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

GAUSSIAN MIXTURE MODEL

Multinomial Distribution

- Binary variable
  - Selecting 0 or 1 → binomial distribution
- How about K options?
  - X=(0,0,1,0,0,0) when K=6 and selecting the third option
  - $\sum_k x_k = 1, P(X|\boldsymbol{\mu}) = \prod_{k=1}^{K} {\mu_k}^{x_k}$ $such\ that\ \mu_k \geq 0, \sum_k \mu_k = 1$
  - A generalization of binomial distribution → Multinomial distribution
- Given a dataset D with N selections, $x_1 \ldots x_n$
  - $P(X|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} {\mu_k}^{x_{nk}} = \prod_{k=1}^{K} {\mu_k}^{\sum_{n=1}^{N} x_{nk}} = \prod_{k=1}^{K} {\mu_k}^{m_k}$
    - When $m_k = \sum_{n=1}^{N} x_{nk}$
    - Number of selecting k^{th} *option* out of N selections
  - How to determine the maximum likelihood solution of $\boldsymbol{\mu}$?
    - Maximize $P(X|\boldsymbol{\mu}) = \prod_{k=1}^{K} {\mu_k}^{m_k}$
    - Subject to $\mu_k \geq 0, \sum_k \mu_k = 1$

Lagrange Method

Maximize $P(X|\mu) = \prod_{k=1}^{K} \mu_k^{m_k}$
Subject to $\mu_k \geq 0, \sum_k \mu_k = 1$
When $m_k = \sum_{n=1}^{N} x_{nk}$

- Method of finding a local maximum subject to constraints
  - Maximize *f(x,y)*
  - Subject to *g(x,y)=c*
  - Assuming that *f* and *g* have continuous partial derivatives
  - 1) Lagrange function and multiplier (do you recall this?)
    - $L(x,y,\lambda) = f(x,y) + \lambda(g(x,y) - c)$
    - $L(\mu,m,\lambda) = \sum_{k=1}^{K} m_k \ln\mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1)$
      - Using the log likelihood
  - 2) Take the partial first-order derivative of variables, and set it to be zero
    - $\frac{d}{d\mu_k} L(\mu,m,\lambda) = \frac{m_k}{\mu_k} + \lambda = 0 \rightarrow \mu_k = -\frac{m_k}{\lambda}$
  - 3) Utilize the constraint to get the optimized value
    - $\sum_k \mu_k = 1 \rightarrow \sum_k -\frac{m_k}{\lambda} = 1 \rightarrow \sum_k m_k = -\lambda \rightarrow \sum_k \sum_{n=1}^{N} x_{nk} = -\lambda \rightarrow N = -\lambda$
    - $\mu_k = \frac{m_k}{N}$: MLE parameter of multinomial distribution