K-Means Clustering and Gaussian Mixture Model

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

Weekly Objectives

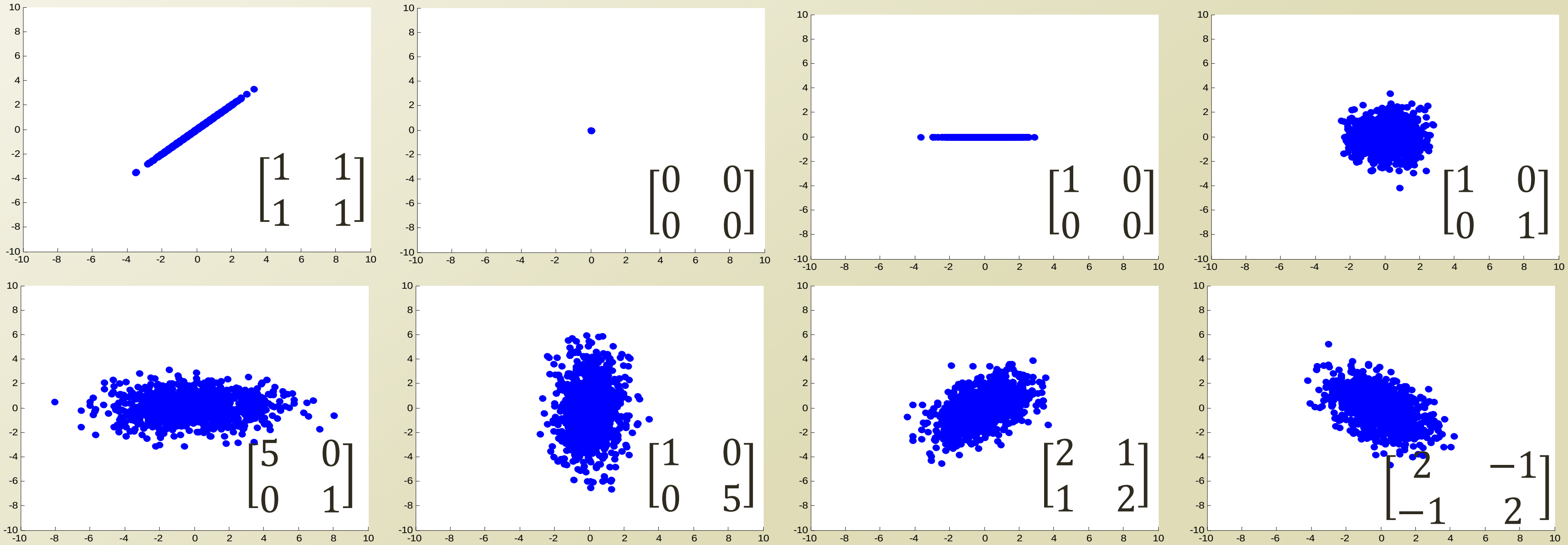
- Understand the clustering task and the K-means algorithm
  - Know what the unsupervised learning is
  - Understand the K-means iterative process
  - Know the limitation of the K-means algorithm
- Understand the Gaussian mixture model
  - Know the multinomial distribution and the multivariate Gaussian 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

Multivariate Gaussian Distribution

- Probability density function of the Gaussian distribution
  - $N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
  - $\boldsymbol{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$
    - $\mathbf{ln}\boldsymbol{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{C}$
    - $\mathbf{ln}\boldsymbol{N}(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) + \boldsymbol{C}$
      - $\propto -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} Tr[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T]$
      - $= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}Tr[\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\left((\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T\right)]$
    - $\frac{d}{d\mu}\ln N(X|\mu,\Sigma) = 0 \rightarrow -\frac{1}{2}\times 2\times -1\times \boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}}) = 0 \rightarrow \widehat{\boldsymbol{\mu}} = \frac{\sum_{n=1}^{N}\boldsymbol{x}_n}{N}$
    - $\frac{d}{d\boldsymbol{\Sigma}^{-1}}\ln N(X|\mu,\Sigma) = 0 \rightarrow \widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})(\boldsymbol{x}_n-\widehat{\boldsymbol{\mu}})^{\boldsymbol{T}}$
      - Beyond the scope of the course
      - Use “trace trick” and 1) $\frac{d}{dA}\log|A| = A^{-T}$, 2) $\frac{d}{dA}Tr[AB] = \frac{d}{dA}Tr[BA] = B^T$

Samples of Multivariate Gaussian Distribution

- Samples of multivariate Gaussian distributions
  - With various covariance matrixes
  - Covariance matrix should a positive-definite matrix
    - $z^T \Sigma z > 0$ for every non-zero column vector z
    - $[a\ b]\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix} = \begin{bmatrix}a & 0\\0 & b\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix} = a^2 + b^2 > 0$ when a,b are non-zero



Mixture Model

- Imagine that the samples are drawn from three different normal distributions
  - Subpopulation
  - The conventional distributions cannot explain the distribution accurately
  - We need to mix the three normal distribution → Create a new distribution adapted to the samples
  - Mixture distribution
- $P(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k)$
  - Mixing coefficients, π_k: A normal distribution is chosen out of K options with probability
    - Works as weighting
    - $\sum_{k=1}^{K} \pi_k = 1, 0 \le \pi_k \le 1$
    - This is a probability (as well as weighting!)
    - Then, which distribution?
    - New variable? Let's say Z!
  - Mixture component, $N(x|\mu_k, \sigma_k)$: A distribution for the subpopulation
- $P(x) = \sum_{k=1}^{K} P(z_k) P(x|z)$
  - Why this ordering of variables?