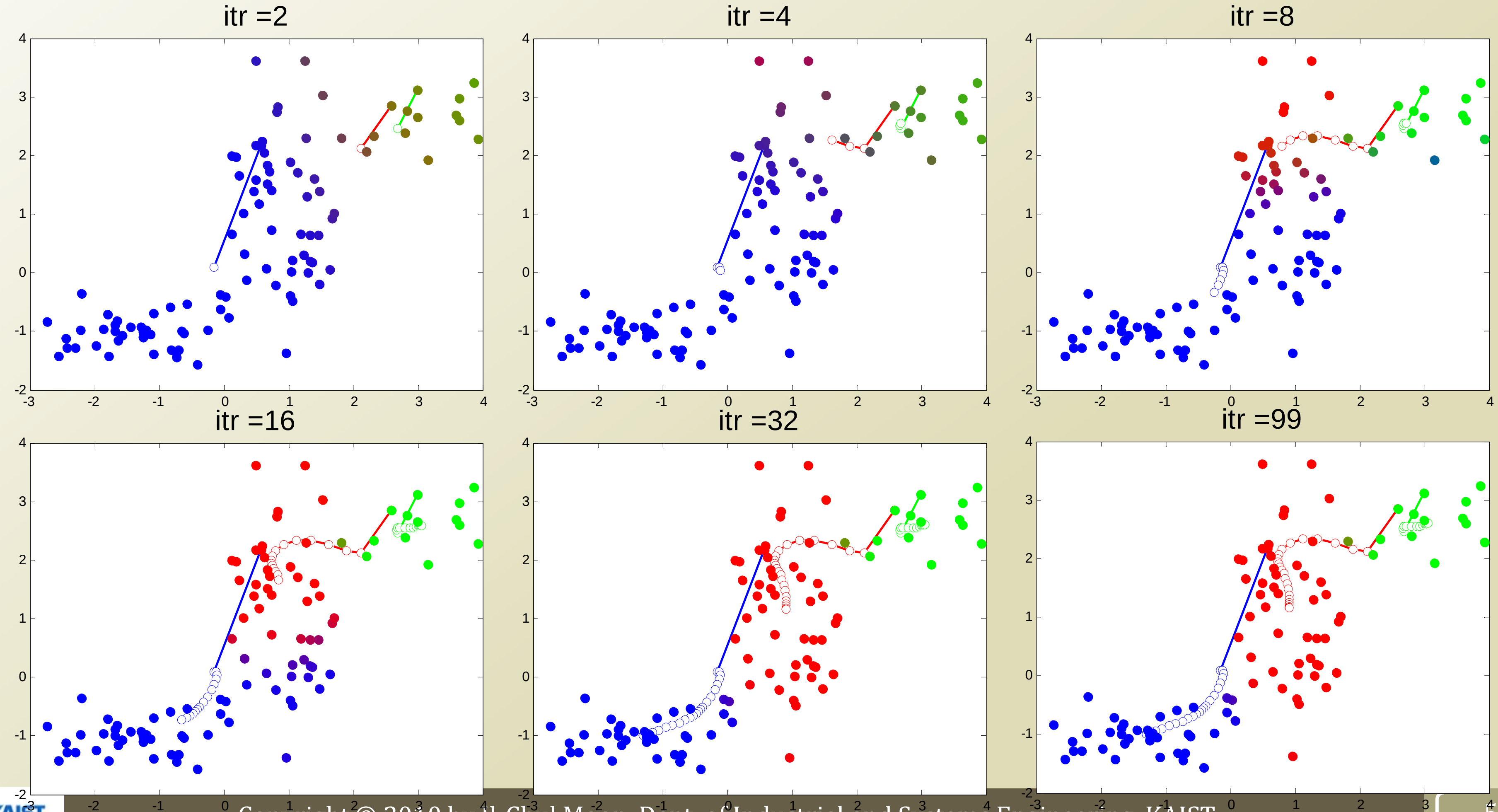
$$\rightarrow \lambda = -N \rightarrow \pi_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})}{N}$$

Progress of GMM

- Soft clustering
  - Estimated parameters
  - Soft assignment of data points to clusters







Properties of GMM

- Pros and cons of Gaussian mixture model
  - Pros
    - More information
      - Soft clustering
      - Not a simple and discrete assignment
        - Information loss
    - More and more information
      - Learn the latent distribution
      - Distance is not always the answer of the distribution
  - Cons
    - Long computation time
      - Why?
    - Falling into local maximum
    - Deciding K
- Anyways to mitigate the disadvantage?
  - Fast K-means and slow GMM