Gaussian Mixture Models

Tuan Nguyen

Ngày 8 tháng 2 năm 2023

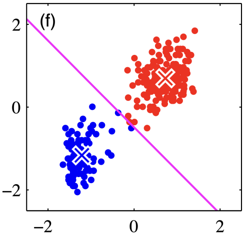
Overview

Soft assignment

Normal distribution

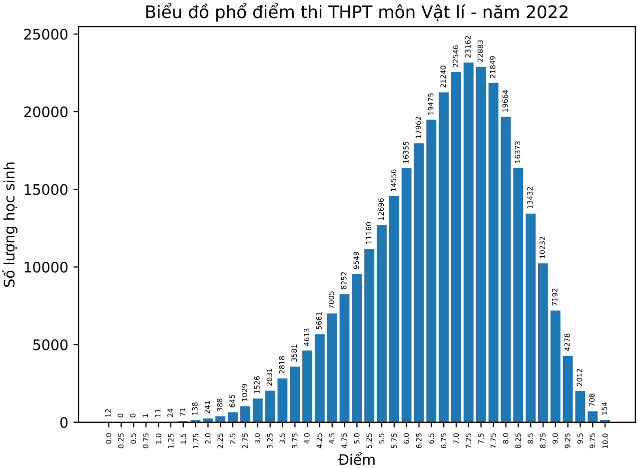
Generative model

Hard assignment vs Soft assignment

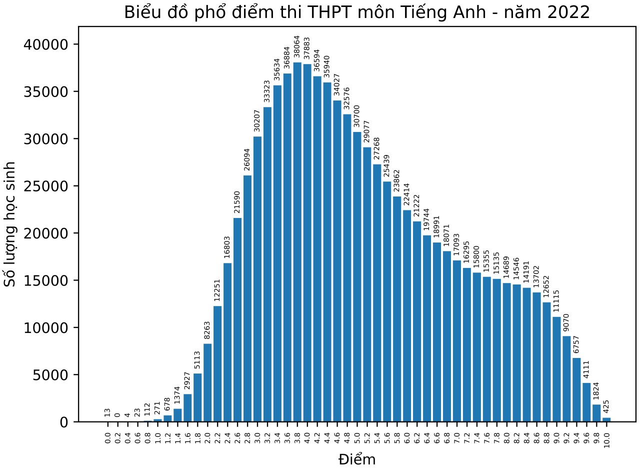


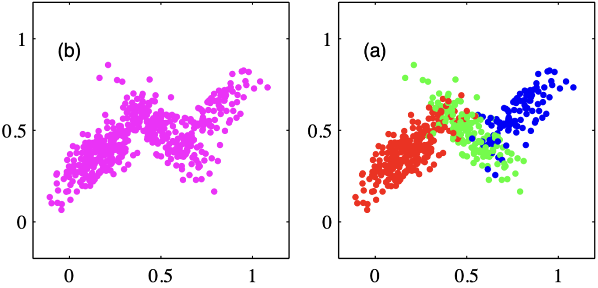
- In the K-means algorithm, a hard assignment of points to clusters is made.
- However, for points near the decision boundary, this may not be such a good idea.
- Instead, we could think about making a soft assignment of points to clusters.

Normal distribution



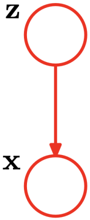
Mix Normal distribution

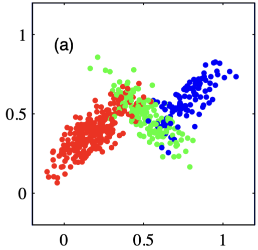




- The Gaussian mixture model (or mixture of Gaussians MoG) models the data as a combination of Gaussians
- Above shows a dataset generated by drawing samples from three different Gaussians

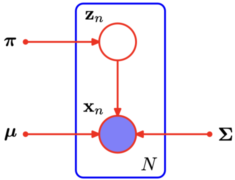
Generative model

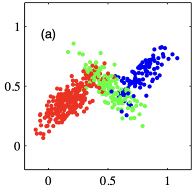




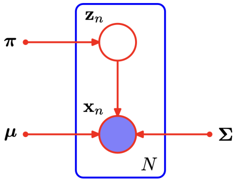
- The mixture of Gaussians is a generative model
- To generate a datapoint xn, we first generate a value for a discrete variable $z_n \in \{1, ..., K\}$
- We then generate a value $x_n \sim \mathrm{N}(x|\mu_k, \Sigma_k)$ for the corresponding Gaussian

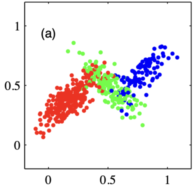
Graphical model





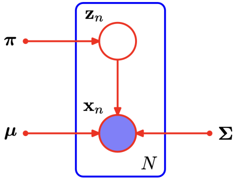
- z_n is a latent variable, unobserved
- Need to give conditional distributions $p(z_n)$ and $p(x_n|z_n)$
- The one-of-K representation is helpful here: $z_{nk} \in \{0, 1\}$, $z_n = (z_{n1}, ..., z_{nK})$

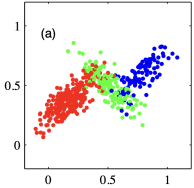




- Use a Bernoulli distribution for $p(z_n)$
  - $\pi_k = p(z_{nk} = 1)$
  - Parameters to this distribution π_k
  - Must have $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$
- $p(z_n) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$

Graphical Model - Observed Variable

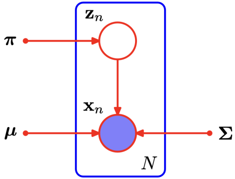


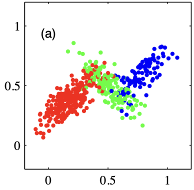


- Use a Gaussian distribution for $p(x_n|z_n)$

$$p(x_n|z_{nk}=1) = \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

$$p(x_n|z_n) = \prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}$$





- The full joint distribution is given by

$$
\begin{aligned}
p(x, z) &= \prod_{n=1}^{N} p(z_n) p(x_n | \mu_k, \Sigma_k) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}}
\end{aligned}
$$

MoG Marginal over Observed Variables

- The marginal distribution $p(x_n)$ for this model is

$$p(x_n) = \sum_{z_n} p(x_n, z_n) = \sum_{z_n} p(z_n)p(x_n|z_n)$$
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

- A mixture of Gaussians $\Rightarrow$ model parameters are $\{\pi_k, \mu_k, \Sigma_k\}$
- Similar to k-means
  - If we know the latent variables z_n, fitting the Gaussians is easy
  - If we know the Gaussian μ_k, Σ_k, finding the latent variables is easy

MoG Marginal over Observed Variables (cont.)

- Rather than latent variables, we will use responsibilities $p(z_{nk} = 1|x_n)$

$$\gamma(z_{nk}) = p(z_{nk} = 1|x_n) \tag{1}$$

$$= \frac{p(z_{nk}1)p(x_n|z_{nk} = 1)}{\sum_{j=1}^{K} p(z_{nj} = 1)p(x_n|z_{nj} = 1)} \tag{2}$$

$$= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \tag{3}$$

- $\gamma(z_{nk})$ is the responsibility of component k for datapoint n

MoG Maximum Likelihood Learning

- Model parameters are $\theta = \{\pi_k, \mu_k, \Sigma_k\}$
- We can use the maximum likelihood criterion

$$\begin{aligned}
\theta_{ML} &= \arg\max_{\theta} \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \\
&= \arg\max_{\theta} \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right)
\end{aligned}$$

- Unfortunately, closed-form solution not possible this time log of sum rather than log of product

ML for Gaussian Mixtures

- Consider the log-likelihood function

$$l(\theta) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)$$

- We can try taking derivatives and setting to zero, even though no closed form solution exists

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

- All depend on $\gamma(z_{nk})$, which depends on all 3 $\Rightarrow$ iterative scheme can be used

EM for Gaussian Mixtures

- Initialize parameters, then iterate:
  - E-step: Calculate responsibilities using current parameters

$$\gamma(z_{nk}) = \frac{p(z_{nk}1)p(x_n|z_{nk}=1)}{\sum_{j=1}^{K} p(z_{nj}=1)p(x_n|z_{nj}=1)}$$

  - M-step: Re-estimate parameters using these $\gamma(z_{nk})$

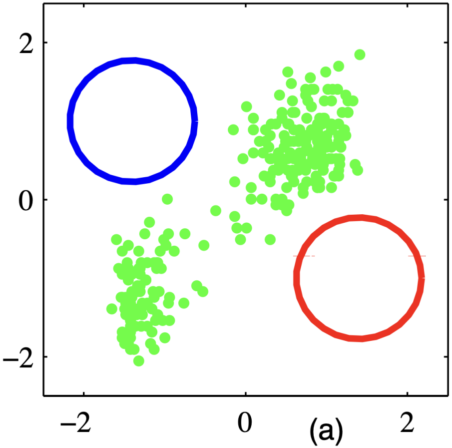
$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

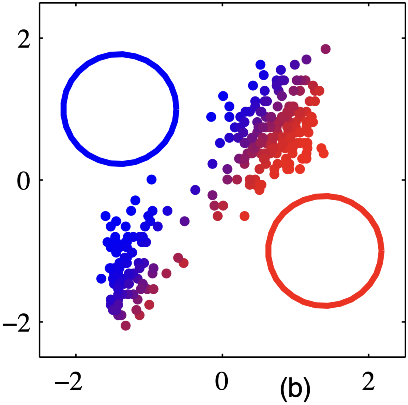
- This algorithm is known as the expectation-maximization algorithm (EM)

Example

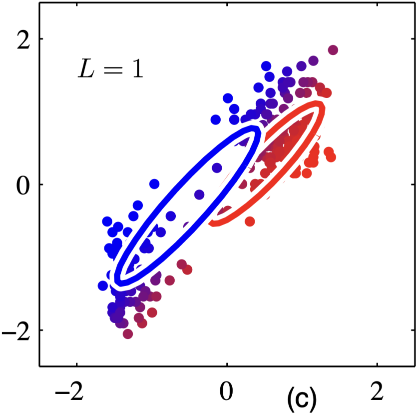


Same initialization as with K-means before

Example (cont.)

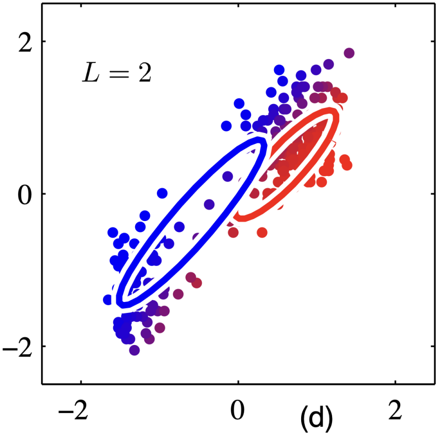


Calculate responsibilities $\gamma(z_{nk})$



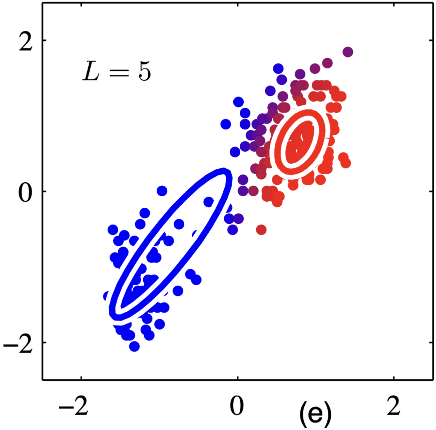
Calculate model parameters $\{\pi_k, \mu_k, \Sigma_k\}$ using these responsibilities

Example (cont.)



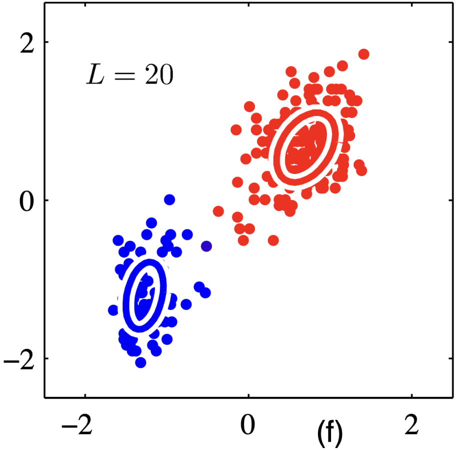
Iteration 2

Example (cont.)



Iteration 5

Example (cont.)



Iteration 20 - converged