Machine Learning for Data Science (CS4786)
Lecture 13

Mixture Models

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016fa/

Probabilistic Models

- Θ consists of set of possible parameters
- We have a distribution P_θ over the data induced by each $\theta \in \Theta$
- Data is generated by one of the $\theta \in \Theta$
- Learning: Estimate value or distribution for $\theta^* \in \Theta$ given data

Maximum Likelihood Principal

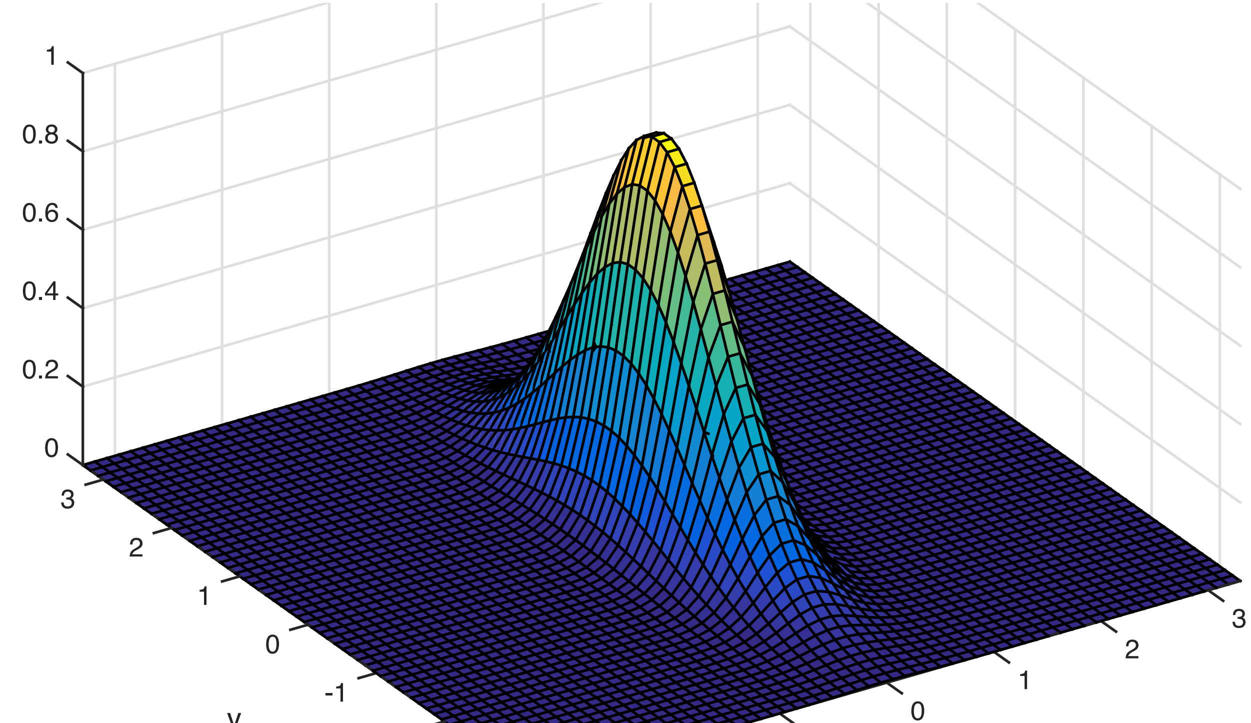
Pick $\theta \in \Theta$ that maximizes probability of observation

$$\theta_{MLE} = \text{argmax}_{\theta \in \Theta} \log \underbrace{P_\theta(x_1, \ldots, x_n)}_{\text{Likelihood}}$$

Multivariate Gaussian

- Two parameters:
  - Mean $\mu \in \mathbb{R}^d$
  - Covariance matrix Σ of size dxd

$$p(x; \mu, \Sigma) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma (x-\mu)\right)$$

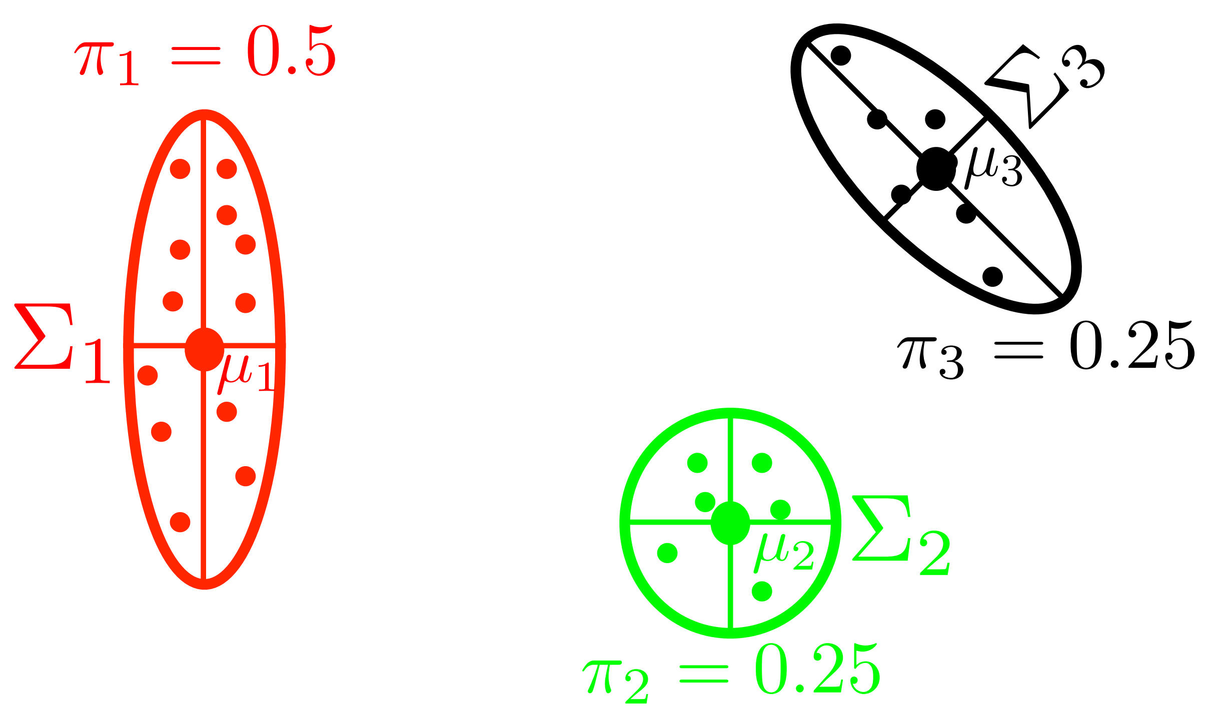


Gaussian Mixture Models

Each $\theta \in \Theta$ is a model.

- Gaussian Mixture Model
  - Each θ consists of mixture distribution $\pi = (\pi_1, \ldots, \pi_K)$, means $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K$
  - For each t, independently:

$$c_t \sim \pi, \qquad x_t \sim N(\mu_{c_t}, \Sigma_{c_t})$$



Towards EM Algorithm

- Latent variables can help, but we have a chicken and egg problem

Given all variables including latent variables, finding optimal parameters is easy

Given model parameter, optimizing/finding distribution over the latent variables is easy

EM ALGORITHM FOR GMM

1. Initialize model parameters $\pi^{(0)}, \mu_1^{(0)}, \ldots, \mu_K^{(0)}$ and $\Sigma_1^{(0)}, \ldots, \Sigma_K^{(0)}$
2. For i = 1 until convergence or bored
   1. $Q_t^{(i)}(k) \propto p(\mathbf{x}_t; \mu_k^{(i-1)}, \Sigma_k^{(i-1)}) \cdot \pi_k^{(i-1)}$
   2. For every $k \in [K]$,

$$\mu_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t(k)}, \quad \Sigma_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k)\left(x_t - \mu_k^{(i)}\right)\left(x_t - \mu_k^{(i)}\right)^\top}{\sum_{t=1}^n Q_t(k)}$$

(weighted centroid) (weighted covariance)

$$\pi_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

3. End For

Demo

Why should EM work?

A very high level view:

- Performing E-step will never decrease log-likelihood (or log a posteriori)
- Performing M-step will never decrease log-likelihood (or log a posteriori)

WHY SHOULD EM WORK?

- Likelihood never decreases
- So whenever we converge we converge to a local optima
- However problem is non-convex and can have many local optimal
- In general no guarantee on rate of convergence
- In practice, do multiple random initializations and pick the best one!

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n)$$

Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)$$

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\begin{aligned}
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) &= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right)
\end{aligned}$$

Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\begin{aligned}
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) &= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} Q^{(i)}(c_t) \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right)
\end{aligned}$$

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\begin{aligned}
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) &= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right) \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} Q^{(i)}(c_t) \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right) \\
&\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
\end{aligned}$$

Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log\left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)}\right)$$

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right)$$

Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right)$$

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t)$$

Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left(\frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t | x_t)} \right)$$

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t)$$

$$= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)$$

Mixture of Multinomials

- Eg. Model purchases of each customer $(x_t = \text{ one of the d items bought})$
- K-types of customers, each designated with distribution over the d items to buy
- Generative model:
  - π is mixture distribution over the K-types of buyers
  - $p_1, \ldots, p_K$ are the K distributions over the d items, one for each customer type
  - Generative process, each round draw customer type $c_t \sim \pi$
  - Next given c_t draw list of purchases as $x_t \sim \text{multinomial}(p_{c_t})$

EM Algorithm for Mixture of Multinomials

1. Initialize model parameters $\pi^{(0)}$ and $p_1^{(0)}, \ldots, p_K^{(0)}$.
2. For i = 1 until convergence or bored
    1. $Q_t^{(i)}(k) \propto p_k^{(i-1)}[x_t] \cdot \pi_k^{(i-1)}$
    2. For every $k \in [K]$,

$$p_k^{(i)}[j] = \frac{\sum_{t=1}^n Q_t^{(i)}(k) \mathbf{1}\{x_t = j\}}{\sum_{t=1}^n Q_t(k)}, \quad \pi_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

3. End For

Mixture Models

- π is mixture distribution over the K-types
- $\gamma_1, \ldots, \gamma_K$ are parameters for K distributions
- Generative process:
  - Draw type $c_t \sim \pi$
  - Next given c_t, draw $x_t \sim \text{Distribtuion}(\gamma_{c_t})$

EM Algorithm for Mixture Models

For $i = 1$ to convergence

(E step) For every t, define distribution Q_t over the latent variable c_t as:

$$Q_t^{(i)}(c_t) \propto \mathrm{PDF}(x_t; \gamma_{c_t}^{(i-1)}) \cdot \pi^{(i-1)}[c_t]$$

(M step) For every $k \in \{1, \ldots, K\}$

$$\pi_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}[k]}{n}, \quad \gamma_k^{(i)} = \underset{\gamma}{\mathrm{argmin}} \sum_{t=1}^n Q_t[k] \log(\mathrm{PDF}(x_t; \gamma))$$

- x_t observation, c_t latent variable.