Lecture 22
Missing Data Problem

ECEN5283 Computer Vision

Dr. Guoliang Fan

Oklahoma State University

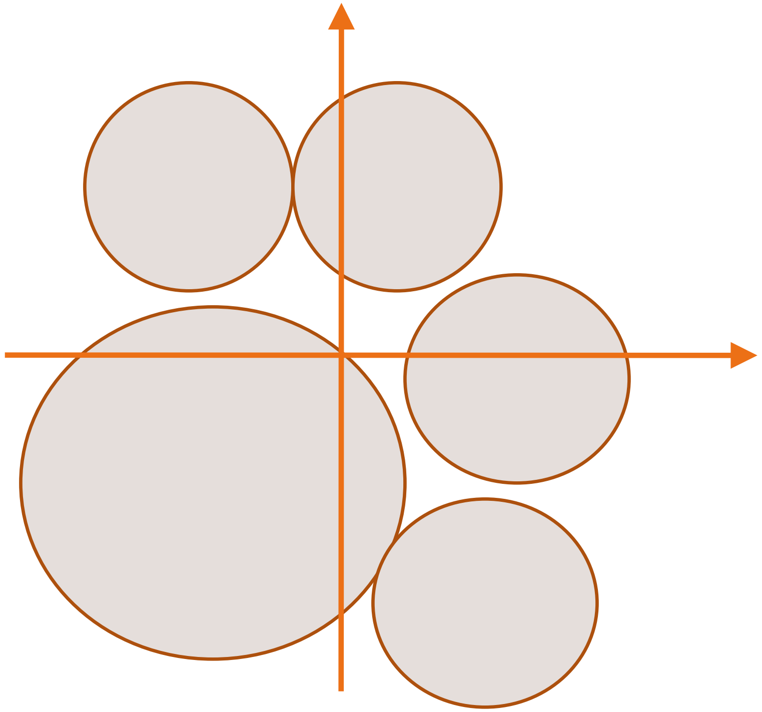
Goals

To re-visit some basic issues of clustering

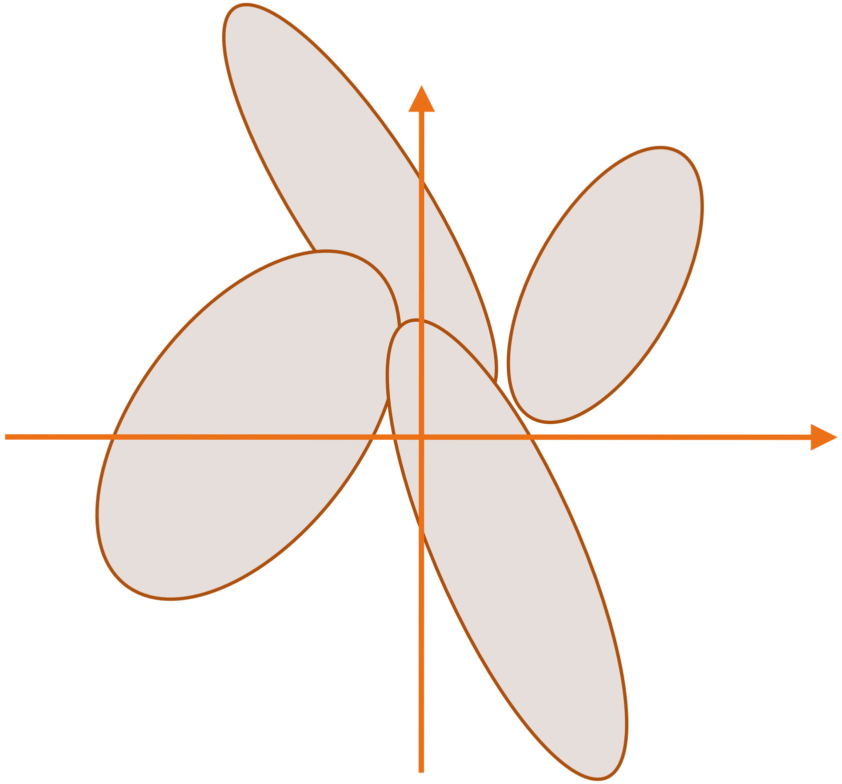
To introduce the missing data problem for classification

To formulate the missing data problem probabilistically with two basic issues, *parameter estimation* and *data classification*

Underlying Assumption of K-means



All feature distributions are isotropic and equally probable

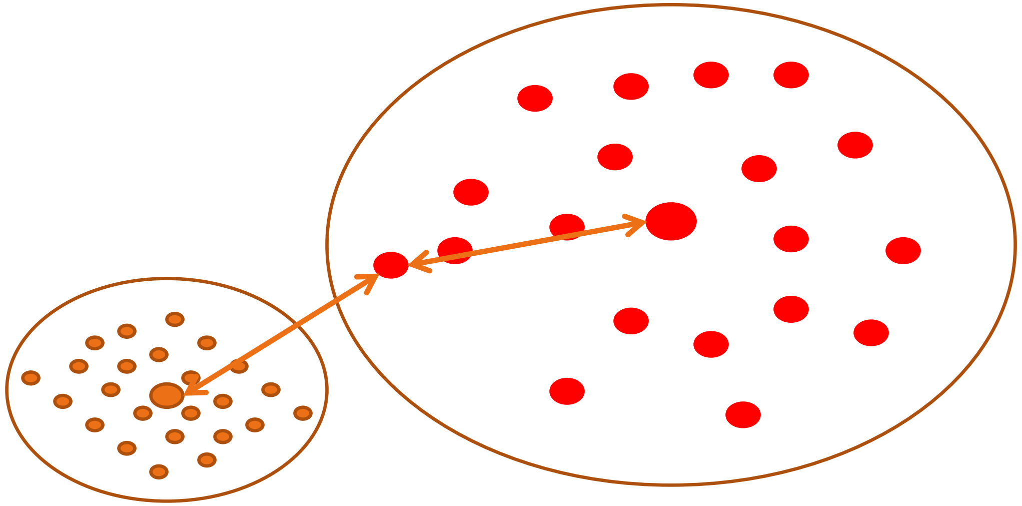


K-mean does not work well for the case of non-isotropic feature distributions

K-means: Limitations

There are three major limitations in K-means

- Does NOT consider the spread (variance) of different clusters.
- Does NOT consider the structure (distribution) of each cluster
- Does NOT consider the proportion (prior) of different clusters.

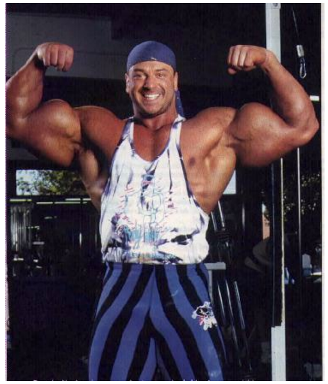


Missing Data Problem: Example

Let us consider a missing data problem example

- Assume that people can be classified into three groups according to the physical size, big, median, and small people.
- Each group is characterized by the population percentage and a 2-D Gaussian showing the distribution of weight-height.
- The reason for using Gaussian distributions instead of hard-thresholds is due to the uncertainty or error for weight-height measurement.

$\mathrm{N}(\mu_1, \Sigma_1)$
$\alpha_1 = 25\%$



$\mathrm{N}(\mu_2, \Sigma_2)$
$\alpha_2 = 60\%$

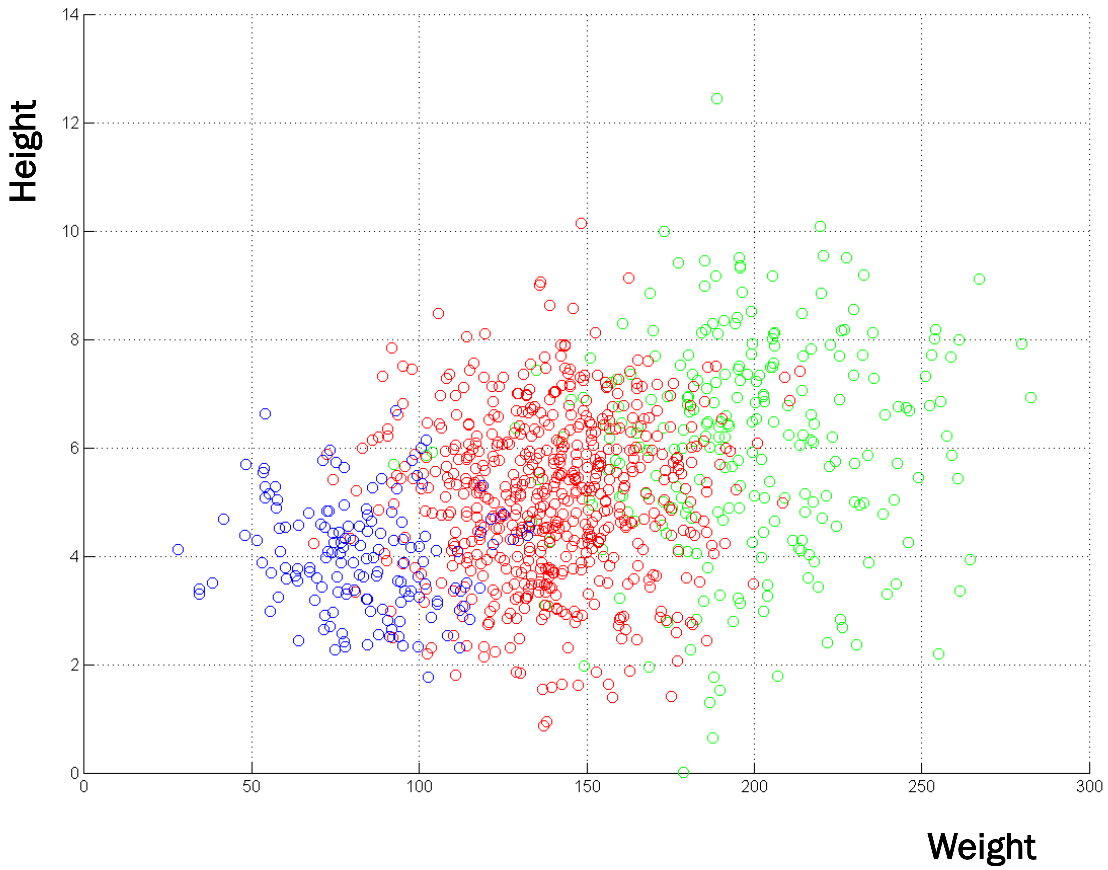
$\mathrm{N}(\mu_3, \Sigma_3)$
$\alpha_3 = 15\%$

If we don't miss anything …

$N(\mu_1, \Sigma_1), \alpha_1 = 25\%$
(big people)

$N(\mu_2, \Sigma_2), \alpha_2 = 60\%$
(median people)

$N(\mu_3, \Sigma_3), \alpha_3 = 15\%$
(small people)



Why data sets are overlapped?

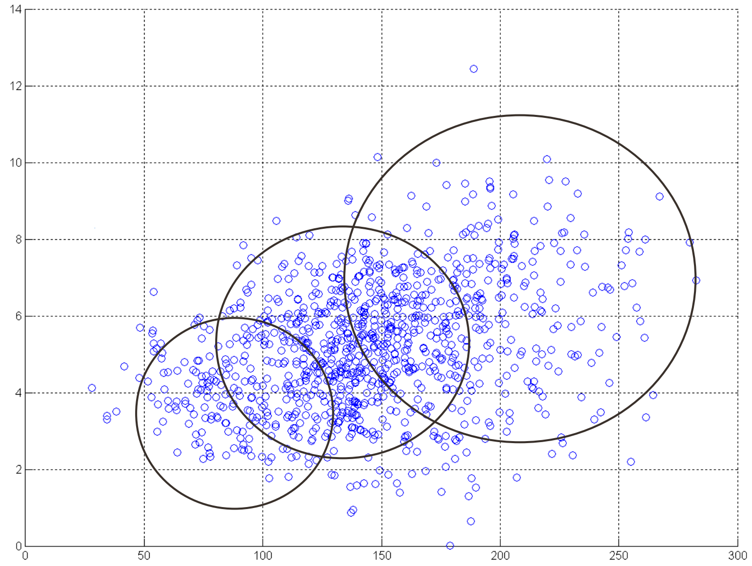
However, in reality we have …

Now you are given the statistics of a certain population, and you are given two tasks:

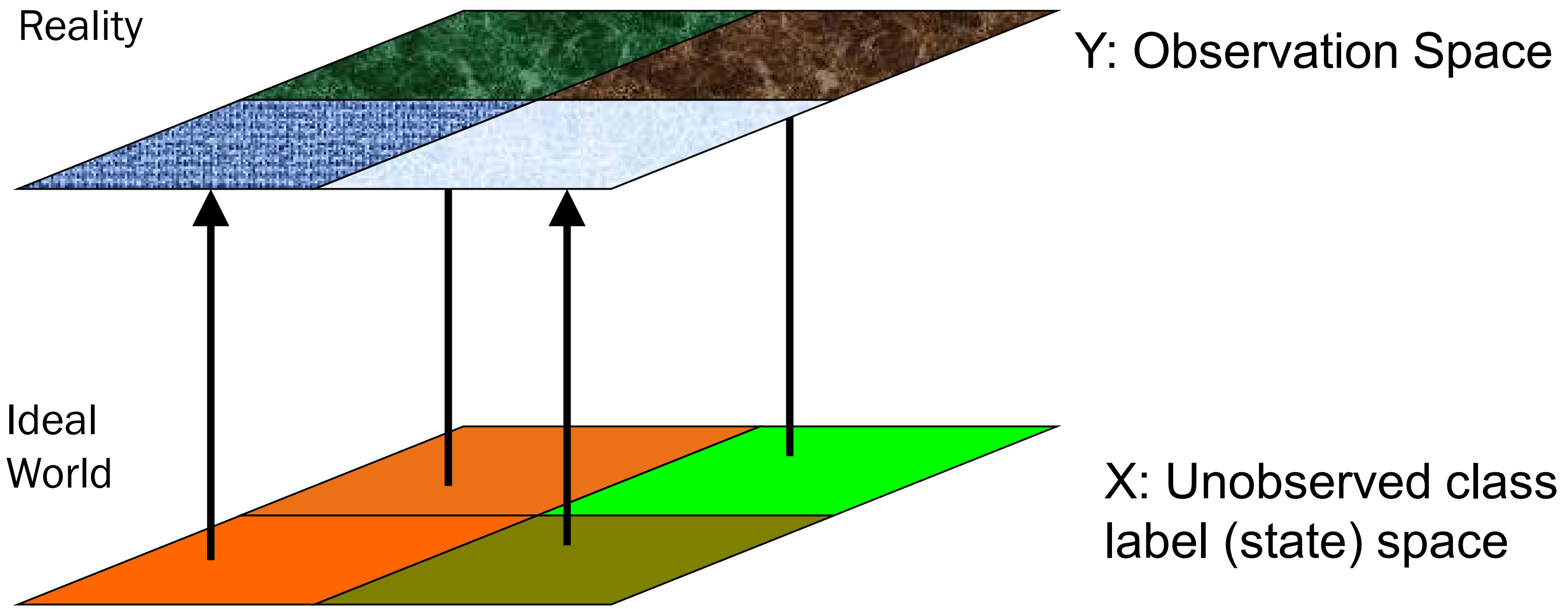
- Estimate the model parameters for each class
- Classify each data point into one of three classes

That means the class labels are missing in the data we have collected, and we need to find them.

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



Missing Data Problem Restatement



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

Inference $Y \to X$ is needed.

Probabilistic Formulation

Prior probability: something you know before you even see the data or the observation. It is like your prior knowledge.

$$\alpha_i = p_X(x = i) \quad \text{(prior probability)}$$

Likelihood function: something to evaluate how likely a D-dimensional sample y is generated from a certain class modeled by Gaussian ($\mathrm{N}(y|\mu_i, \Sigma_i)$)

$$p_{Y|X}(y|x = i) = \mathrm{N}(y|\mu_i, \Sigma_i) \qquad N(y|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right)$$

Posterior probability: based on what you see (likelihood) and what you know (prior probability) what is the probability of a data sample y belonging to certain class label. It is like the estimate of the missing data.

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

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)

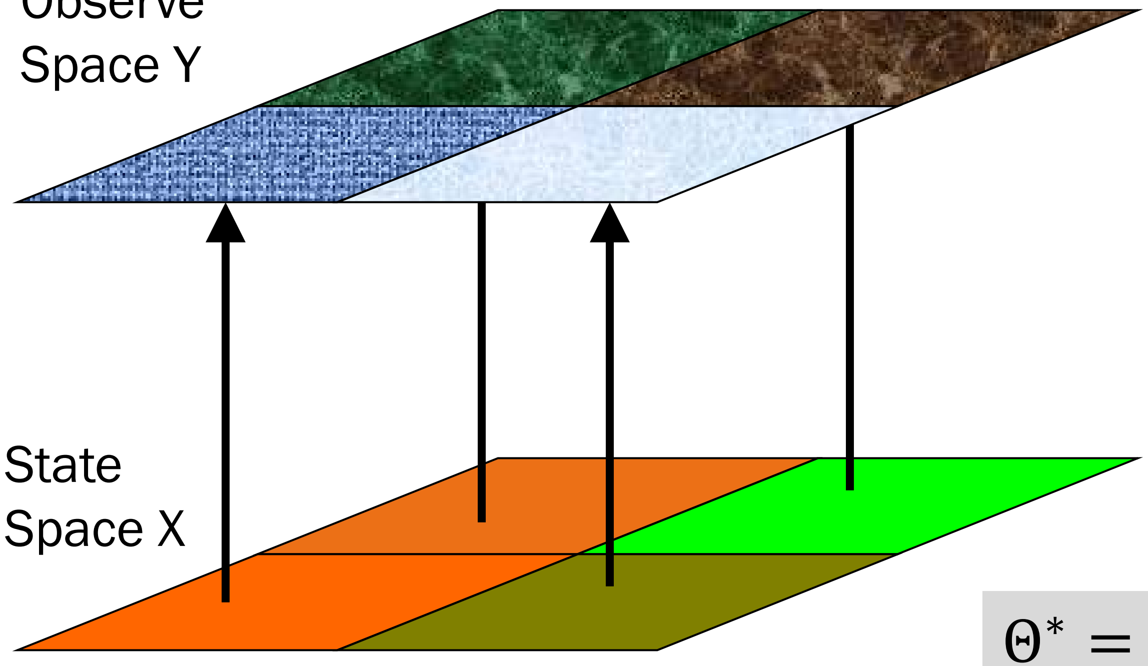
Joint probability $$p_{XY}(x, y) = p_{Y|X}(y|x)p_X(x)$$

Bayes' Law of posteriori probability

$$p_{X|Y}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \quad \text{(Bayes' law)}$$

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



Parameter Space

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

$$\downarrow$$

$$p_Y(\boldsymbol{y}|\Theta) = \prod_{y \in \boldsymbol{y}} p_Y(y|\Theta): \text{ data likelihood}$$

(the model–data fitness)

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

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

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)

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