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

And as you can see, usually the Gaussian is able to capture one mode, one hump in your data distribution. And to make this model more flexible, people considered having a mixture, instead of having one Gaussian model, now you have multiple Gaussians to model multiple modes, and multiple humps in your data distribution.

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

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

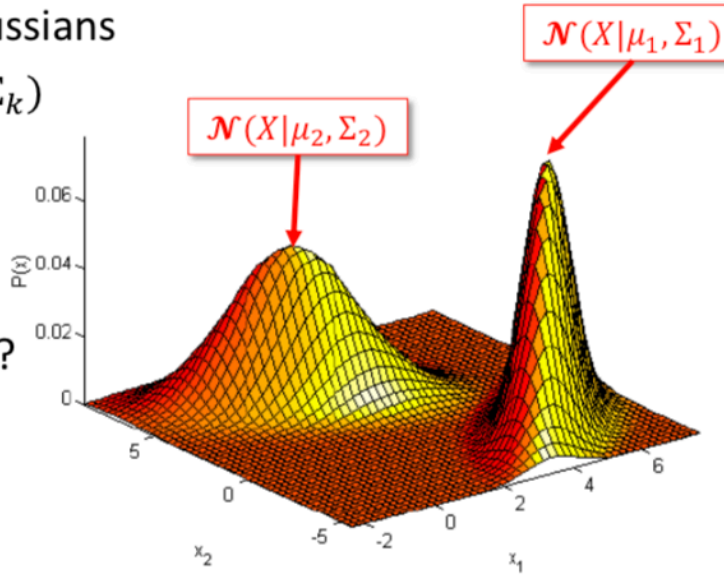
Each Gaussian component will have a different mean and different covariance matrix.

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



Georgia Tech

EM algorithm

- Associate each data and each component with a τ_k^i
- Initialize (π_k, μ_k, Σ_k), $k = 1 \dots K$
- Iterate the following two steps till convergence:
  - Expectation step (E-step): update τ_k^i given current (π_k, μ_k, Σ_k)

So, first we'll introduce a variable called tau ik. We're gonna associate each data point, i, and each Gaussian component, k, this variable tau ik.And this variable is going to indicate the probability that the i sample comes from the case component.

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

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

So you can see this is a probability of the case component plugging the i sample, the likelihood, and then you normalize it by summing this over all the case components. So after we have obtained this E-step, we're gonna use these pi, tau ik's to update the parameters, the mixing proportion.

  - Maximization step (M-step): update (π_k, μ_k, Σ_k) given τ_k^i

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

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

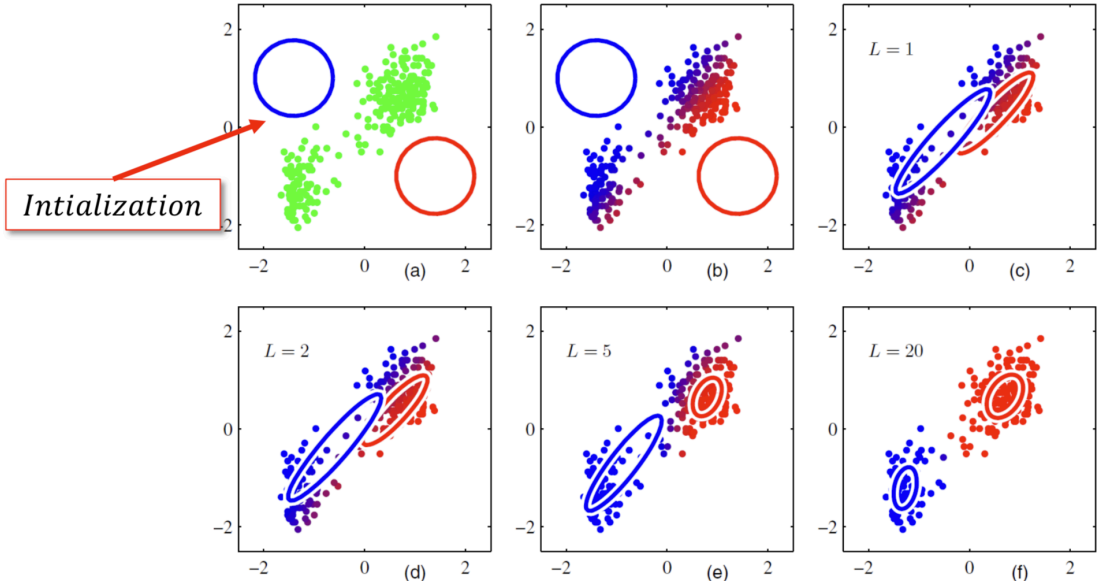
$$(k = 1 \dots K)$$

As you can see, this is the sum of the tau ik over all the samples divided by m, and then updating the mu, and updating the sigma correspondingly.

Georgia Tech

Expectation-Maximization Iterations

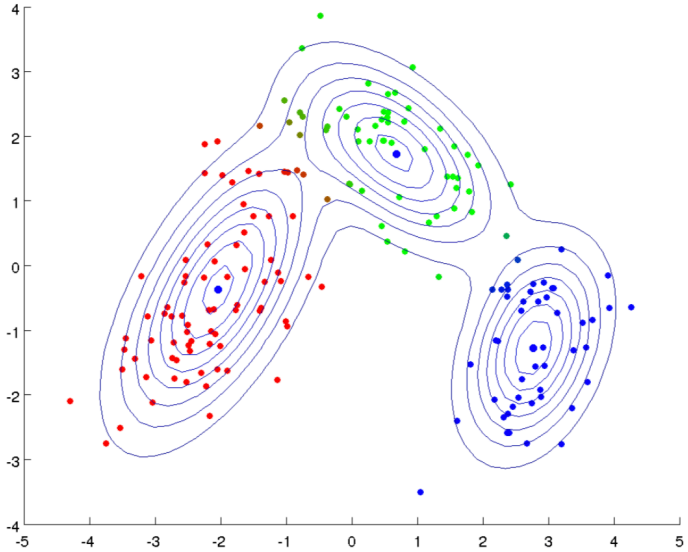
- $k = 1\ or\ 2$
- Use τ_1^i as the proportion of red, and τ_2^i proportion of blue
- Draw only one contour for each Gaussian component



Georgia Tech

Mixture of 3 Gaussians

- First run PCA to reduce the dimension to 2
- $k = 1\ or\ 2\ or\ 3$
- Use τ_1^i as the proportion of red, τ_2^i proportion of green, and τ_3^i proportion of green



Georgia Tech

Image a generative process for data points

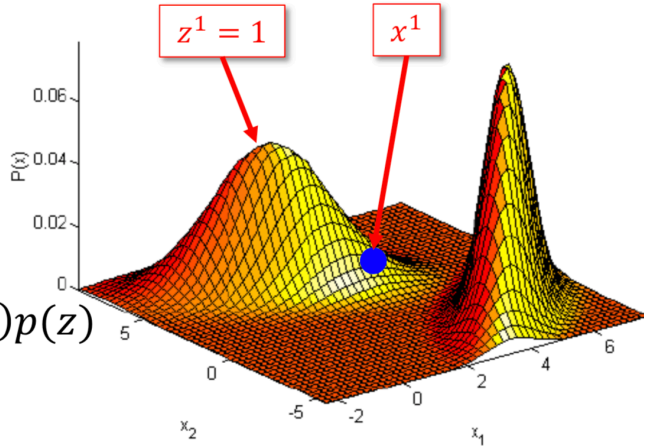
- For each data point x^i:
  - Randomly choose a mixture component, $z^i = \{1,2, \dots K\}$, with probability π_{z^i}
  - Then sample the actual value of x^i from a Gaussian distribution $\mathcal{N}(x|\mu_{z^i}, \Sigma_{z^i})$

- 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 $\theta = (\pi_k, \mu_k, \Sigma_k), k = 1 \dots K$

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

- Use our generative process

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

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

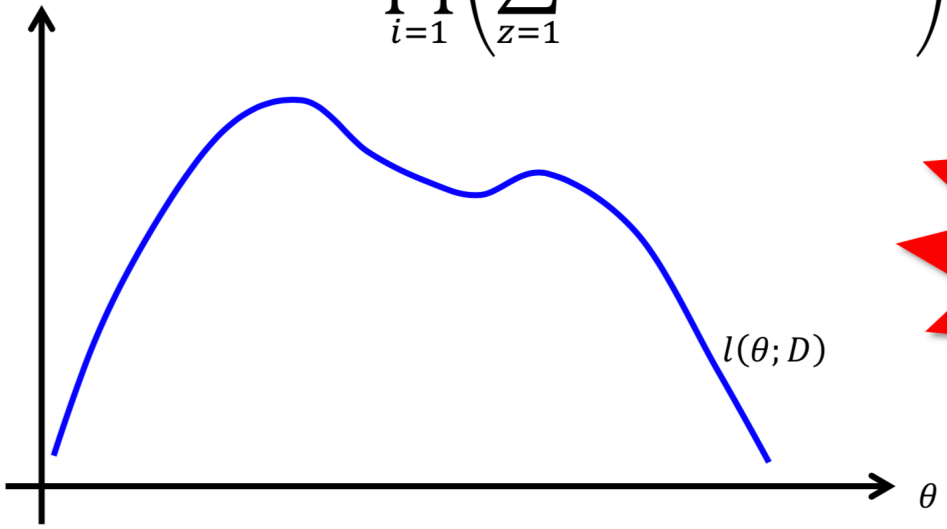
Georgia Tech

Why is learning hard?

- With latent variables z, likelihood of the data becomes

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

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



Nonconvex
Difficult!

Georgia Tech

Details of EM

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

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

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

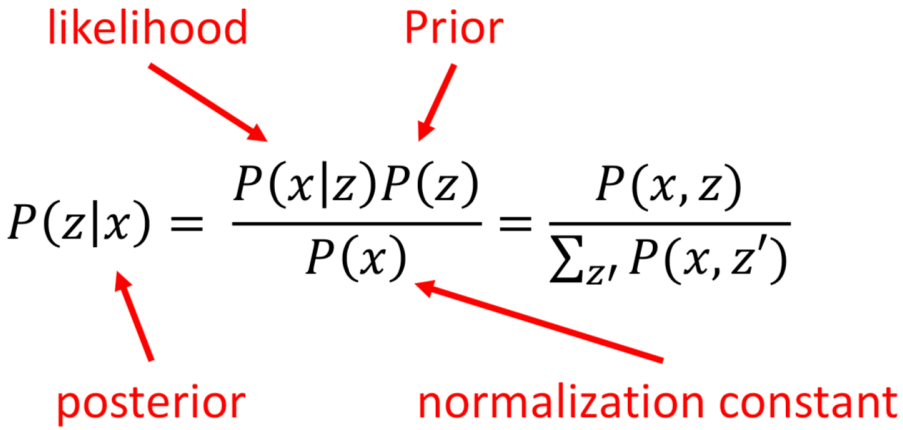
$$l(\theta; D) \geq f(\theta) = E_{q(z^1, z^2, .., z^m)}[\log \prod_{i=1}^{m} p(x^i, z^i | \theta)]$$

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

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

Georgia Tech

Bayes rule



Prior: $p(z) = \pi_z$

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

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

Georgia Tech

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

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

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

Georgia Tech

E-step: compute the expectation

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

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

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

- 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 $f(\theta)$

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

- For instance, we want to find π_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

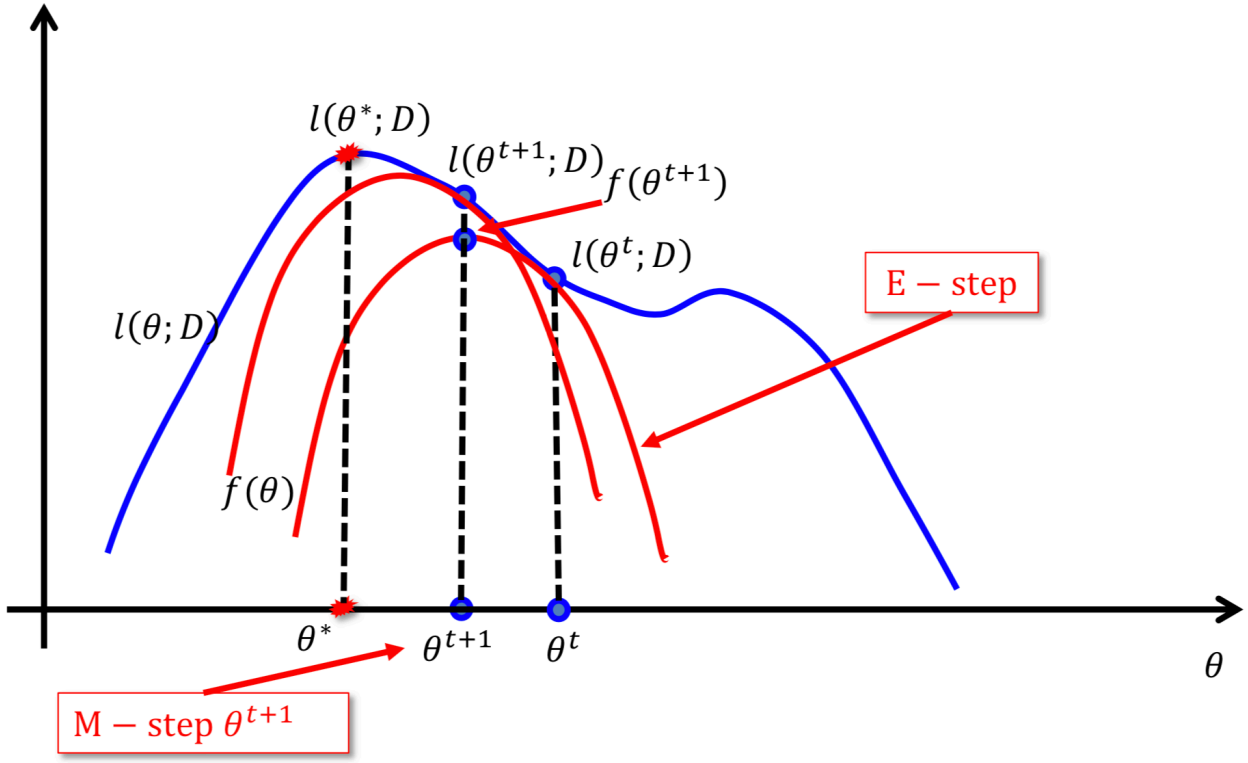
$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{m} \frac{\tau_k^i}{\pi_k} - \lambda = 0$$

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

$$\Rightarrow \lambda = m$$

Georgia Tech

EM graphically



Georgia Tech

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:
    - $\mu_k = \frac{\sum_i \delta(z^i,k) x^i}{\sum_i \delta(z^i,k)}$
    - $\Sigma_k = \frac{\sum_i \delta(z^i,k)(x^i-\mu_k)\,(x^i-\mu_k)^T}{\sum_i \delta(z^i,k)}$

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