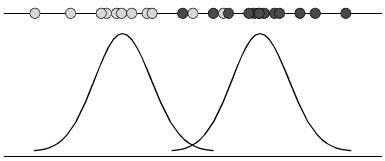
CS522 Advanced Database Systems
Clustering: Model-Based Methods

Chengyu Sun
California State University, Los Angeles

Basic Idea

- Data in a cluster is generated by a random process according to some probability distribution



Model-Based Clustering

- Make an assumption about the probability distributions
  - Usually *multivariate normal distribution* is used
- Estimate the parameters of the distributions
  - E.g. μ and σ for a normal distribution
- Calculate the probability of a data point belonging to a cluster

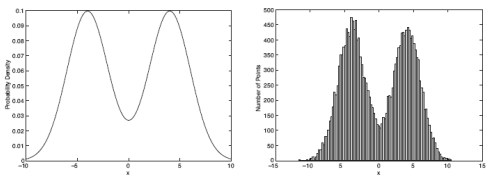
Parameter Estimation – Single Cluster

- Data: $\{x_1, x_2, x_3, \ldots, x_n\}$
- Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ=?? σ=??

Mixture Models



(a) Probability density function for the mixture model.

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

Figure 9.2. Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.

Notations

D	The dataset
x_i	An object in the dataset
N	The number of objects in the dataset
C_j	A cluster
θ_j	The parameters of the jth distribution
Θ	The set of all parameters, i.e. $\{\theta_1, \theta_2 \ldots, \theta_m\}$
M	The number of clusters/distributions

Probabilities Under Mixture Models

The probability of the `j`th distribution is chosen to generate an object: w_j , $\sum_{j=1}^{M} w_j = 1$

The probability of the `i`th object if it comes from the `j`th distribution: $P(x_i \mid \theta_j)$

The probability of an object `x`: $p(x_i \mid \Theta) = \sum_{j=1}^{M} w_j P(x_i \mid \theta_j)$

The probability of the dataset `D`: $p(D \mid \Theta) = \prod_{i=1}^{N} \sum_{j=1}^{M} w_j P(x_i \mid \theta_j)$

Maximum Likelihood Estimation (MLE)

- Choose parameters such that `P(D|Θ)` is maximized, i.e. the given data is the most probable

EM (Expectation-Maximization) Algorithm

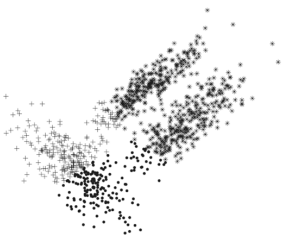
1. Select an initial set of parameters
2. Expectation Step: calculate each `P(x_i∈C_j)`
3. Maximization Step: use the probabilities from the previous step to recalculate the parameters
4. Repeat 2 until the parameters do not change

Maximization Example

- Recalculate the mean of the kth cluster

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i P(x_i \in C_k)}{\sum_{j=1}^{M} P(x_i \in C_j)}$$

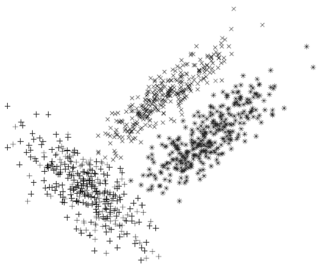
EM vs. K-Means …



(b) Clusters produced by K-means clustering.

Figure 9.6. Mixture model and K-means clustering of a set of two-dimensional points.

… EM vs. K-Means



(a) Clusters produced by mixture model clustering.

Readings

- Textbook 7.8.1