Lecture 22
Expectation Maximization (EM)
ECEN 5283 Computer Vision

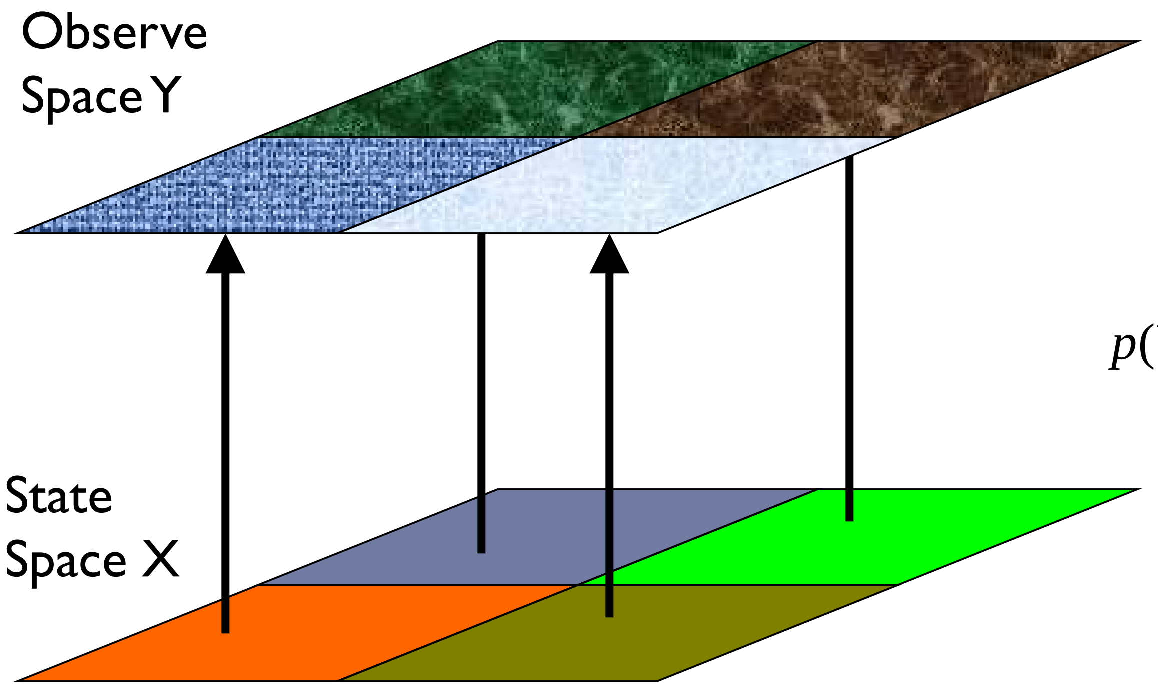
Dr. Guoliang Fan
School of Electrical and Computer Engineering
Oklahoma State University

Goals

- To review the missing data problem and its two major issues.

- To introduce a soft-clustering algorithm, i.e., Expectation Maximization (EM) algorithm.

Issue (1) Parameter Estimation



Parameter Space

$$\Theta = \{\alpha_i, \boldsymbol{\mu}_i, \Sigma_i \mid i = 1,2,3,4\}$$

$$p(Y \mid \Theta) = \prod_{y \in Y} p(y) : \text{data likelihood}$$

(the model - data fitness)

$$\Theta^* = \arg\max_{\Theta} p(Y \mid \Theta) \quad \text{(parameter estimation)}$$

To find the parameter Θ that can best explain the current observation Y.

Parameter Estimation Example

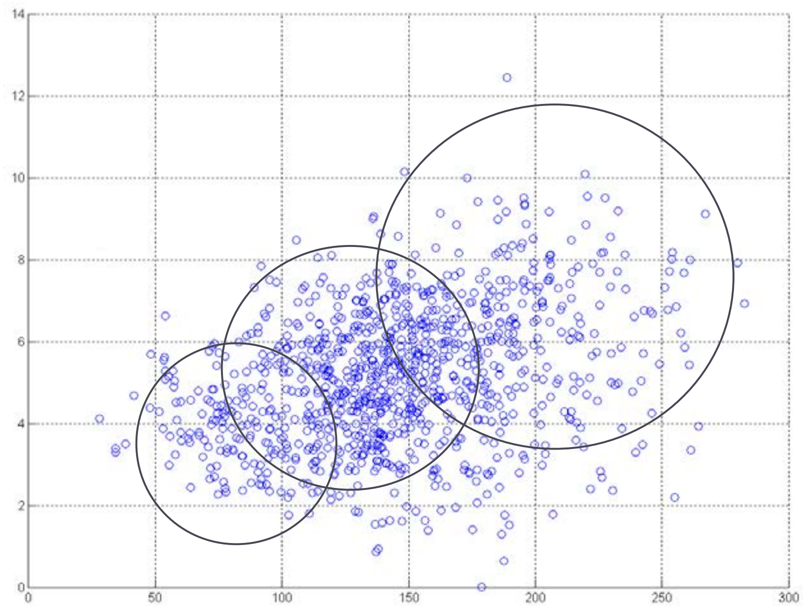
Labels and Data

$$(X,Y) = \left((x_l, y_l) \mid l = 1,\ldots,N\right)$$

x_l : label, y_l : observation

Parameter Space

$$\Theta = \{\alpha_i, \boldsymbol{\mu}_i, \Sigma_i \mid i = 1,2,3\}$$



$$\Theta^* = \arg\max_{\Theta} p(Y \mid \Theta)$$

(maximum likelihood estimation)

$$\alpha_i = P(x = i)$$

(prior probability)

$$p(y \mid x = i) = \mathrm{N}(y \mid \boldsymbol{\mu}_i, \Sigma_i)$$

(likelihood function)

Missing Data Problem: Issue (2) Data Classification

- After parameter estimation, we need to decide the class label of each data sample according to estimated parameters.

$$Y=(y_1,y_2,\ldots,y_N) \xrightarrow{\Theta=\{\alpha_i,\boldsymbol{\mu}_i,\Sigma_i \mid i=1,\ldots,k\}} X=(x_1,x_2,\ldots,x_N)$$

- We need to compute the posterior probability of data sample y belonging to class x. (This is the estimate of the missing data)

$$\alpha_i=\Pr(x=i)=p(x)$$

(prior probability)

$$p(y \mid x=i)=\mathrm{N}(y \mid \boldsymbol{\mu}_i,\Sigma_i)$$

(likelihood function)

$$p(x \mid y)=\frac{p(x,y)}{p(y)}=\frac{p(y \mid x)\,p(x)}{\sum_{i=1}^{k} p(y \mid x=i)\,p(x=i)} \quad \text{(posterior probability)}$$

$$x^*=\arg_{x\in X}\max p(x \mid y) \quad \text{(maximum } \textit{a posteriori} \text{ or MAP)}$$

Data Classification Example

- If we have six samples and three classes, the missing data indicates the class label for each pixel. Hopefully, the estimated missing data will be close it.

$$\mathbf{I}^0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}_{6\times 3} \qquad \mathbf{I} = \begin{pmatrix} 0.1 & 0.8 & 0.1 \\ 0.3 & 0.6 & 0.1 \\ 0.4 & 0.3 & 0.3 \\ 0.7 & 0.2 & 0.1 \\ 0.3 & 0.2 & 0.5 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}_{6\times 3}$$

The true missing data **The estimated missing data**

$$x_l = \arg_{m\in\{1,\ldots,g\}} \max \mathbf{I}(l,m)$$

EM Formulation: Objective Function

- *How does a tailor make a cloth?*
  - To make a cloth that fits best to the body
- *How to estimate the parameters Θ?*
  - Estimating Θ that best fits the data:

$$Y = (y_1, y_2, \ldots, y_N)$$

- *How to evaluate the fitness between Θ and Y?*
  - The fitness between Θ (model) and Y (data) is reflected by the likelihood of Y given Θ. Therefore, parameter estimation is:

$$\Theta^* = \arg\max_{\Theta} p(Y \mid \Theta) = \arg\max_{\Theta} \prod_{i=1}^{N} p(y_i \mid \Theta) = \arg\max_{\Theta} \sum_{i=1}^{N} \log\big(p(y_i \mid \Theta)\big)$$

$$p(Y \mid \Theta) = \prod_{i=1}^{N} p(y_i \mid \Theta) \ \text{(under the independent assumption)}$$

EM Formulation: Data log-likelihood

$$\log p(Y \mid \Theta) = \log \prod_{j=1}^{N} p(y_j \mid \Theta) = \sum_{j=1}^{N} \log p(y_j \mid \Theta)$$

$$= \sum_{j=1}^{N} \log\left(\sum_{i=1}^{k} p(y_j, x_j = i \mid \Theta) \right)$$

$$= \sum_{j=1}^{N} \log\left(\sum_{i=1}^{k} p(y_j \mid x_j = i, \Theta)\, p(x_j = i \mid \Theta) \right)$$

$$= \sum_{j=1}^{N} \log\left(\sum_{i=1}^{k} \underbrace{p(y_j \mid x_j = i, \Theta)}_{\text{Likelihood function}} \underbrace{\alpha_i}_{\text{prior}} \right)$$

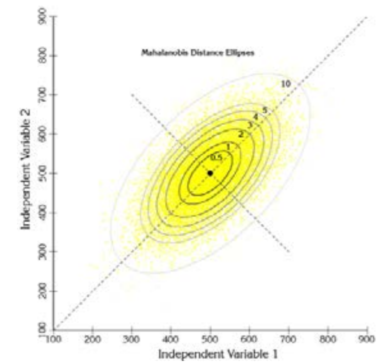
EM Formulation: Likelihood Function

- What is a likelihood function?
  - The likelihood function indicates how likely a particular distribution is to produce an observed sample. It is like a ruler for the tailor.
- What it can do for EM?
  - It allows us to estimate unknown parameters based on known outcomes.

$$p(y \mid \theta_m) = \frac{1}{(2\pi)^{d/2} \det(\Sigma_m)^{1/2}} \exp\left\{ -\frac{1}{2}(y-\mu_m)^T \Sigma_m^{-1} (y-\mu_m) \right\}$$

The dimension of y

Mahalanobis Distance



$$p(y \mid \theta_m) = p(y \mid x = m) \ \text{(likelihood function)}$$

EM Formulation: Two-Step Iteration

- W use a two-step iteration to solve the missing data problem
  - Initialize the parameters (it is like to initialize the centers in k-means)

$$\Theta = \{\alpha_i, \mu_i, \Sigma_i \mid i = 1, \ldots, k\}$$

  - Step 1: Estimate the missing data (x) in terms of the posterior probability of each data sample (y) (it is like to classify each data point in k-means.)

$$p(x = i \mid y) \ \{i = 1, \ldots, k\}$$

(estimate the probability of each sample belonging to different classes)

  - Step 2: From the estimated missing data, to obtain the maximum likelihood estimate of the parameters (it is like to update the centers in k-means)

$$\Theta^* = \arg\max_{\Theta} \log p(Y \mid \Theta)$$

(update the parameters to better fit the data and the model.)

EM Algorithm: E-step

- **Initialization**: set s=0 and

$$\Theta^0 = (\alpha_1^{(0)}, \alpha_2^{(0)}, \ldots, \alpha_g^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_g^{(0)}).$$

- **Expectation** (E-step):

$$\mathbf{I}(l,m) = \frac{\alpha_m^{(s)} p(y_l \mid \theta_m^{(s)})}{\sum_{i=1}^{k} \alpha_i^{(s)} p(y_l \mid \theta_i^{(s)})} = p(x_l = m \mid y_l, \Theta^{(s)})$$

Posterior probability

$$p(x \mid y) = \frac{p(x,y)}{p(y)} = \frac{p(y \mid x) p(x)}{\sum_x p(y \mid x) p(x)}$$

$$p(y_l \mid \theta_m^{(s)}) = \frac{\exp\left\{-\frac{1}{2}(y_l - \mu_m)^T \Sigma_m^{-1} (y_l - \mu_m)\right\}}{(2\pi)^{d/2} \det(\Sigma_l)^{1/2}}$$

Likelihood function

EM Algorithm: M-step

- Maximization (M-step): $\Theta^* = \arg\max_{\Theta} p(Y \mid \Theta)$

$$\frac{\partial \log p(\mathbf{Y} \mid \Theta^{(s+1)})}{\partial \alpha_m} = 0 \longrightarrow \alpha_m^{(s+1)} = \frac{1}{N}\sum_{l=1}^{N} p(x_l = m \mid y_l, \Theta^{(s)})$$

$$\frac{\partial \log p(\mathbf{Y} \mid \Theta^{(s+1)})}{\partial \mu_m} = 0 \longrightarrow \mu_m^{(s+1)} = \frac{\sum_{l=1}^{N} y_l\, p(x_l = m \mid y_l, \Theta^{(s)})}{\sum_{l=1}^{N} p(x_l = m \mid y_l, \Theta^{(s)})}$$

$$\frac{\partial \log p(\mathbf{Y} \mid \Theta^{(s+1)})}{\partial \Sigma_m} = 0 \longrightarrow \Sigma_m^{(s+1)} = \frac{\sum_{l=1}^{N} p(x_l = m \mid y_l, \Theta^{(s)}) \left\{ (y_l - \mu_m^{(s)})(y_l - \mu_m^{(s)})^T \right\}}{\sum_{l=1}^{N} p(x_l = m \mid y_l, \Theta^{(s)})}$$