Lecture 6. Probabilistic Model-Based Clustering Methods

Lecture 6. Probabilistic Model-Based Clustering Methods

- ❑ Basic Concepts of Probabilistic Model-Based Clustering
- ❑ Mixture Models for Cluster Analysis
- ❑ Gaussian Mixture Models
- ❑ The Expectation-Maximization (EM) Algorithm (Univariate)
- ❑ The Expectation-Maximization (EM) Algorithm (Multivariate)
- ❑ Analysis of the Mixture Model Methods
- ❑ Summary

Session 1: Basic Concepts of Probabilistic Model-Based Clustering

Probabilistic Model-Based Clustering: Basic Concepts

- Probabilistic model
  - Model the data from a *generative process*
  - Assume the data are generated by a mixture of underlying probability distributions
  - Attempt to optimize the fit between the observed data and some mathematical model using a probabilistic approach
- Probabilistic model-based clustering
  - Each cluster can be represented mathematically by a parametric probability distribution (e.g., Gaussian or Poisson distribution)
  - Cluster: Data points (or objects) that most likely belong to the same distribution
  - Clustering: Parameter estimation so that they will have a *maximum likelihood fit* to the model by a mixture of *K* component distributions (i.e., K clusters)
- Broad applications
  - Image segmentation, document clustering, topic modeling, etc.

Typical Probabilistic Model-Based Clustering Methods

- **Mixture models**
  - Assume observations to be clustered are drawn from one of several components
  - Infer the parameters of these components (i.e., clusters) and assign data points to specific components of the mixture
- The **Expectation-Maximization (EM)** algorithm
  - A general technique to find maximum likelihood estimations in mixture models
  - The EM algorithm for Gaussian mixture model
- **Probabilistic topic models** for text clustering and analysis (to be covered in the "Text Mining" course)
  - Probabilistic latent semantic analysis (PLSA)
  - Latent Dirichlet allocation (LDA)

Session 2: Mixture Model for Cluster Analysis

Model-Based Clustering

- A set C of k probabilistic clusters $C_1, \dots, C_k$ with probability density functions $f_1, \dots, f_k$, respectively, and their probabilities $\omega_1, \dots, \omega_k$
- Probability of an object o generated by cluster C_j is $P(o, C_j) = P(C_j)P(o|C_j) = \omega_j f_j(o)$
- Probability of o generated by the set of cluster $\boldsymbol{C}$ is $P(o|\mathbf{C}) = \sum_{j=1}^{k} \omega_j f_j(o)$
- Since objects are assumed to be generated independently, for a data set $D = \{o_1, \dots, o_n\}$, we have

$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

- Task: Find a set C of k probabilistic clusters so that $P(D|\boldsymbol{C})$ is maximized
  - Maximizing $P(D|\boldsymbol{C})$ is often intractable since the probability density function of a cluster can take an arbitrarily complicated form
  - To make it computationally feasible (as a compromise), assume the probability density functions are some parameterized distributions

Parametric Mixed Models

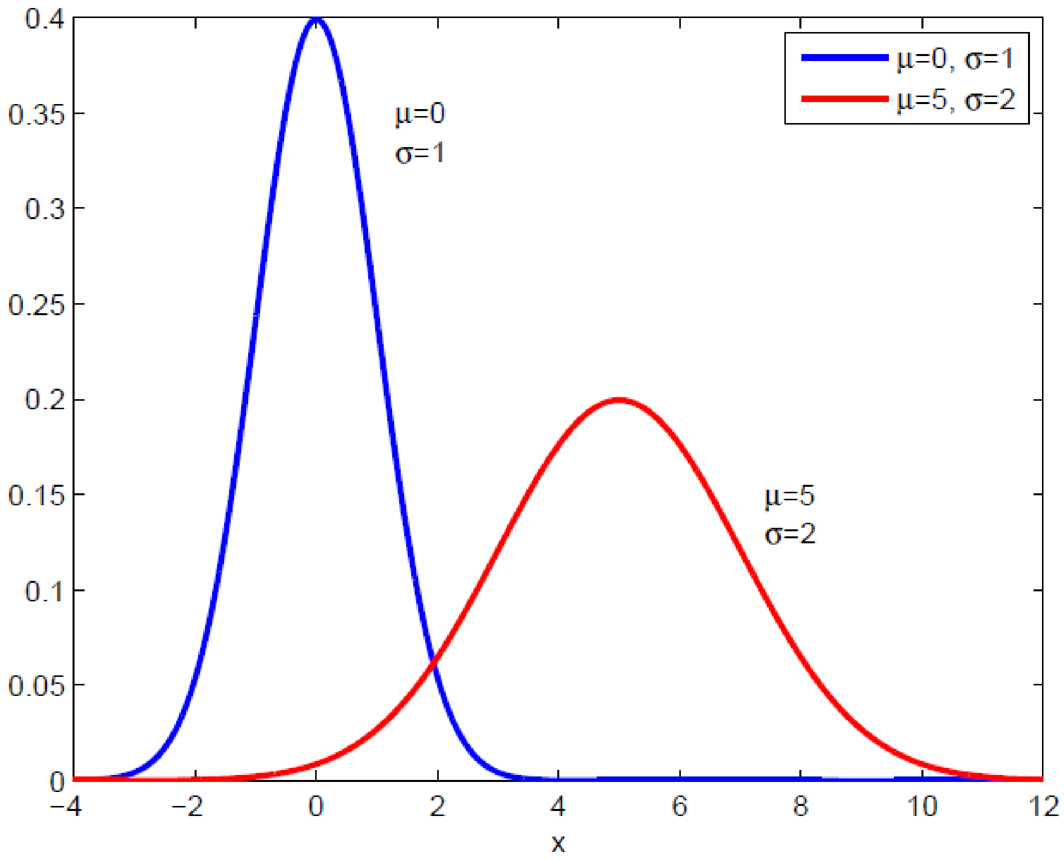
- **Our task** is to infer a set of K probabilistic clusters that is mostly likely to generate D
  - The values of the discrete latent variables can be interpreted as the assignments of data points to specific components (i.e., clusters) of the mixture
- Each cluster is mathematically represented by a **parametric distribution**
- In principle, the mixtures can be constructed with any types of components, and we could still have a perfectly good mixture model
- In practice, a lot of effort is given over to **parametric mixture models**, where all components are from the same parametric family of distributions but with different parameters
  - Ex. All Gaussians with different means and variances, all Poisson distributions with different means, or all power laws with different exponents
- Two most common mixtures: Mixture of **Gaussian** (continuous) and mixture of **Bernoulli** (discrete) distributions

Session 3: Gaussian Mixture Models

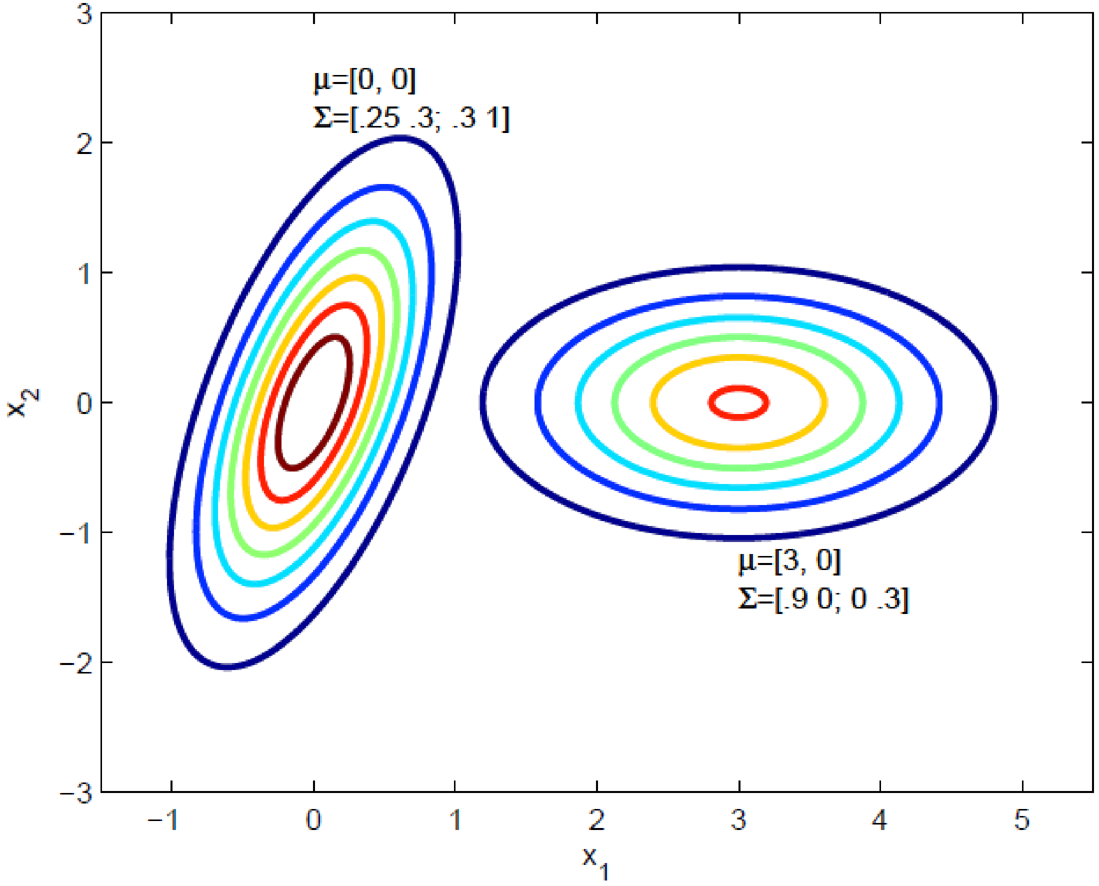
Univariate and Multivariate Gaussian Distributions

- Plots and contours for Gaussian distributions for various parameters

Plots of the univariate Gaussian distribution for various parameters of μ and σ



Contours of the multivariate (2-D) Gaussian distribution for various parameters of μ and Σ



Gaussian Mixture Model

- We assume each cluster C_i is characterized by a multivariate normal distribution

$$f_i(\mathrm{x}) = f(\mathrm{x} \mid \mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_\mathrm{i} \mid}} \exp\{-\frac{(\mathrm{x}-\mu_i)^T \Sigma_i^{-1}(\mathrm{x}-\mu_i)}{2}\}$$

where the cluster mean μ_i and covariance matrix $\sum_i$ are unknown parameters, and $f_i(x)$ is the probability density x attributable to cluster C_i

- We assume the probability density function of ***X*** is given as a Gaussian mixture model over all the k cluster normals defined as

$$f(\mathrm{x}) = \sum_{i=1}^{k} f_i(\mathrm{x}) P(C_i) = \sum_{i=1}^{k} f(\mathrm{x} \mid \mu_\mathrm{i}, \Sigma_\mathrm{i}) P(C_i)$$

where the prior probabilities $P(C_i)$ (called mixture parameters) must satisfy

$$\sum_{i=1}^{k} P(C_i) = 1$$

Maximum Likelihood Estimation of Gaussian Mixture Model

- Maximum Likelihood Estimation (MLE)
  - Given the dataset D, the likelihood of the model parameters ϑ is:

$$P(\mathbf{D}\,|\,\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\mathrm{x}_j) \quad \text{or written as} \quad \ln P(\mathbf{D}\,|\,\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln f(\mathrm{x}_j) = \sum_{j=1}^{n} \ln \sum_{i=1}^{k} f(\mathrm{x}_j|\mu_i,\Sigma_i) P(C_i)$$

  - MLE is to choose parameters ϑ: $\theta^* = \arg\max_{\theta}\{P(D\,|\,\theta)\}$
    - or maximize the log-likelihood: $\theta^* = \arg\max_{\theta}\{\ln P(D\,|\,\theta)\}$
- Directly maximizing the log-likelihood over ϑ is hard
- We can use EM approach for finding the maximum likelihood estimation for the parameters ϑ
- **Expectation step**: Given current estimates for ϑ, compute the cluster posterior probability $P(C_i|x_j)$ via Bayes theorem: $$P(C_i\,|\,x_j) = \frac{f_i(x_j)\cdot P(C_i)}{\sum_{a=1}^{k} f_a(x_j)\cdot P(C_a)}$$
- **Maximization step:**
  - Using weight $P(C_i|x_j)$ re-estimate ϑ, i.e., re-estimate μ_i, Σ_i and $P(C_i)$ for each cluster C_i

Session 4: The Expectation-Maximization (EM) Algorithm (Univariate)

The Expectation-Maximization Framework for K-Means and EM

- The *k*-means algorithm has two steps at each iteration
  - **Expectation Step** (E-step): Given the current cluster centers, each object is assigned to the cluster whose center is closest to the object. An object is *expected to belong to the closest cluster.*
  - **Maximization Step** (M-step): Given the cluster assignment, the algorithm *adjusts the center* for each cluster so that *the sum of distance* from the objects assigned to this cluster and the new center is minimized
- **The (EM) algorithm:** A framework to approach maximum likelihood or maximum a posteriori estimates of parameters in statistical models
  - **E-step** assigns objects to clusters according to the current parameters of probabilistic clusters
  - **M-step** finds the new clustering or parameters that minimize the sum of squared errors (SSE) or the expected likelihood

Expectation-Maximization for One Dimension (Univariate)

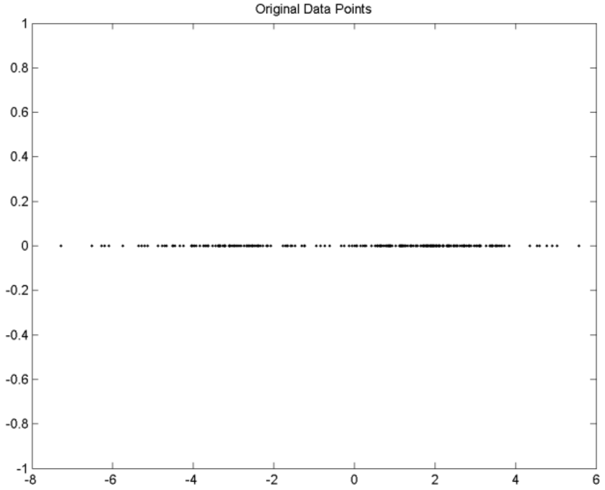
- Consider a dataset ***D*** consisting of a single attribute X, where each point x_i (i = 1, …, n) is a random sample from X
- For the mixture model, we use univariate normals for each cluster

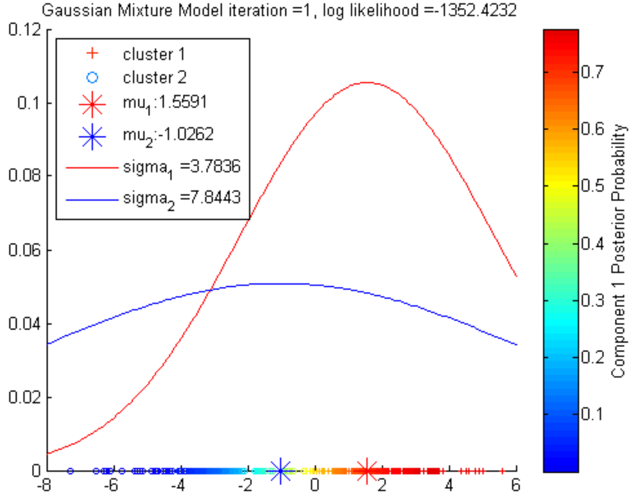
$$f_i(x) = f(x \mid \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\}$$

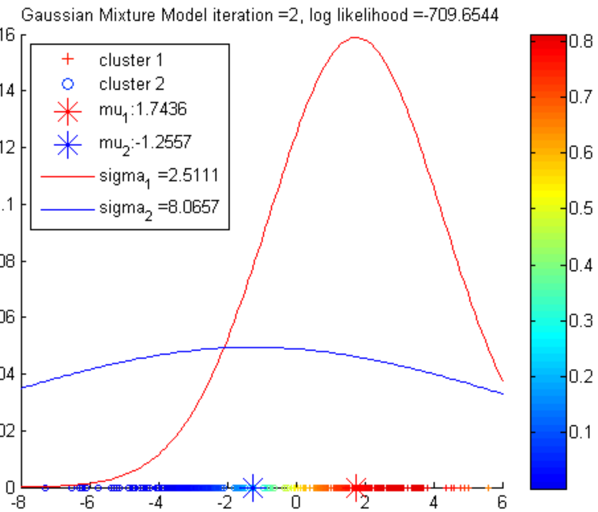
- Initialization:
  - For each cluster C_i, with i = 1, …, k, randomly initialize cluster parameters:
    - μ_i is selected uniformly at random; $\sigma_i^2 = 1$; $P(C_i) = 1/k$ (each cluster has equal prob.)
- Expectation step:
  - Calculate the posterior probability $P(C_i|x_j)$: $P(C_i \mid x_j) = \frac{f(x_j \mid \mu_i, \sigma_i^2) \cdot P(C_i)}{\sum_{a=1}^{k} f(x_j \mid \mu_a, \sigma_a^2) \cdot P(C_a)}$
- Maximization step:
  - Compute the maximum likelihood estimates of the cluster parameters by re-estimating μ_i, σ_i^2 and $P(C_i)$ for each cluster C_i

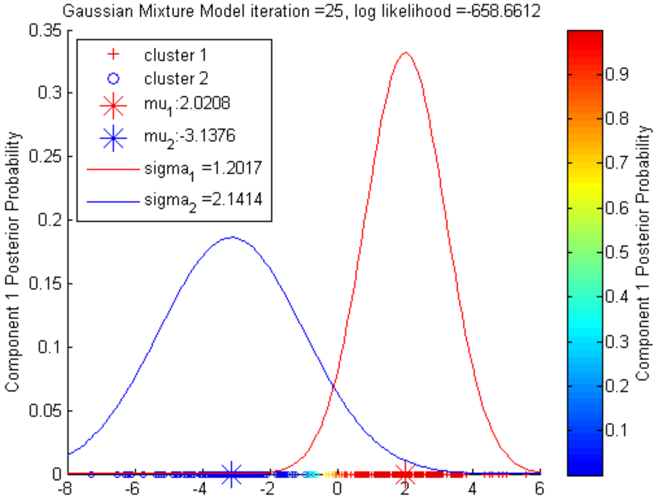
Demonstration of the EM Execution for One Dimensional Data

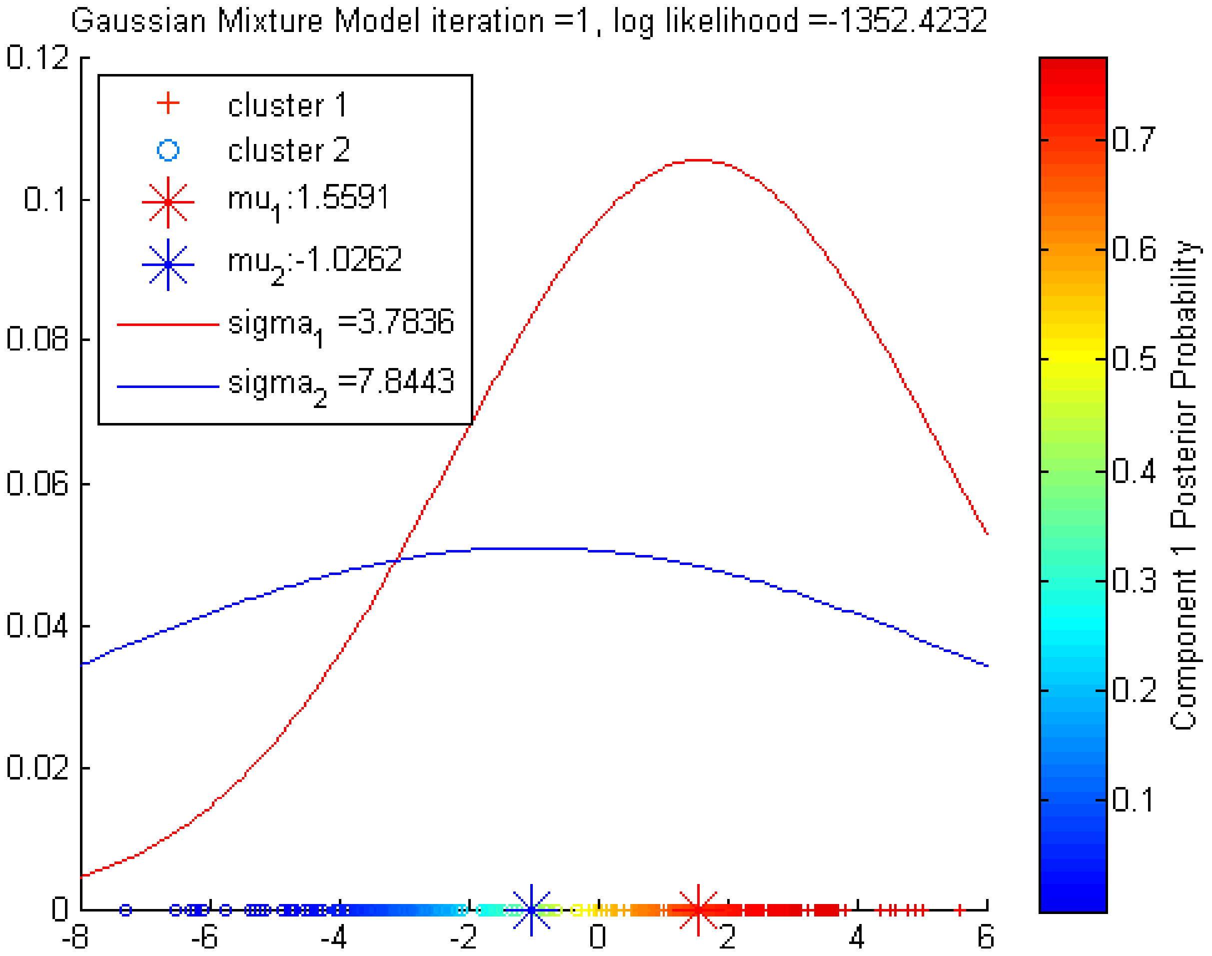
- The execution of the EM Algorithm for Univariate (Single Dimension)











Session 5: The Expectation-Maximization (EM) Algorithm (Multivariate)

The Expectation Maximization Algorithm (Multivariate)

- Randomly initialize $\mu_1, \cdots, \mu_k$; $\Sigma_i \leftarrow \mathbf{I}\ \forall i = 1, \ldots, k$; $P(C_i) \leftarrow 1/k\ \forall i = 1, \ldots, k$ // **Initialization**
- **Repeat**

// **Expectation Step:** Assigns objects to clusters according to the current parameters of probabilistic clusters

- for $i = 1, \ldots, k$ and $j = 1, \ldots, n$ do

$$w_{ij} \leftarrow \frac{f(x_j \mid \mu_i, \Sigma_i) \cdot P(C_i)}{\sum_{a=1}^{k} f(x_j \mid \mu_a, \Sigma_a) \cdot P(C_a)}$$

// Calculate the posterior probability $P(C_i|x_j)$

// **Maximization Step:** Finds the new clustering or parameters that minimize SSE or the expected likelihood

- for $i = 1, \ldots, k$ do

$$\mu_i \leftarrow \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathrm{x}_j}{\sum_{j=1}^{n} w_{ij}} \qquad \Sigma_i \leftarrow \frac{\sum_{j=1}^{n} w_{ij} (\mathrm{x}_j - \mu_i)(\mathrm{x}_j - \mu_i)^T}{\sum_{j=1}^{n} w_{ij}} \qquad P(C_i) \leftarrow \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

// re-estimate mean // re-estimate covariance matrix // re-estimate priors

- **Until** the sum of the changes of the means across two iterations is no greater than threshold ϵ

Demonstration of the EM Execution for Two Dimensional Data

- The execution of the EM algorithm for a two-dimensional data set

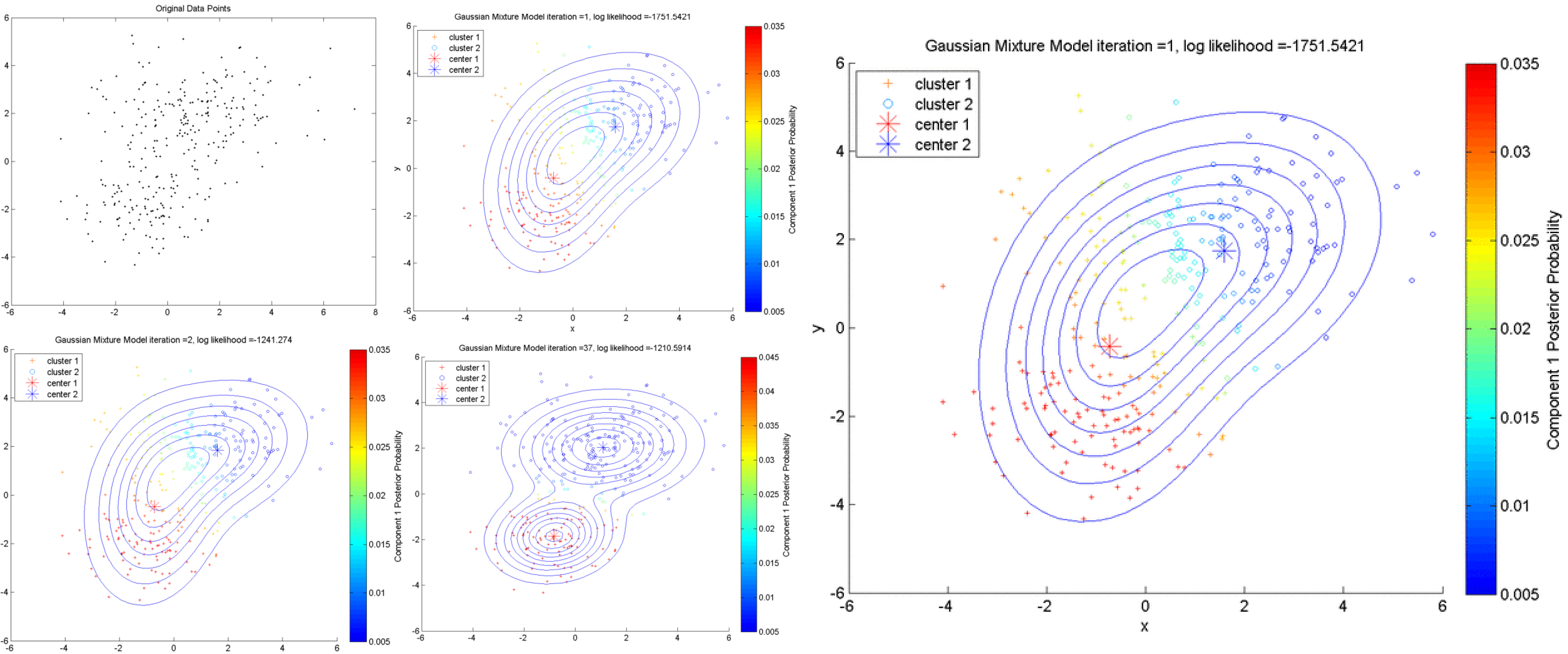
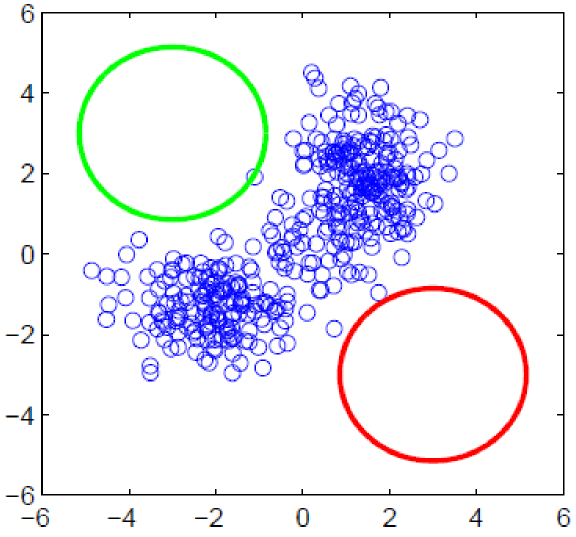
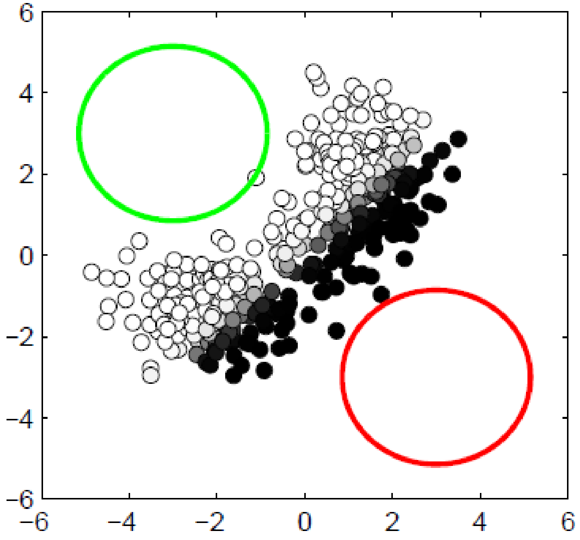


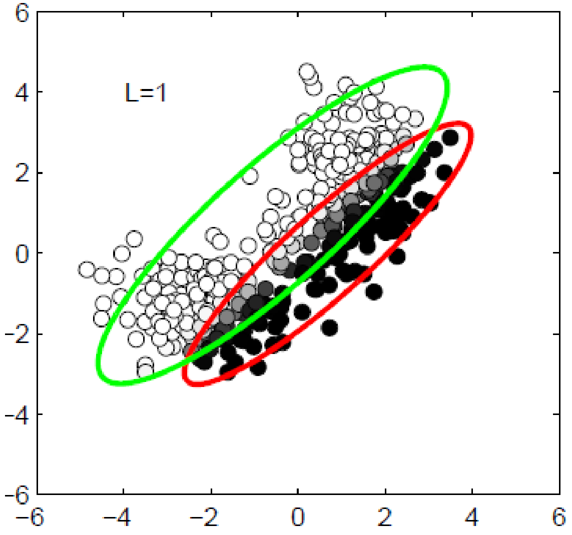
Illustration of the EM Algorithm for Two Gaussian Components



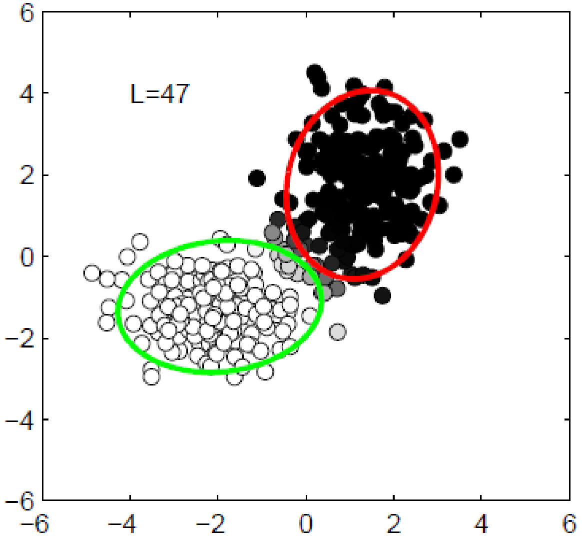
A randomly generated data set (in blue circles). A random initialization of the mixture model: The two Gaussian components are shown as green and red circles



After the initial E-step: Each data point is depicted using a proportion of white ink and black ink according to the posterior probability generated by the corresponding component



After the first M-step: The means and covariances of both components have changed



The results after 47 cycles of EM: Close to convergence

Session 6: Analysis of the Mixture Model Methods

K-Means Can Be Considered as a Special Case of EM

- K-means can be considered as a special case of the EM algorithm, where

$$P(\mathrm{x}_{\mathrm{j}} \mid C_i) = \begin{cases} 1 & \text{if } C_i = \arg\min_{C_a}\{\| x_j - \mu_a \|^2\} \\ 0 & \text{otherwise} \end{cases}$$

$$P(C_i \mid \mathrm{x}_{\mathrm{j}}) = \begin{cases} 1 & \text{if } \mathrm{x}_{\mathrm{j}} \in C_i, \text{i.e., } C = \arg\min_{C_a}\{\| x_j - \mu_a \|^2\} \\ 0 & \text{otherwise} \end{cases}$$

- K-means can be viewed as a hard-EM: In the E-step, we take the local minimum instead of a distribution
- The Gaussian Mixture Model (GMM) is the soft version of k-means
  - We calculate the distribution instead of the most likely one in the E-step and use the weighted sum to compute the new centers in the M-step
  - GMM introduces variance to learning, whereas clusters in k-means have the same variance

Initialization and Speed-Up of Expectation-Maximization

- Hard vs. soft clustering assignments
  - K-Means: Hard assignment clustering—Each point can belong to only one cluster
  - Probabilistic clustering: Soft assignment of points to clusters—Each point has a probability of belonging to each cluster
- Compared with K-means algorithm, the EM algorithm for Gaussian mixture model (GMM) takes many more iterations to reach convergence
- To find a suitable initialization and speed up the convergence for a GMM:
  - First run the *K*-means algorithm, and then choose the means and covariances of the clusters and the fractions of data points assigned to the respective clusters for initializing μ_k, Σ_k and P(C_i), respectively
- A Gaussian component collapses onto a particular data point (called: singularity)
  - When detecting a Gaussian component is collapsing, reset its mean and covariance, and then continue with the optimization

Strengths and Weaknesses of Mixture Models

- Strengths
  - Mixture models are more general than partitioning and fuzzy clustering
  - Clusters can be characterized by a small number of parameters
  - The results may satisfy the statistical assumptions of the generative models
- Weaknesses
  - Converge to local optimal (overcome: run multiple times with random initialization)
  - Computationally expensive if the number of distributions is large or the data set contains very few observed data points
  - Need large data sets
  - Hard to estimate the number of clusters

Session 7: Summary

Summary: Probabilistic Model-Based Clustering Methods

- Basic Concepts of Probabilistic Model-Based Clustering
- Mixture Models for Cluster Analysis
- Gaussian Mixture Models
- The Expectation-Maximization (EM) Algorithm (Univariate)
- The Expectation-Maximization (EM) Algorithm (Multivariate)
- Analysis of the Mixture Model Methods
- Summary

Recommended Readings

- A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society.* 1977
- G. J. McLachlan and K. E. Bkasford. *Mixture Models: Inference and Applications to Clustering*. John Wiley & Sons, 1988
- K. Burnham and D. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Verlag, 2002
- C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006
- M. J. Zaki and W. Meira, Jr.. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014
- H. Deng and J. Han, *Probabilistic Models for Clustering*, in (Chapter 3) C. Aggarwal and C. K. Reddy (eds.), *Data Clustering: Algorithms and Applications*. CRC Press, 2014