Lecture 23

Expectation Maximization (EM)

ECEN5283 Computer Vision

Dr. Guoliang Fan

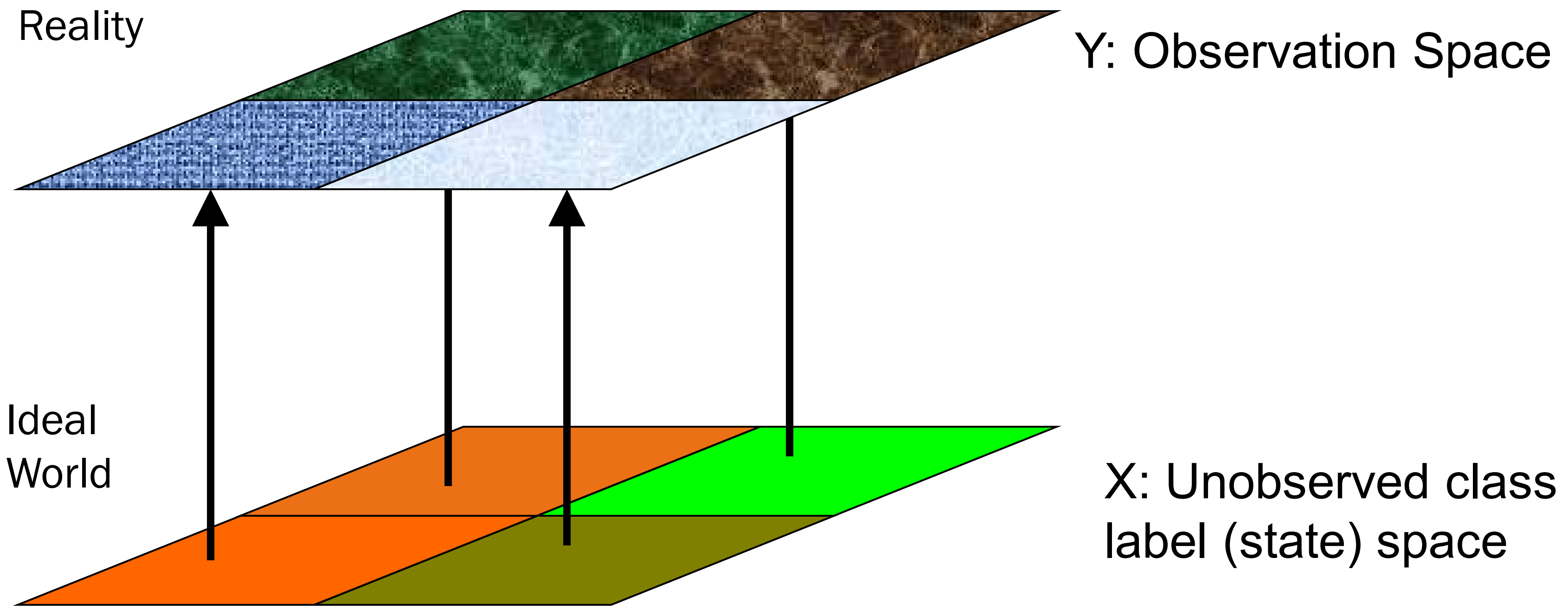
Oklahoma State University

Goals

To review the missing data problem and its two major issues

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

Missing Data Problem Restatement



Mapping $X \rightarrow Y$ loses the class label information.

Inference $Y \rightarrow X$ is needed.

Some Review of Probability Theory

Where do we start? $\alpha_i = p_X(x = i)$ (prior probability) $p_{Y|X}(y|x = i) = \mathrm{N}(y|\mu_i, \Sigma_i)$ (likelihood function)

Joint probability $p_{XY}(x, y) = p_{Y|X}(y|x)p_X(x)$

Bayes' Law of posteriori probability

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}$$

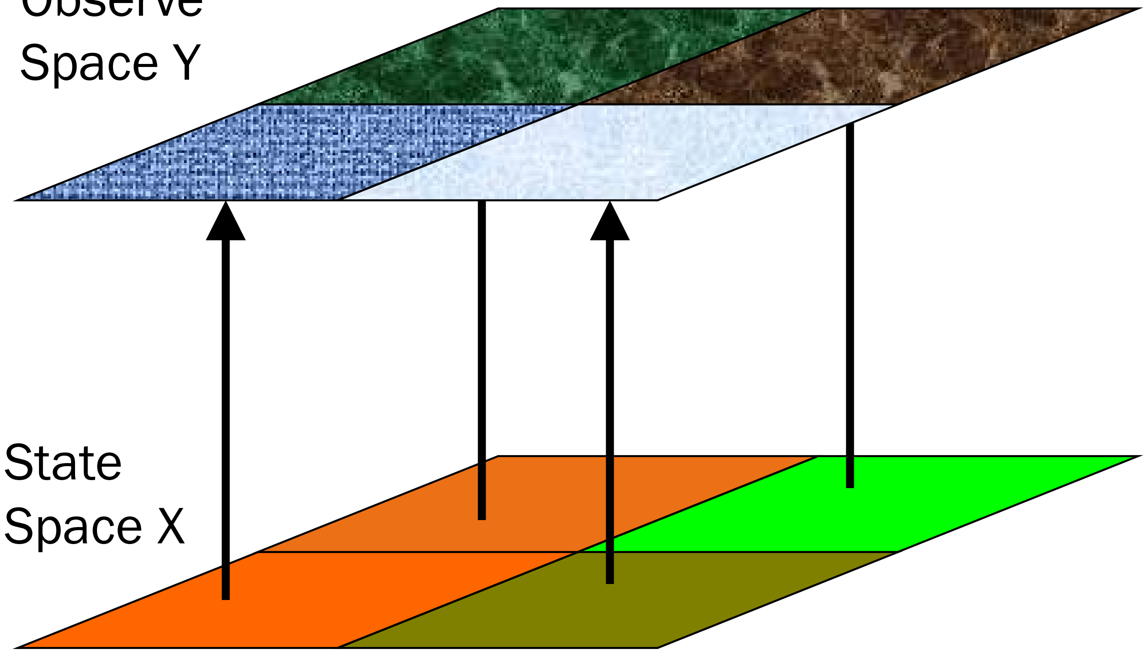
Y： observation (e.g., symptoms)

X： label (e.g., diseases)

Marginalization probability $p_Y(y) = \sum_x p_{XY}(x, y) = \sum_x p_{Y|X}(y|x)p_X(x)$

Issue (1) Parameter Estimation

Observe Space Y



State Space X

$\boldsymbol{y} = (y_1, y_2, \dots, y_N)$

Observation data

$\downarrow$

Parameter Space

$\Theta = \{\alpha_i, \mu_i, \Sigma_i | i = 1,2,3,4\}$

$p_Y(\boldsymbol{y}|\Theta)$ data likelihood

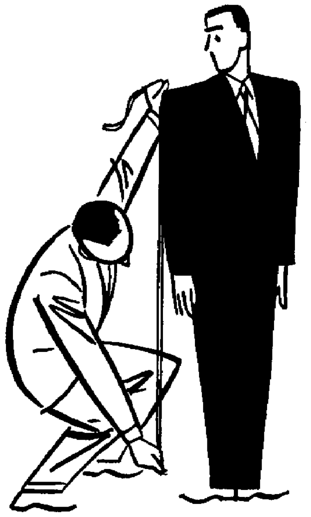
(the model–data fitness)

$$\Theta^* = \arg\max_{\Theta} p_Y(\boldsymbol{y}|\Theta)$$

(parameter estimation)

To find the parameter Θ that can best explain the current observation $\boldsymbol{y}$.

Formulation: Parameter Estimation



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 set $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$.

How to evaluate the fitness between Θ and $\boldsymbol{y}$?

- The fitness between Θ (model) and $\boldsymbol{y}$ (data) is reflected by the likelihood of $\boldsymbol{y}$ given Θ. Therefore, parameter estimation is:

$$\Theta^* = \arg\max_{\Theta} p_Y(\boldsymbol{y}|\Theta) = \arg\max_{\Theta} \prod_{i=1}^{N} p_Y(y_i|\Theta) = \arg\max_{\Theta} \sum_{i=1}^{N} \log(p_Y(y_i|\Theta))$$

$$p_Y(\boldsymbol{y}|\Theta) = \prod_{i=1}^{N} p_Y(y_i|\Theta)$$

(independent assumption)

Issue (2) Data Classification

To classify data, we need to compute the probability of data sample y belonging to class x, i.e., the posterior probability $p(x|y)$, which is computed during parameter estimation.

$\alpha_i = p_X(x = i)$
(prior probability)

$p_{Y|X}(y|x = i) = \mathrm{N}(y|\mu_i, \Sigma_i)$
(likelihood function)

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{i=1}^{k} p_{Y|X}(y|x = i)p_X(x = i)} \quad \text{(posterior probability)}$$

(Bayes' law)

$$x^* = \arg_{x\in X} \max p_{X|Y}(x|y) \quad \text{(maximum } a\ posteriori \text{ or MAP)}$$

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.

$$\mathbf{I}^0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}_{6\times 3}$$

The true missing data

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

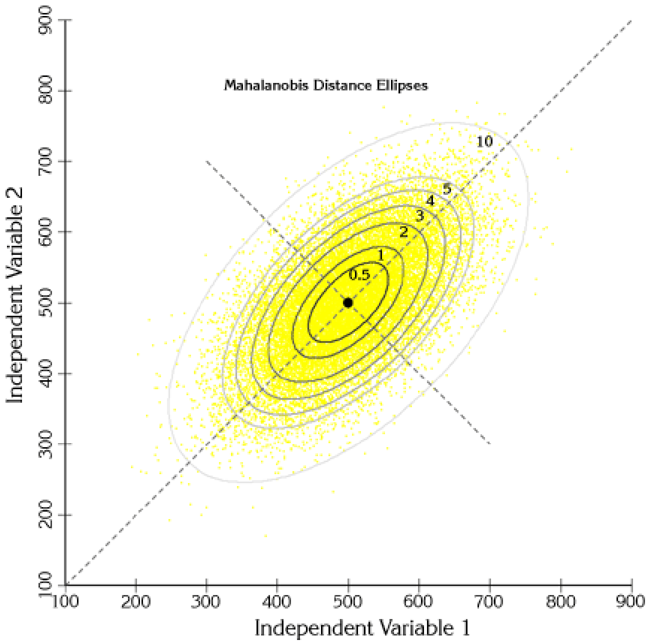
The estimated missing data

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

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.
- For cluster i, we can assume a d-dimensional Gaussian PDF as the likelihood function of data samples in that class that is defined by mean μ_i and covariance matrix Σ_i



$$p_{Y|X}(y|x=i) = N(y|\mu_i, \Sigma_i)$$

Mahalanobis Distance

$$p_Y(y|\theta_i) = \frac{1}{(2\pi)^{d/2} \det(\Sigma_i)^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu_i)^T \Sigma_i^{-1}(y-\mu_i)\right\}$$

$$\theta_i = \{\mu_i, \Sigma_i\}$$

EM Formulation: Two-Step Iteration

We use a two-step iteration to solve the missing data problem

- Initialize the parameters (it is like to initialize the centers in k-means)

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

- 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.)

$$p_{X|Y}(x = i|y) \ \{i = 1, \ldots, k\}$$

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

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

$$\Theta^* = \arg\max_{\Theta} \log p_Y(\boldsymbol{y}|\Theta)$$

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

EM Algorithm: E-step

Initialization: set $s = 0$ and

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

Expectation (E-step):

$$\mathbf{I}(l, m) = \frac{\alpha_m^{(s)} p_Y\left(y_l \middle| \theta_m^{(s)}\right)}{\sum_{i=1}^{k} \alpha_i^{(s)} p_{Y|X}\left(y_l \middle| \theta_i^{(s)}\right)} = p_{X|Y}(x_l = m | y_l, \Theta^{(s)})$$

Posterior probability

$$p_{X|Y}(x = i|y) = \frac{p_{Y|X}(y|x) p_X(x)}{\sum_{i=1}^{k} p_{Y|X}(y|x = i) p_X(x = i)}$$

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

EM Algorithm: M-step

Maximization (M-step): $\Theta^* = \arg\max_{\Theta} \log p(\boldsymbol{y}|\Theta)$

$$\alpha_m^{(s+1)} = p_X(x) = \frac{1}{N}\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)})$$

$$\theta_m^{(s+1)} \begin{cases} \mu_m^{(s+1)} = \dfrac{\sum_{l=1}^{N} y_l \cdot p_{X|Y}(x_l = m|y_l, \Theta^{(s)})}{\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)})} \\[2ex] \Sigma_m^{(s+1)} = \dfrac{\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)}) \left\{(y_l - \mu_m^{(s)})(y_l - \mu_m^{(s)})^T\right\}}{\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)})} \end{cases}$$