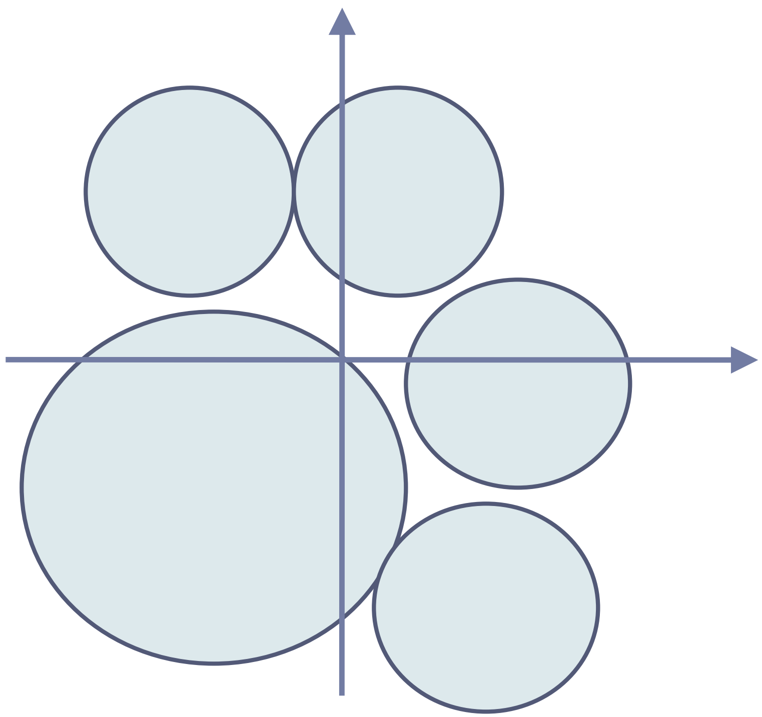
Lecture 21
Missing Data Problem
ECEN 5283 Computer Vision

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

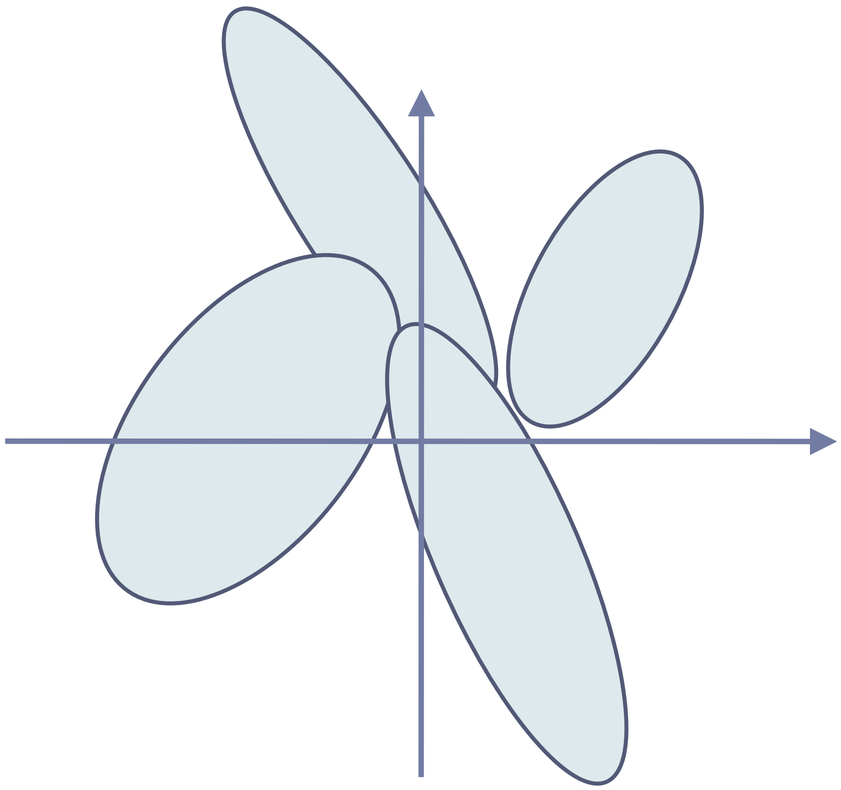
Goals

- 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



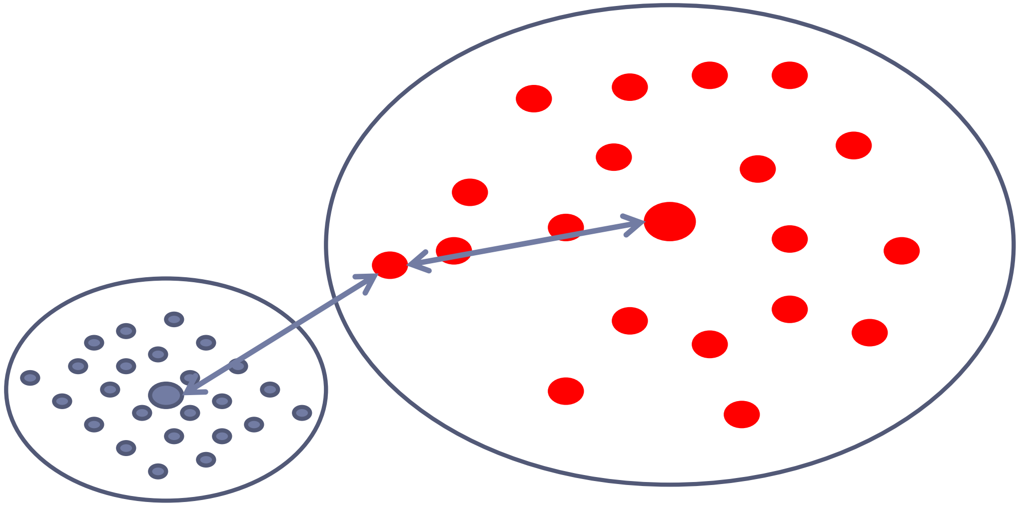
All feature distributions are isotropic due to the nature of Euclidean distance



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

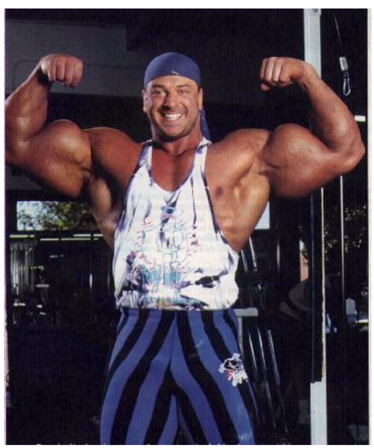
K-means: Limitations

- There are three major limitations in K-means
  - Does NOT consider the spread (variance) of different clusters.
  - Does NOT consider the structure 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.



$N(\boldsymbol{\mu}_1, \Sigma_1)$

$\alpha_1 = 25\%$

$N(\boldsymbol{\mu}_2, \Sigma_2)$

$\alpha_2 = 60\%$

$N(\boldsymbol{\mu}_3, \Sigma_3)$

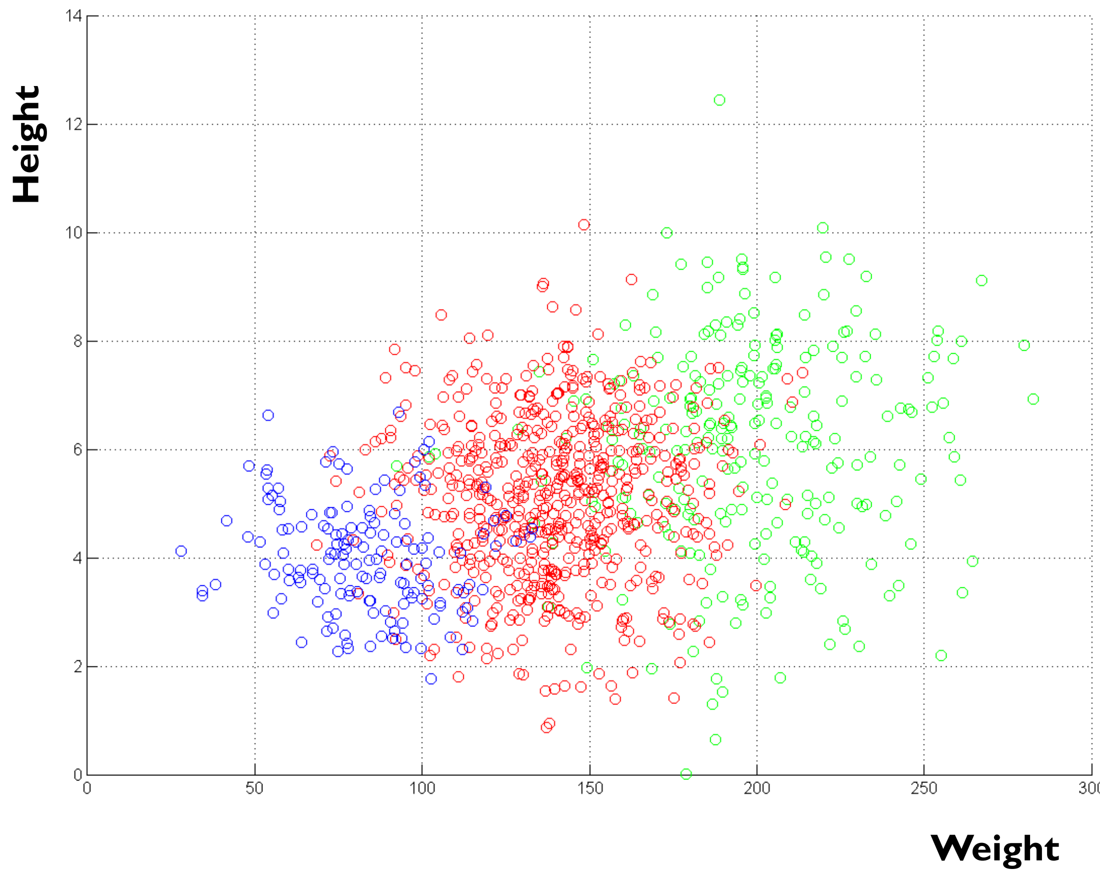
$\alpha_3 = 15\%$

If we don't miss anything …

$N(\mathbf{\mu}_1, \Sigma_1), \alpha_1 = 25\%$
(big people)

$N(\mathbf{\mu}_2, \Sigma_2), \alpha_2 = 60\%$
(median people)

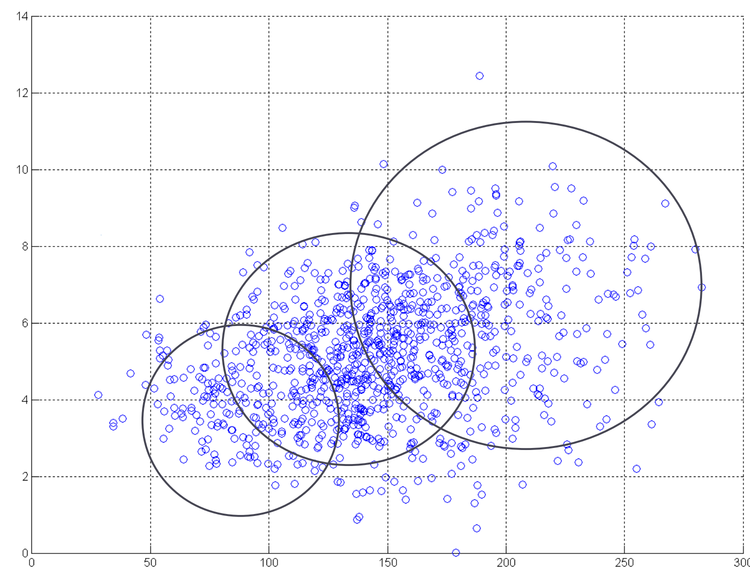
$N(\mathbf{\mu}_3, \Sigma_3), \alpha_3 = 15\%$
(small people)



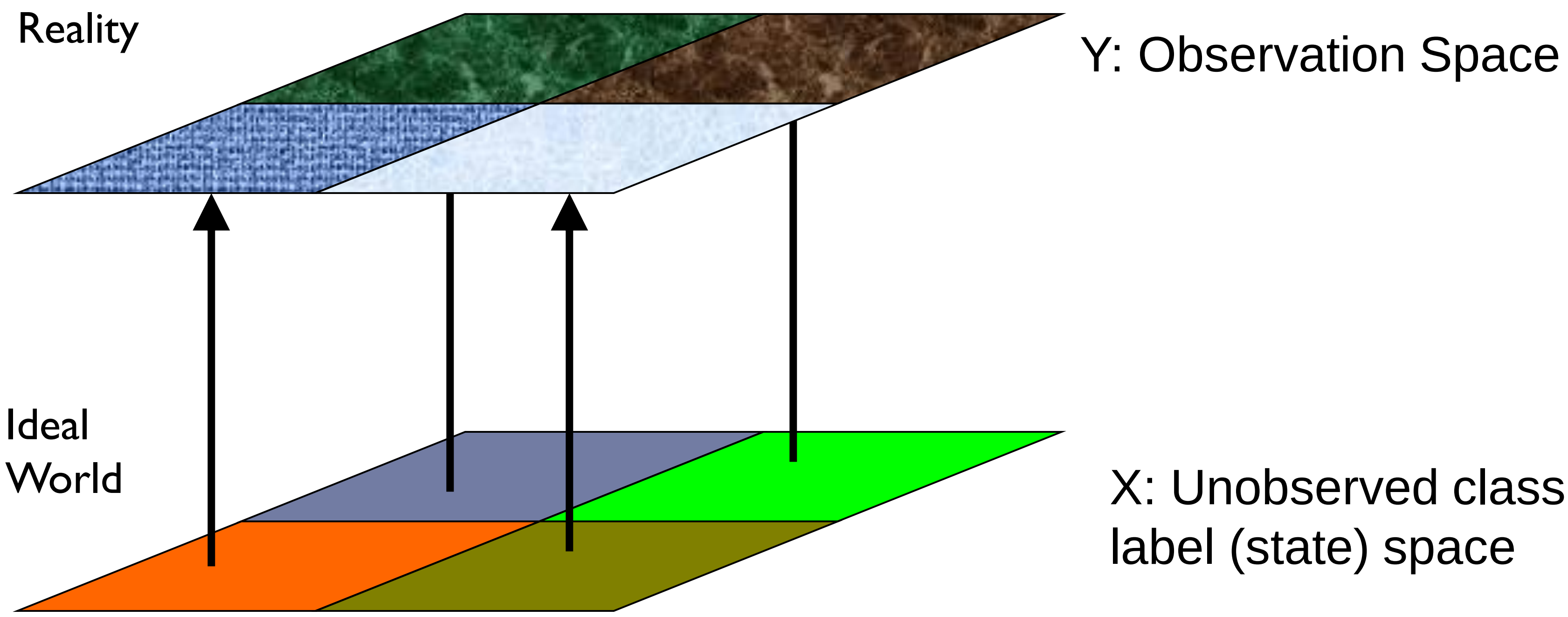
However, in reality we have …

- 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.

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



Missing Data Problem Restatement



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

Inference $Y \rightarrow X$ is needed.

Probabilistic Formulation

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

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

- **Likelihood function:** something to evaluate how likely a data sample is generated from a certain class. It is like your observed evidence.

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

- **Posterior probability:** based on what you see (likelihood) and 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.

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

Some review of probabilistic theory

- Where do we start? $\alpha_i = p(x=i)$ (prior probability) $p(y \mid x=i) = \mathrm{N}(y \mid \boldsymbol{\mu}_i, \Sigma_i)$ (likelihood function)

- Joint probability

$$p(x, y) = p(y \mid x) p(x)$$

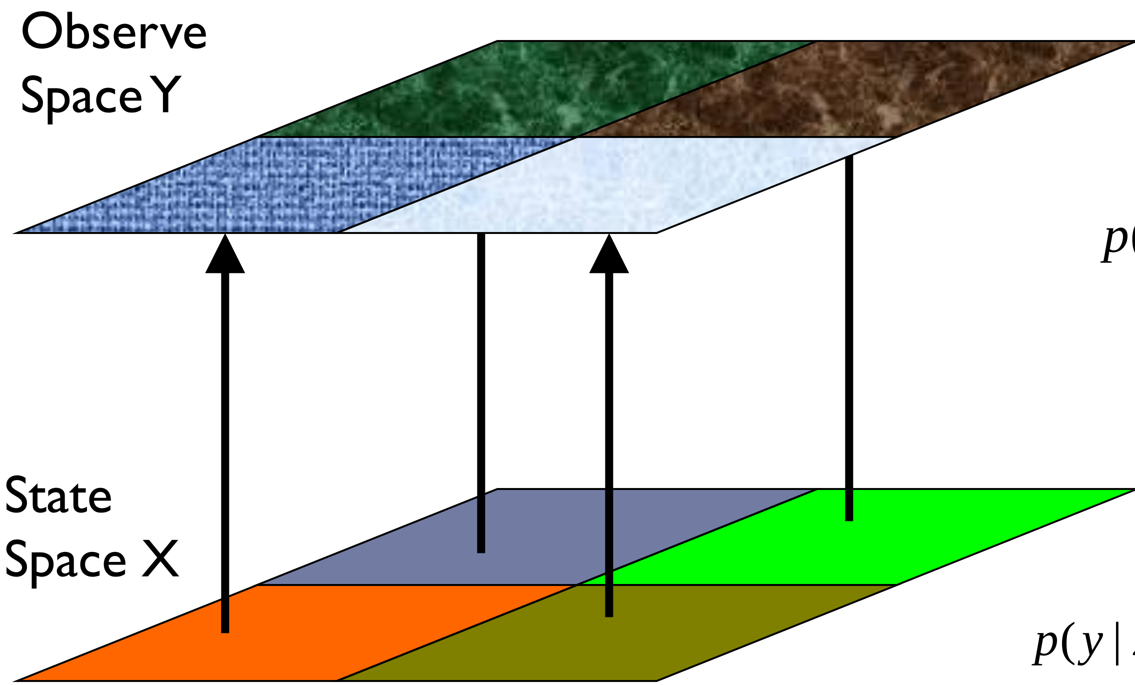
- Bayes' Law of posteriori probability

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(y \mid x) p(x)}{p(y)} \quad \text{(Bayes' law)}$$

- Marginalization probability

$$p(y) = \sum_x p(x, y) = \sum_x p(y \mid x) p(x)$$

Issue (1) Parameter Estimation



Parameter Space

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

$$\downarrow$$

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

(the model - data fitness)

$$p(y) = \sum_x p(y \mid x)\, p(x)$$

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

likelihood function

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

(prior probability)

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

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

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.

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

(prior probability)

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

(likelihood function)

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{p(y / x) p(x)}{\sum_{i=1}^{k} p(y \mid x = i) p(x = i)}$$

(Bayes' law) (posterior probability)

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