Computational Data Analysis

# Machine Learning


**Yao Xie, Ph.D.**

*Associate Professor*

Harold R. and Mary Anne Nash Early Career Professor

H. Milton Stewart School of Industrial and Systems Engineering


## Gaussian Mixture Model and EM Algorithm

# Gaussian mixture model

- A density model $p(X)$ may be multi-modal: model it as a mixture of uni-modal distributions (e.g. Gaussians)

$$\mathcal{N}(X|\mu_k, \Sigma_k) := \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}} exp\left(-\frac{1}{2}(X-\mu)^{\top}\Sigma^{-1}(X-\mu)\right)$$

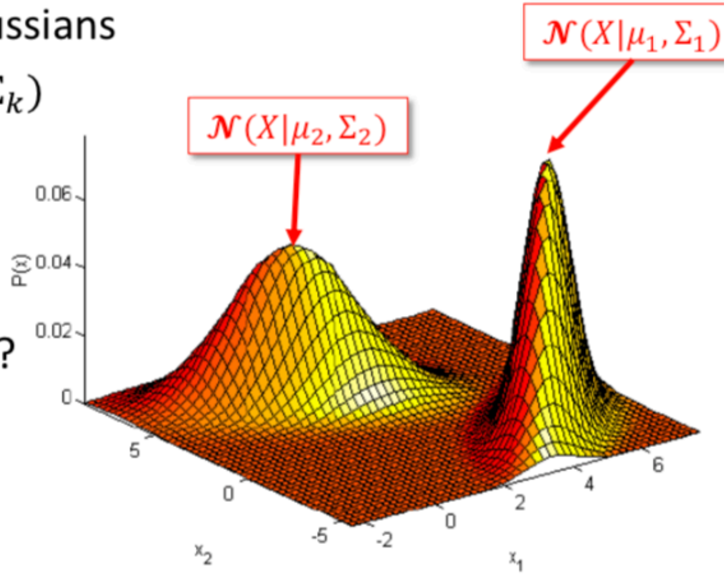
- Consider a mixture of $K$ Gaussians
  - $p(X) = \sum_{k=1}^{K} \pi_k \mathcal{N}(X|\mu_k, \Sigma_k)$

mixing proportion

mixture Component

- Parametric or noparametric?

- Learn $\pi_k \in (0,1), \mu_k, \Sigma_k$;



# EM algorithm

- Associate each data and each component with a $\tau_k^i$
- Initialize $(\pi_k, \mu_k, \Sigma_k)$, $k = 1 \ldots K$
- Iterate the following two steps till convergence:
  - Expectation step (E-step): update $\tau_k^i$ given current $(\pi_k, \mu_k, \Sigma_k)$

$$\tau_k^i = p\left(z_k^i = 1 \middle| D, \mu, \Sigma\right) = \frac{\pi_k \boldsymbol{\mathcal{N}}(x_i | \mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \boldsymbol{\mathcal{N}}(x_i | \mu_{k'}, \Sigma_{k'})}$$

$$(k = 1 \ldots K, i = 1 \ldots m)$$

  - Maximization step (M-step): update $(\pi_k, \mu_k, \Sigma_k)$ given $\tau_k^i$

$$\pi_k = \frac{\sum_i \tau_k^i}{m}, \qquad \mu_k = \frac{\sum_i \tau_k^i \, x^i}{\sum_i \tau_k^i}$$

$$\Sigma_k = \frac{\sum_i \tau_k^i \left(x^i - \mu_k\right)\left(x^i - \mu_k\right)^T}{\sum_i \tau_k^i}$$

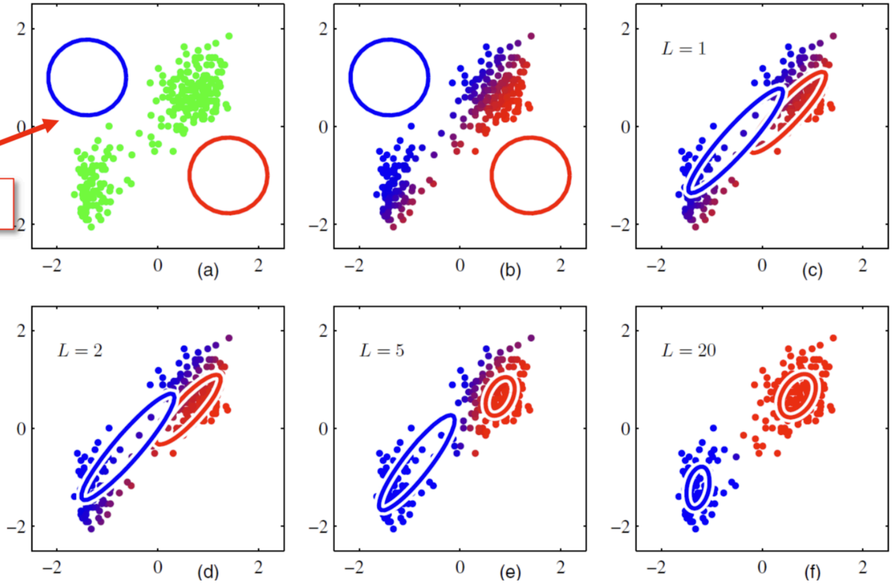
$$(k = 1 \ldots K)$$

Georgia Tech

# Expectation-Maximization Iterations

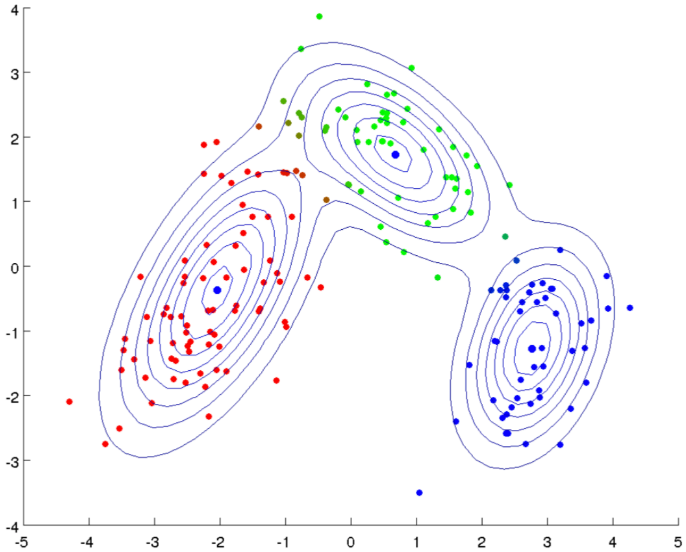
- $k = 1\ or\ 2$
- Use $\tau_1^i$ as the proportion of red, and $\tau_2^i$ proportion of blue
- Draw only one contour for each Gaussian component

$Intialization$



# Mixture of 3 Gaussians

- First run PCA to reduce the dimension to 2
- $k = 1\ or\ 2\ or\ 3$
- Use $\tau_1^i$ as the proportion of red, $\tau_2^i$ proportion of green, and $\tau_3^i$ proportion of green




Georgia Tech

# Image a generative process for data points

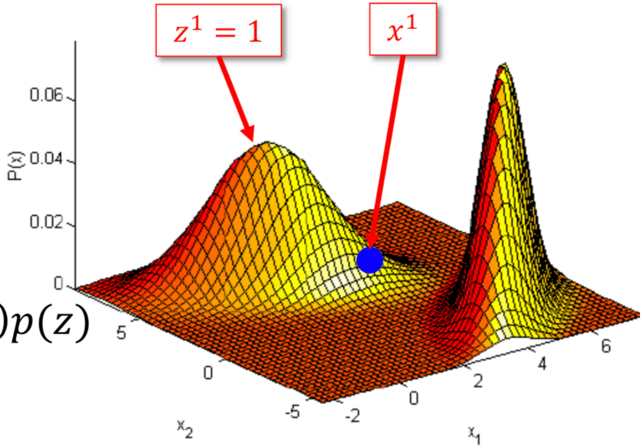
- For each data point $x^i$:
  - Randomly choose a mixture component, $z^i = \{1,2, \ldots K\}$, with probability $\pi_{z^i}$
  - Then sample the actual value of $x^i$ from a Gaussian distribution $\mathcal{N}\left(x \middle| \mu_{z^i}, \Sigma_{z^i}\right)$

- Joint distribution over $p(x, z)$

$$p(x, z) = \pi_z \mathcal{N}(x|\mu_z, \Sigma_z)$$

- Marginal distribution $p(x)$

$$p(x) = \sum_{z=1}^{K} p(x, z) = \sum_{z=1}^{K} p(x|z)p(z)$$




Georgia Tech

# Learning the Parameters

- How to learn?

- Maximum likelihood learning (let $\theta = (\pi_k, \mu_k, \Sigma_k)$, $k = 1 \dots K$

$$\theta^* = \operatorname{argmax} l(\theta; D) = \log \prod_{i=1}^{m} p(x^i)$$

- Use our generative process

$$l(\theta; D) = \log \prod_{i=1}^{m} \left( \sum_{z^i=1}^{K} p(x^i, z^i | \theta) \right)$$

$$= \log \prod_{i=1}^{m} \left( \sum_{z^i=1}^{K} p(x^i | \mu_{z^i}, \Sigma_{z^i}) p(z^i | \pi) \right)$$

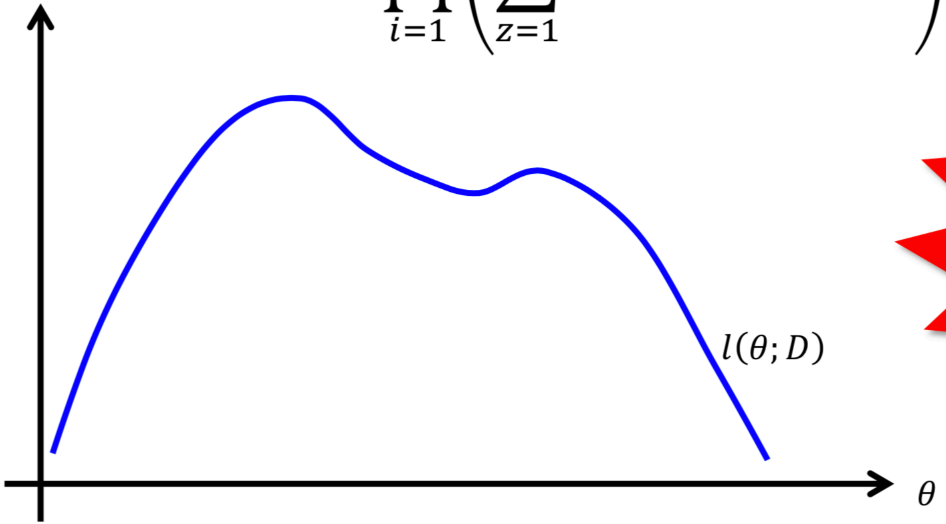
Georgia Tech

# Why is learning hard?

- With latent variables $z$, likelihood of the data becomes

$$l(\theta; D) = \log \prod_{i=1}^{m} \left( \sum_{z^i=1}^{K} p(x^i|\mu_{z^i}, \Sigma_{z^i}) p(z^i|\pi) \right)$$

$$= \log \prod_{i=1}^{m} \left( \sum_{z=1}^{K} \pi_{z^i} \mathcal{N}(x|\mu_{z^i}, \Sigma_{z^i}) \right)$$




Nonconvex
Difficult!

# Details of EM

- We intend to learn the parameters that maximizes the log–likelihood of the data

$$l(\theta; D) = \log \prod_{i=1}^{m} \left( \sum_{z^i=1}^{K} p(x^i, z^i | \theta) \right)$$

Nonconvex
Difficult!

- Expectation step (E-step): What do we take expectation over?

$$l(\theta; D) \geq f(\theta) = E_{q(z^1, z^2, \ldots, z^m)}[\log \prod_{i=1}^{m} p(x^i, z^i | \theta)]$$

- Maximization step (M-step): how to maximize?

$$\theta^{t+1} = argmax_{\theta} \ f(\theta)$$

# Bayes rule

likelihood

Prior

$$P(z|x) = \frac{P(x|z)P(z)}{P(x)} = \frac{P(x,z)}{\sum_{z'} P(x,z')}$$

posterior

normalization constant

Prior: $p(z) = \pi_z$

Likelihood: $p(x|z) = \boldsymbol{\mathcal{N}}(x|\mu_z, \Sigma_z)$

Posterior: $p(z|x) = \frac{\pi_z \boldsymbol{\mathcal{N}}(x|\mu_z,\Sigma_z)}{\sum_{z'} \pi_{z'} \boldsymbol{\mathcal{N}}(x|\mu_{z'},\Sigma_{z'})}$

# E-step: what is $q(z^1, z^2, .., z^m)$

- $q(z^1, z^2, .., z^m)$: posterior distribution of the latent variables

$$q(z^1, z^2, .., z^m) = \prod_{i=1}^{m} p(z^i | x^i, \theta^t)$$

- For each data point $x^i$, compute $p(z^i = k | x^i)$ for each $k$

$$\tau_k^i = p\left(z^i = k \middle| x^i\right) = \frac{p(z^i = k\ , x^i)}{\sum_{k'=1..K} p(z^i = k', x^i)}$$

$$= \frac{\pi_k \mathcal{N}\left(x^i \middle| \mu_k, \Sigma_k\right)}{\sum_{k'=1..K} \pi_{k'} \mathcal{N}(x^i | \mu_{k'}, \Sigma_{k'})}$$

Georgia Tech

# E-step: compute the expectation

$$f(\theta) = E_{q(z^1, z^2, \dots, z^m)}\left[\log \prod_{i=1}^{m} p(x^i, z^i|\theta)\right]$$

$$= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}\left[\log p(x^i, z^i|\theta)\right]$$

$$= \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}\left[\log \pi_{z^i} \mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})\right]$$

- Expand log of Gaussian $\log \mathcal{N}(x^i|\mu_{z^i}, \Sigma_{z^i})$

$$f(\theta) = \sum_{i=1}^{m} E_{p(z^i|x^i,\theta^t)}\left[\log \pi_{z^i} - (x^i - \mu_{z^i})^\top \Sigma_{z^i}^{-1} (x^i - \mu_{z^i}) - \log|\Sigma_{z^i}| + c\right]$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{K} \tau_k^i \left[\log \pi_k - (x^i - \mu_k)^\top \Sigma_k^{-1} (x^i - \mu_k) - \log|\Sigma_k| + c\right]$$

Georgia Tech

# M-step: maximize $f(\theta)$

- $f(\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} \ \tau_i^k \left[\log \pi_k - (x^i - \mu_k)^\top \Sigma_k^{-1} (x^i - \mu_k) \text{ - } \log|\Sigma_k| + c\right]$

- For instance, we want to find $\pi_k$, and $\sum_{i=1}^{K} \pi_k = 1$
  - Form Lagrangian

$$L = \sum_{i=1}^{m} \sum_{k=1}^{K} \ \tau_k^i [\log \pi_k + other\ terms] \ + \lambda(1 - \sum_{i=1}^{K} \pi_k)$$

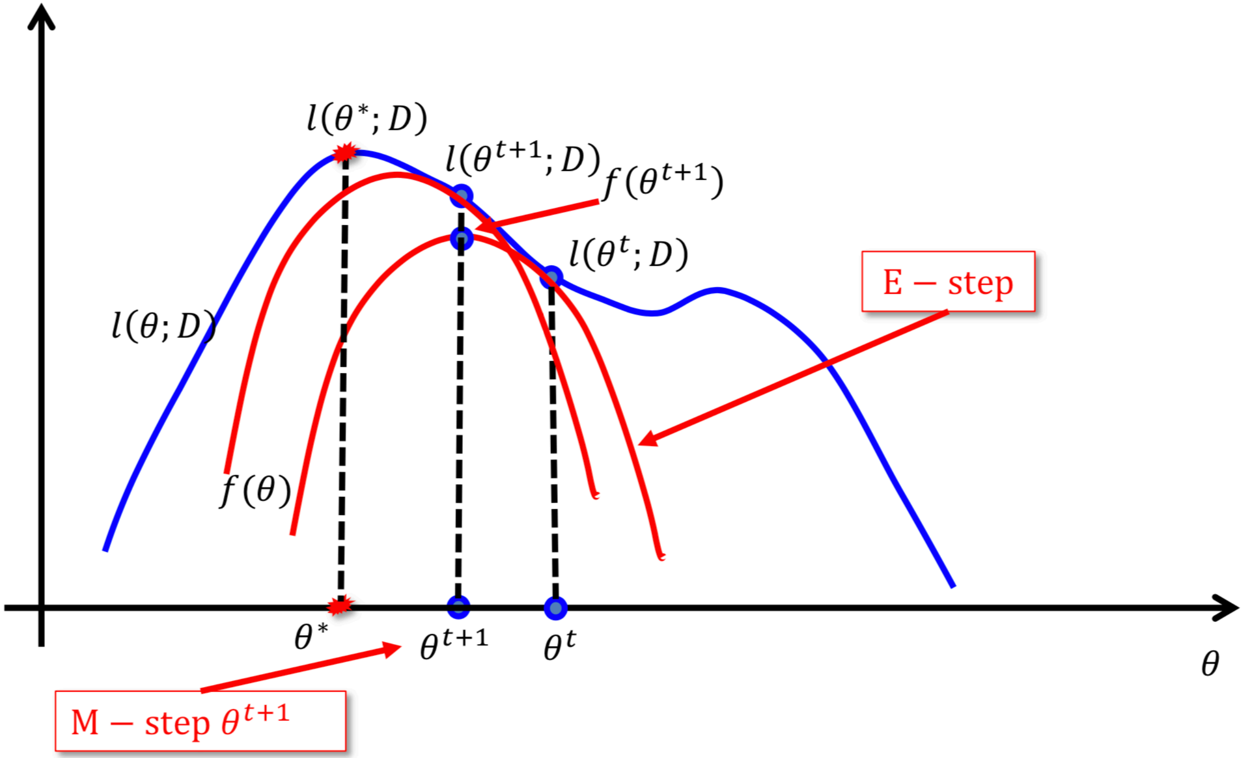
  - Take partial derivative and set to 0

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{m} \frac{\tau_k^i}{\pi_k} - \lambda = 0$$

$$\Rightarrow \pi_k = \frac{1}{\lambda} \sum_{i=1}^{m} \tau_k^i$$

$$\Rightarrow \lambda = m$$

# EM graphically



# EM vs. modified K-means

- The EM algorithm for mixture of Gaussian is like a soft clustering algorithm

- K-means:
  - "E-step", we do hard assignment:
    - $z^i = argmax_k (x^i - \mu_k)\, \Sigma_k^{-1} (x^i - \mu_k)$
  - "M-step", we update the means and covariance of cluster using maximum likelihood estimate:
    - $\mu_k = \dfrac{\sum_i \delta(z^i, k) x^i}{\sum_i \delta(z^i, k)}$
    - $\Sigma_k = \dfrac{\sum_i \delta(z^i, k)(x^i - \mu_k)\,(x^i - \mu_k)^T}{\sum_i \delta(z^i, k)}$

$\delta(z^i, k) = 1 \; if \; z^i = k$; otherwise 0.