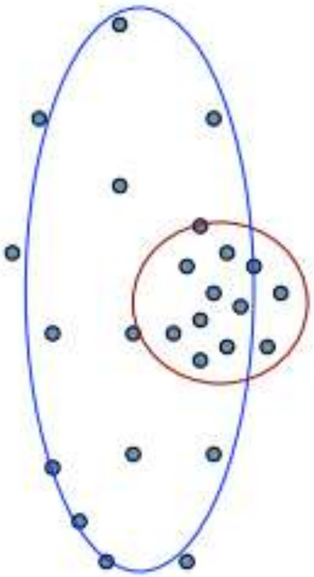
Data Mining

UNIT- V
Cluster Analysis

The Evils of "**Hard Assignments**"?



- Clusters may overlap
- Some clusters may be "wider" than others
- Distances can be deceiving!

Clustering Using Mixture Models

- This clustering approach is based on statistical models.
- Mixture Models
  - Mixture models view the data as a set of observations from a mixture of different probability distributions.
- Mixture models correspond to process of generating data.
  - Given several distributions
  - Randomly select one of these distributions
  - Generate object from it

Clustering Using Mixture Models

- Mathematical representation
  - K distributions and m objects

$$X = \{x_1, x_2, x_3, \dots, x_m\}$$

  - *Let $\boldsymbol{\theta}$ be the set of all parameters*

$$\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \dots, \theta_k\}$$

  - The probability of an object x is given by

$$prob(\mathbf{x}|\Theta) = \sum_{j=1}^{K} w_j p_j(\mathbf{x}|\theta_j)$$

$$\sum_{j=1}^{K} w_j = 1$$

Clustering Using Mixture Models

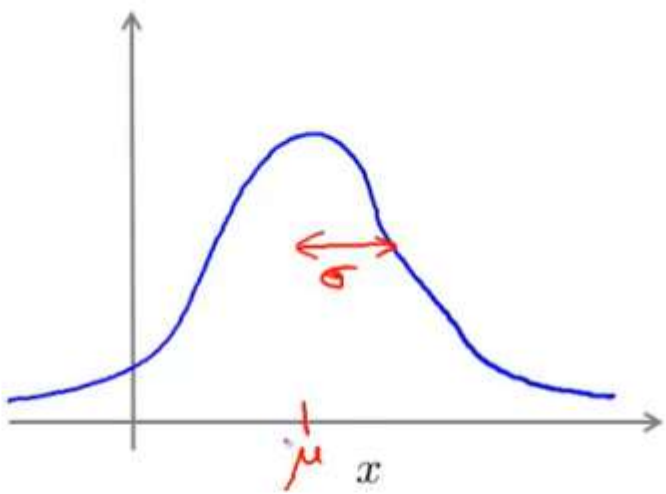
- If the objects are generated in an independent manner

$$prob(\mathcal{X}|\Theta) = \prod_{i=1}^{m} prob(\mathbf{x}_i|\Theta) = \prod_{i=1}^{m}\sum_{j=1}^{K} w_j p_j(\mathbf{x}_i|\theta_j)$$

Clustering Using Mixture Models

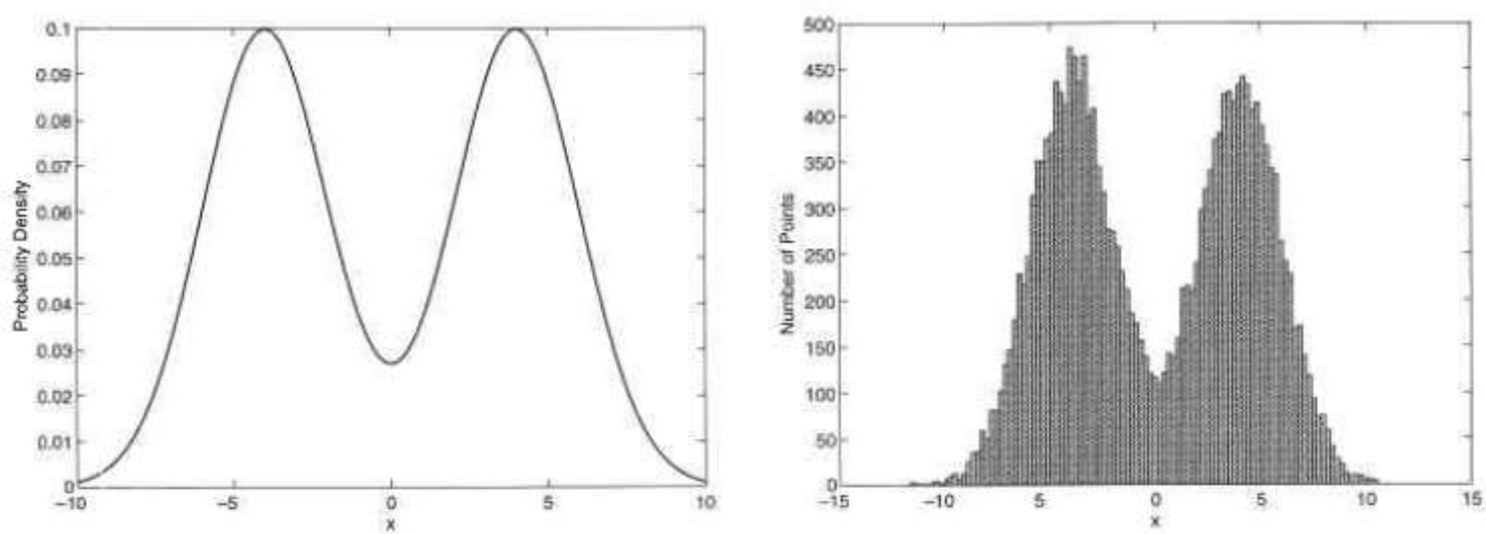
- Univariate Gaussian Mixture

$$prob(x_i|\Theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Clustering Using Mixture Models

- Assume two Gaussian distributions, with a common standard deviation of 2 and means of -4 and 4, respectively.



(a) Probability density function for the mixture model.

(b) 20,000 points generated from the mixture model.

Clustering Using Mixture Models

- Assume distributions is selected with equal probability, i.e., w1 = w2 = 0.5.

$$prob(x_i|\Theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

$$prob(x|\Theta) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x+4)^2}{8}} + \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-4)^2}{8}}.$$

Clustering Using Mixture Models

- Estimating Model Parameters Using Maximum Likelihood
  - A standard approach used for parameter estimation is maximum likelihood estimation
- MLE
  - To begin, consider a set of *m* points that are generated from a one dimensional Gaussian distribution.
  - Assuming that the points are generated independently

$$prob(\mathcal{X}|\Theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-u)^2}{2\sigma^2}}$$

Clustering Using Mixture Models

$$log\ prob(\mathcal{X}|\Theta) = -\sum_{i=1}^{m} \frac{(x_i - u)^2}{2\sigma^2} - 0.5m \log 2\pi - m \log \sigma$$

- Now we would like to find a procedure to estimate the parameters.

Clustering Using Mixture Models

- In other words, choose the µ and σ that maximize.

$$prob(\mathcal{X}|\Theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-u)^2}{2\sigma^2}}$$

- This approach is known in statistics as the maximum likelihood principle,
- The process of applying this principle to estimate the parameters of a statistical distribution from the data is known as maximum likelihood estimation (MLE).

Clustering Using Mixture Models

- MLE

$$likelihood(\Theta|\mathcal{X}) = L(\Theta|\mathcal{X}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$log\ likelihood(\Theta|\mathcal{X}) = \ell(\Theta|\mathcal{X}) = -\sum_{i=1}^{m} \frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m\log 2\pi - m\log\sigma$$

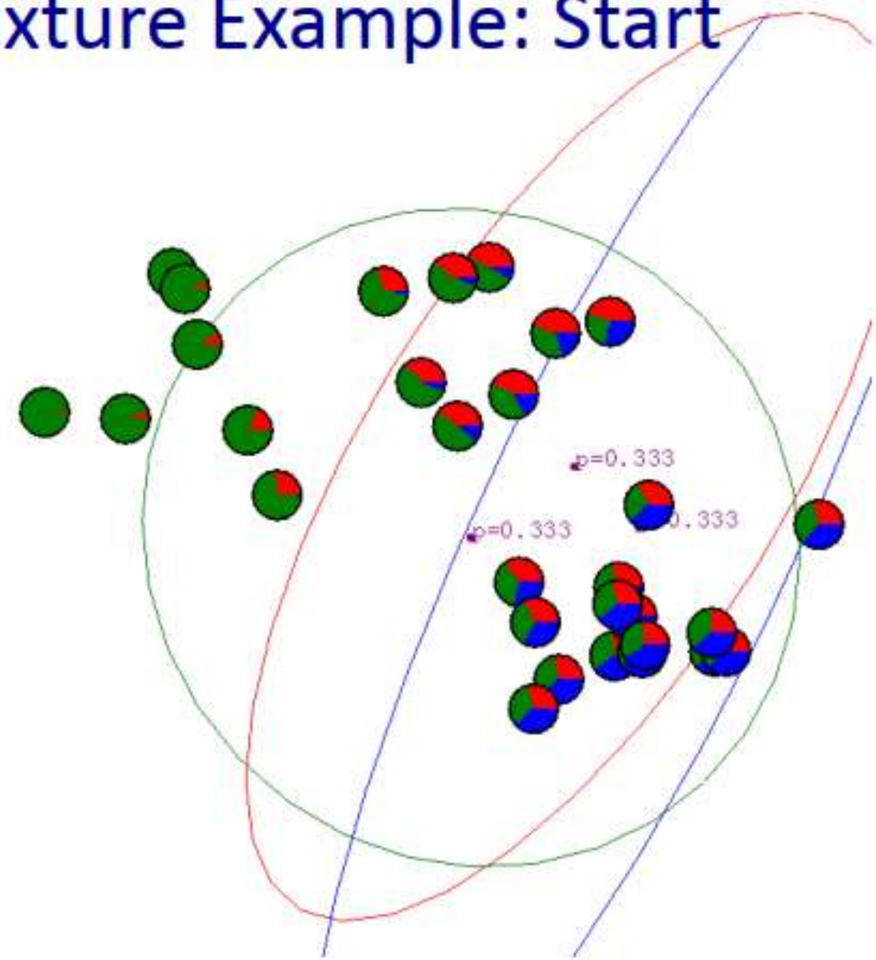
Clustering Using Mixture Models

Algorithm 9.2 EM algorithm.

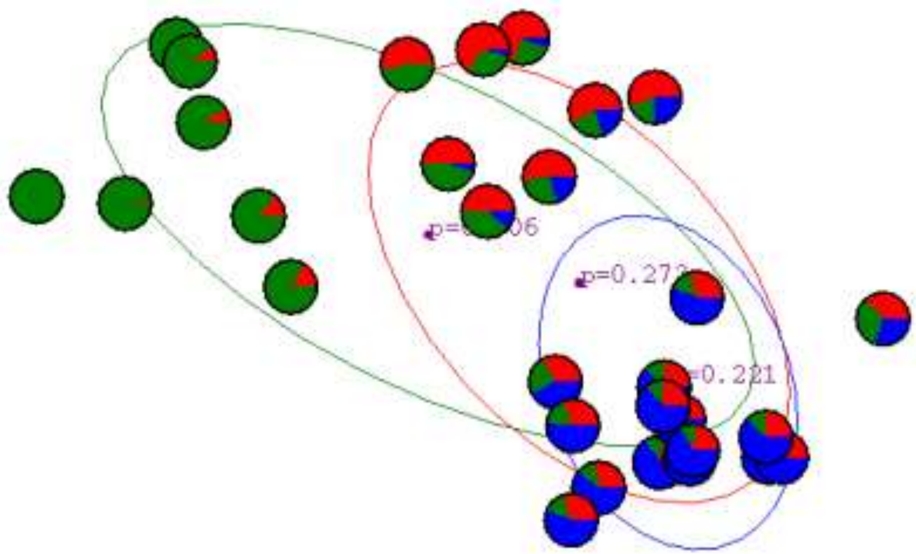
1: Select an initial set of model parameters.
(As with K-means, this can be done randomly or in a variety of ways.)
2: **repeat**
3: **Expectation Step** For each object, calculate the probability that each object belongs to each distribution, i.e., calculate $prob(distribution\ j|\mathbf{x}_i, \Theta)$.
4: **Maximization Step** Given the probabilities from the expectation step, find the new estimates of the parameters that maximize the expected likelihood.
5: **until** The parameters do not change.
(Alternatively, stop if the change in the parameters is below a specified threshold.)

Example

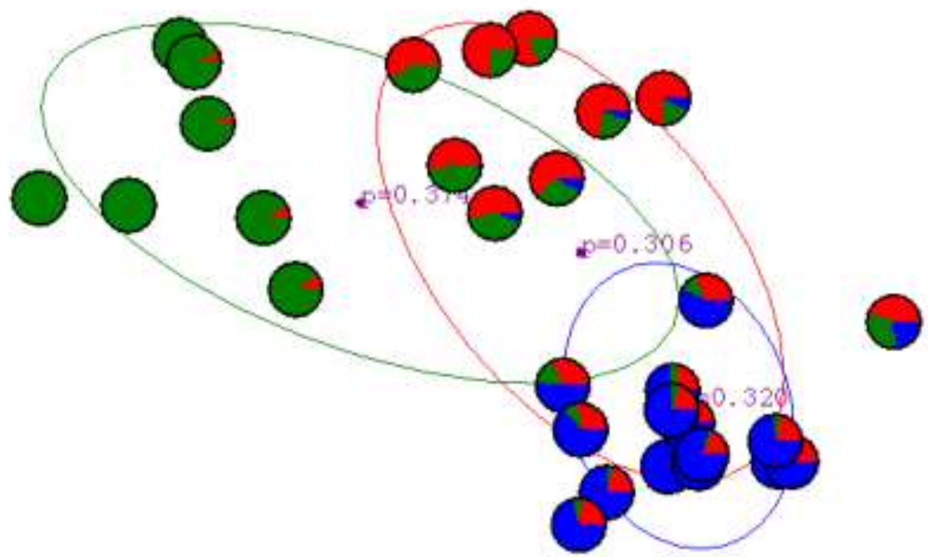
Gaussian Mixture Example: Start



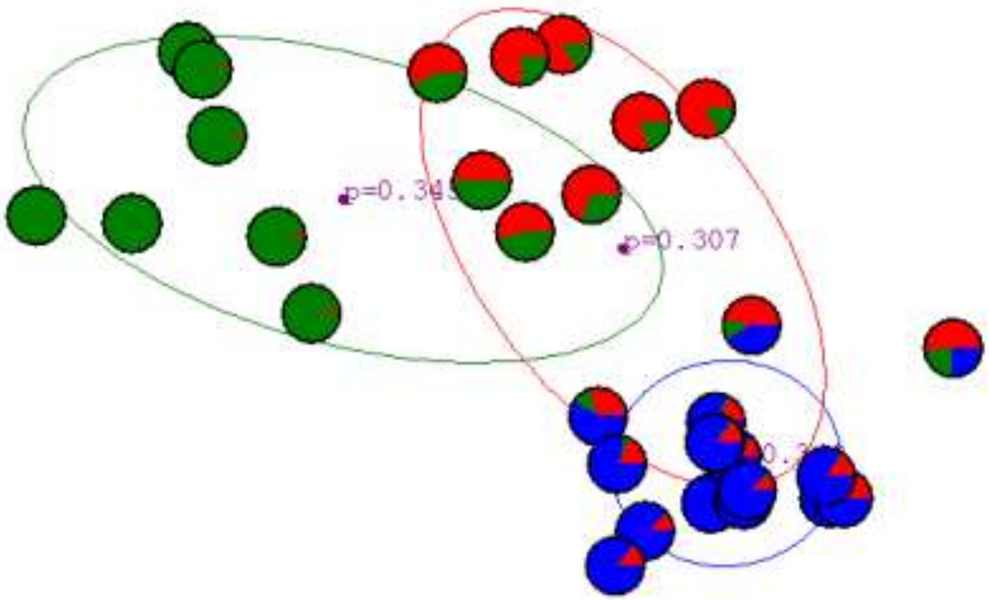
After first iteration



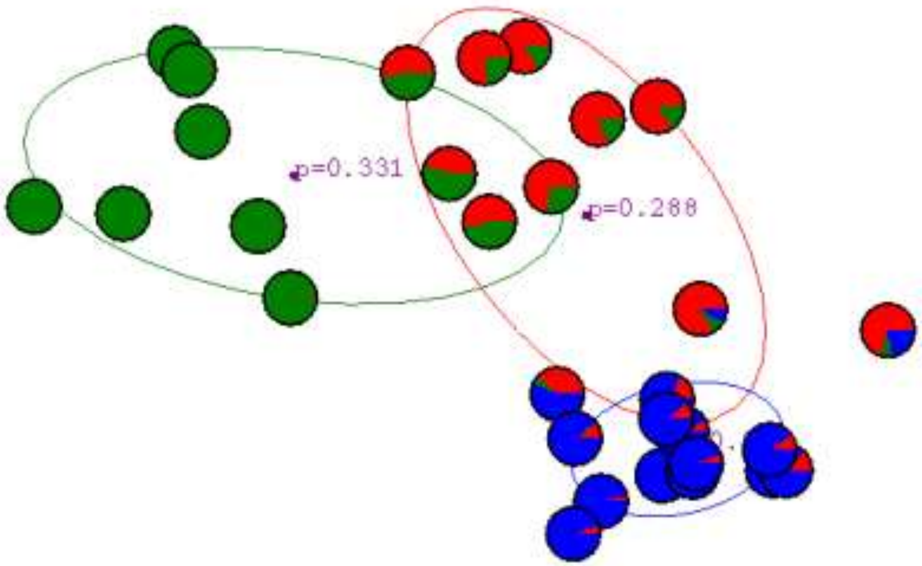
After 2nd iteration



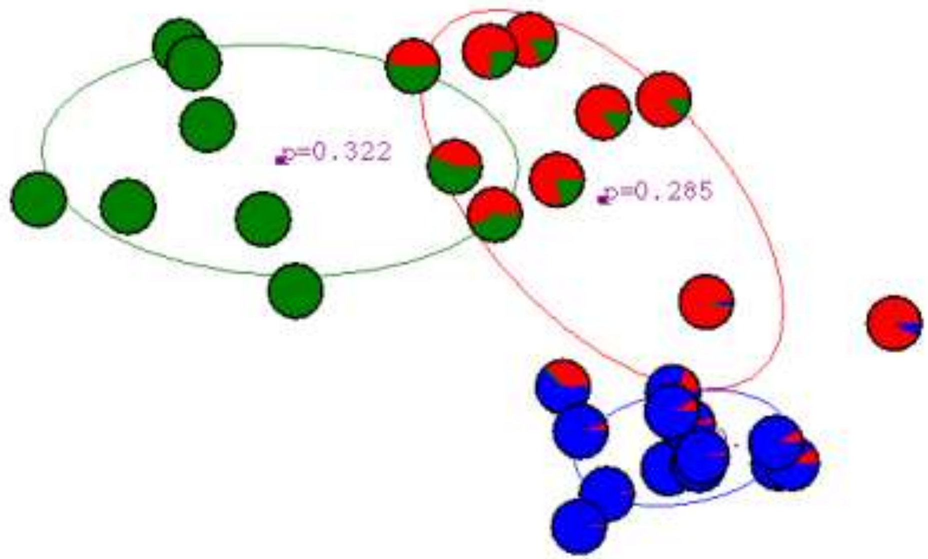
After 3rd iteration



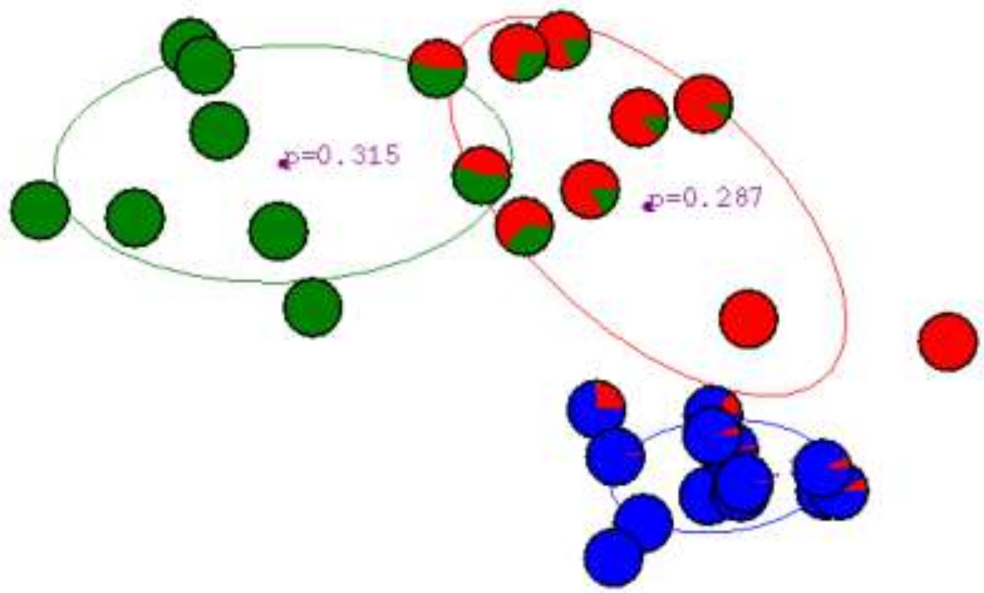
After 4th iteration



After 5th iteration



After 6th iteration



After 20th iteration