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

KAIST

EM ALGORITHM

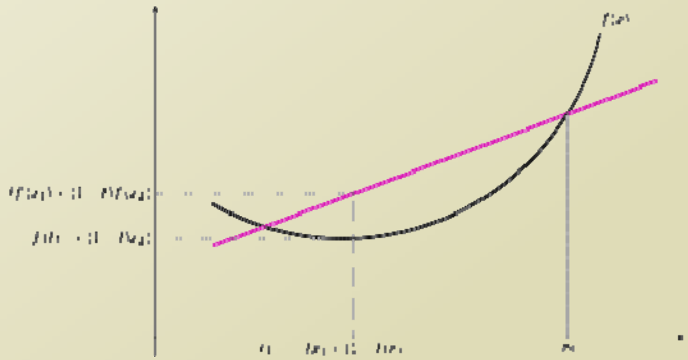
Inference with Latent Variables

- Difference between classification and clustering
- Let's say
  - {X,Z}: complete set of variables
  - X: observed variables
  - Z: hidden (latent) variables
  - θ: parameters for distributions
  - $P(X|\theta) = \sum_Z P(X,Z|\theta) \rightarrow \ln P(X|\theta) = \ln\{\sum_Z P(X,Z|\theta)\}$
    - Any problem here?
    - The locations of summation and log make this complicated
    - Eventually, we want to exchange the locations of the two operators
- What we want to know is
  - The values of Z and θ
    - Optimizing $P(X|\theta) = \sum_Z P(X,Z|\theta)$
  - The interacting terms for the optimization

Probability Decomposition

- $l(\theta) = \ln P(X|\theta) = \ln\{\sum_Z P(X,Z|\theta)\} = \ln\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\}$
  - Use the Jensen's inequality
  - $\ln\{\sum_Z q(Z)\frac{P(X,Z|\theta)}{q(Z)}\} \geq \sum_Z q(Z)\ln\frac{P(X,Z|\theta)}{q(Z)}$
- $= \sum_Z q(Z)\ln P(X,Z|\theta) - q(Z)\ln q(Z)$
  - Recall the second term?
  - $H(X) = -\sum_X P(X=x)log_b P(X=x)$
- $= E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
  - $Q(\theta,q) = E_{q(Z)}\ln P(X,Z|\theta) + H(q)$
  - This hold for any distribution of q
  - This is only the lower bound of $l(\theta)$
    - Need to make it tight!
    - How to?

Jensen's Inequality



When φ(x) is concave

$$\varphi\left(\frac{\sum a_i x_i}{\sum a_j}\right) \geq \frac{\sum a_i \varphi(x_i)}{\sum a_j}$$

When φ(x) is convex

$$\varphi\left(\frac{\sum a_i x_i}{\sum a_j}\right) \leq \frac{\sum a_i \varphi(x_i)}{\sum a_j}$$