Mixture Models

Jia Li

Department of Statistics
The Pennsylvania State University

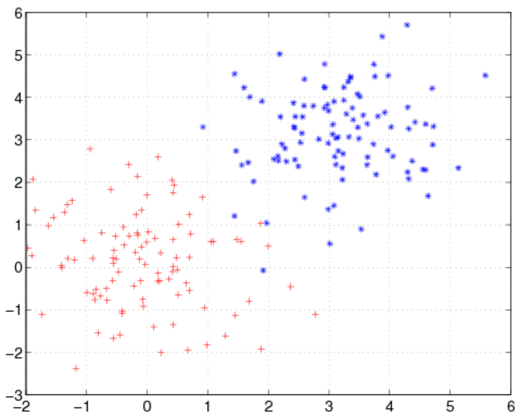
Email: jiali@stat.psu.edu
http://www.stat.psu.edu/∼jiali

Clustering by Mixture Models

- General background on clustering
- Example method: k-means
- Mixture model based clustering
- Model estimation

Clustering

- A basic tool in data mining/pattern recognition:
    - Divide a set of data into groups.
    - Samples in one cluster are close and clusters are far apart.



- Motivations:
  - Discover classes of data in an unsupervised way (unsupervised learning).
  - Efficient representation of data: fast retrieval, data complexity reduction.
  - Various engineering purposes: tightly linked with pattern recognition.

Approaches to Clustering

- Represent samples by feature vectors.
- Define a distance measure to assess the closeness between data.
- "Closeness" can be measured in many ways.
  - Define distance based on various norms.
  - For gene expression levels in a set of micro-array data, "closeness" between genes may be measured by the Euclidean distance between the gene profile vectors, or by correlation.

- Approaches:
  - Define an objective function to assess the quality of clustering and optimize the objective function (purely computational).
  - Clustering can be performed based merely on pair-wise distances. How each sample is represented does not come into the picture.
  - Statistical model based clustering.

K-means

- Assume there are M clusters with centroids
$$\mathcal{Z} = \{z_1, z_2, ..., z_M\} \ .$$
- Each training sample is assigned to one of the clusters. Denote the assignment function by $\eta(\cdot)$. Then $\eta(i) = j$ means the ith training sample is assigned to the jth cluster.
- Goal: minimize the total mean squared error between the training samples and their representative cluster centroids, that is, the trace of the pooled within cluster covariance matrix.
$$\arg\min_{\mathcal{Z},\eta} \sum_{i=1}^{N} \parallel x_i - z_{\eta(i)} \parallel^2$$

- Denote the objective function by

$$L(\mathcal{Z}, \eta) = \sum_{i=1}^{N} \parallel x_i - z_{\eta(i)} \parallel^2 \ .$$

- Intuition: training samples are tightly clustered around the centroids. Hence, the centroids serve as a compact representation for the training data.

Necessary Conditions

- If $\mathcal{Z}$ is fixed, the optimal assignment function $\eta(\cdot)$ should follow the nearest neighbor rule, that is,

$$\eta(i) = \arg\min_{j \in \{1,2,\ldots,M\}} \parallel x_i - z_j \parallel \ .$$

- If $\eta(\cdot)$ is fixed, the cluster centroid z_j should be the average of all the samples assigned to the jth cluster:

$$z_j = \frac{\sum_{i:\eta(i)=j} x_i}{N_j} \ .$$

N_j is the number of samples assigned to cluster j.

The Algorithm

- Based on the necessary conditions, the k-means algorithm alternates the two steps:
  - For a fixed set of centroids, optimize $\eta(\cdot)$ by assigning each sample to its closest centroid using Euclidean distance.
  - Update the centroids by computing the average of all the samples assigned to it.
- The algorithm converges since after each iteration, the objective function decreases (non-increasing).
- Usually converges fast.
- Stopping criterion: the ratio between the decrease and the objective function is below a threshold.

Mixture Model-based Clustering

- Each cluster is mathematically represented by a parametric distribution. Examples: Gaussian (continuous), Poisson (discrete).
- The entire data set is modeled by a mixture of these distributions.
- An individual distribution used to model a specific cluster is often referred to as a component distribution.

- Suppose there are K components (clusters). Each component is a Gaussian distribution parameterized by μ_k, Σ_k. Denote the data by X, $X \in \mathcal{R}^d$. The density of component k is

$$\begin{aligned} f_k(x) &= \phi(x \mid \mu_k, \Sigma_k) \\ &= \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(\frac{-(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)}{2}\right) . \end{aligned}$$

- The prior probability (weight) of component k is a_k. The mixture density is:

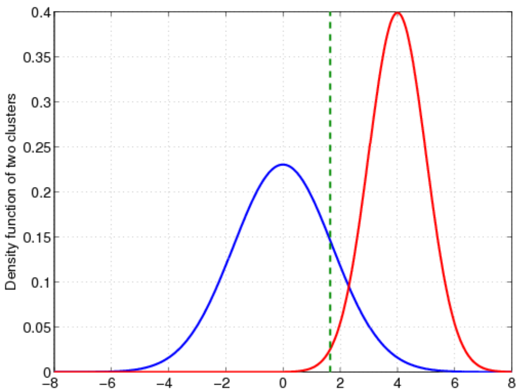
$$f(x) = \sum_{k=1}^{K} a_k f_k(x) = \sum_{k=1}^{K} a_k \phi(x \mid \mu_k, \Sigma_k) .$$

Advantages

- A mixture model with high likelihood tends to have the following traits:
  - Component distributions have high "peaks" (data in one cluster are tight)
  - The mixture model "covers" the data well (dominant patterns in the data are captured by component distributions).

Advantages

- Well-studied statistical inference techniques available.
- Flexibility in choosing the component distributions.
- Obtain a density estimation for each cluster.
- A “soft” classification is available.



EM Algorithm

- The parameters are estimated by the maximum likelihood (ML) criterion using the EM algorithm.
- A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal Royal Statistics Society*, vol. 39, no. 1, pp. 1-21, 1977.
- The EM algorithm provides an iterative computation of maximum likelihood estimation when the observed data are incomplete.

- Incompleteness can be conceptual.
  - We need to estimate the distribution of X, in sample space $\mathcal{X}$, but we can only observe X indirectly through Y, in sample space $\mathcal{Y}$.
  - In many cases, there is a mapping $x \rightarrow y(x)$ from $\mathcal{X}$ to $\mathcal{Y}$, and x is only known to lie in a subset of $\mathcal{X}$, denoted by $\mathcal{X}(y)$, which is determined by the equation $y = y(x)$.
  - The distribution of X is parameterized by a family of distributions $f(x \mid \theta)$, with parameters $\theta \in \Omega$, on x. The distribution of y, $g(y \mid \theta)$ is

$$g(y \mid \theta) = \int_{\mathcal{X}(y)} f(\mathbf{x} \mid \theta) dx \ .$$

- The EM algorithm aims at finding a θ that maximizes $g(y \mid \theta)$ given an observed y.
- Introduce the function

$$Q(\theta' \mid \theta) = E(\log f(x \mid \theta') \mid y, \theta) \ ,$$

that is, the expected value of $\log f(x \mid \theta')$ according to the conditional distribution of x given y and parameter θ. The expectation is assumed to exist for all pairs (θ', θ). In particular, it is assumed that $f(x \mid \theta) > 0$ for $\theta \in \Omega$.

- **EM Iteration**:
  - E-step: Compute $Q(\theta \mid \theta^{(p)})$.
  - M-step: Choose $\theta^{(p+1)}$ to be a value of $\theta \in \Omega$ that maximizes $Q(\theta \mid \theta^{(p)})$.

EM for the Mixture of Normals

- Observed data (incomplete): $\{x_1, x_2, ..., x_n\}$, where n is the sample size. Denote all the samples collectively by $\mathbf{x}$.
- Complete data: $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where y_i is the cluster (component) identity of sample x_i.
- The collection of parameters, θ, includes: a_k, μ_k, Σ_k, $k = 1, 2, ..., K$.
- The likelihood function is:

$$L(\mathbf{x}|\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} a_k \phi(x_i|\mu_k, \Sigma_k) \right) .$$

- $L(\mathbf{x}|\theta)$ is the objective function of the EM algorithm (maximize). Numerical difficulty comes from the sum inside the log.

- The Q function is:

$$\begin{aligned} Q(\theta'|\theta) &= E\left[\log \prod_{i=1}^{n} a'_{y_i}\phi(x_i \mid \mu'_{y_i}, \Sigma'_{y_i}) \mid \mathbf{x}, \theta\right] \\ &= E\left[\sum_{i=1}^{n} \left(\log(a'_{y_i}) + \log \phi(x_i \mid \mu'_{y_i}, \Sigma'_{y_i}\right) \mid \mathbf{x}, \theta\right] \\ &= \sum_{i=1}^{n} E\left[\log(a'_{y_i}) + \log \phi(x_i \mid \mu'_{y_i}, \Sigma'_{y_i}) \mid x_i, \theta\right] . \end{aligned}$$

The last equality comes from the fact the samples are independent.

- Note that when x_i is given, only y_i is random in the complete data (x_i, y_i). Also y_i only takes a finite number of values, i.e, cluster identities 1 to K. The distribution of Y given $X = x_i$ is the posterior probability of Y given X.
- Denote the posterior probabilities of $Y = k$, $k = 1, ..., K$ given x_i by $p_{i,k}$. By the Bayes formula, the posterior probabilities are:

$$p_{i,k} \propto a_k \phi(x_i \mid \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} p_{i,k} = 1 .$$

- Then each summand in $Q(\theta'|\theta)$ is

$$
\begin{aligned}
& E\left[\log(a'_{y_i}) + \log \phi(x_i \mid \mu'_{y_i}, \Sigma'_{y_i}) \mid x_i, \theta\right] \\
= & \sum_{k=1}^{K} p_{i,k} \log a'_k + \sum_{k=1}^{K} p_{i,k} \log \phi(x_i \mid \mu'_k, \Sigma'_k) \ .
\end{aligned}
$$

- Note that we cannot see the direct effect of θ in the above equation, but $p_{i,k}$ are computed using θ, i.e, the current parameters. θ' includes the updated parameters.

- We then have:

$$
\begin{aligned}
Q(\theta'|\theta) \;&=\; \sum_{i=1}^{n}\sum_{k=1}^{K} p_{i,k}\log a'_k + \\
&\sum_{i=1}^{n}\sum_{k=1}^{K} p_{i,k}\log \phi(x_i \mid \mu'_k, \Sigma'_k)
\end{aligned}
$$

- Note that the prior probabilities a'_k and the parameters of the Gaussian components μ'_k, Σ'_k can be optimized separately.

- The a'_k's subject to $\sum_{k=1}^K a'_k = 1$. Basic optimization theories show that a'_k are optimized by

$$a'_k = \frac{\sum_{i=1}^n p_{i,k}}{n} \,.$$

- The optimization of μ_k and Σ_k is simply a maximum likelihood estimation of the parameters using samples x_i with weights $p_{i,k}$. Basic optimization techniques also lead to

$$\mu'_k = \frac{\sum_{i=1}^n p_{i,k} x_i}{\sum_{i=1}^n p_{i,k}}$$

$$\Sigma'_k = \frac{\sum_{i=1}^n p_{i,k} (x_i - \mu'_k)(x_i - \mu'_k)^t}{\sum_{i=1}^n p_{i,k}}$$

- After every iteration, the likelihood function L is guaranteed to increase (may not strictly).
- We have derived the EM algorithm for a mixture of Gaussians.

EM Algorithm for the Mixture of Gaussians

Parameters estimated at the pth iteration are marked by a superscript (p).

1. Initialize parameters
2. E-step: Compute the posterior probabilities for all $i = 1, ..., n$, $k = 1, ..., K$.

$$p_{i,k} = \frac{a_k^{(p)} \phi(x_i \mid \mu_k^{(p)}, \Sigma_k^{(p)})}{\sum_{k=1}^{K} a_k^{(p)} \phi(x_i \mid \mu_k^{(p)}, \Sigma_k^{(p)})} .$$

3. M-step:

$$a_k^{(p+1)} = \frac{\sum_{i=1}^{n} p_{i,k}}{n} , \quad \mu_k^{(p+1)} = \frac{\sum_{i=1}^{n} p_{i,k} x_i}{\sum_{i=1}^{n} p_{i,k}}$$

$$\Sigma_k^{(p+1)} = \frac{\sum_{i=1}^{n} p_{i,k} (x_i - \mu_k^{(p+1)})(x_i - \mu_k^{(p+1)})^t}{\sum_{i=1}^{n} p_{i,k}}$$

4. Repeat step 2 and 3 until converge.

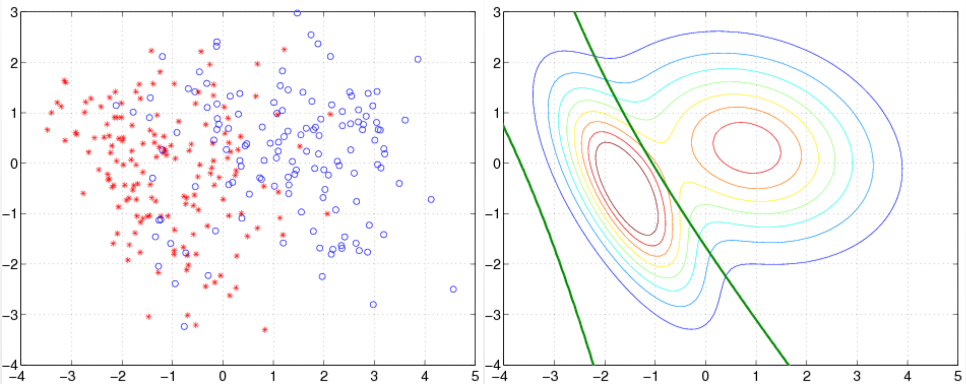
- Comment: for mixtures of other distributions, the EM algorithm is very similar. The E-step involves computing the posterior probabilities. Only the particular distribution ϕ needs to be changed. The M-step always involves parameter optimization. Formulas differ according to distributions.

Computation Issues

- If a different Σ_k is allowed for each component, the likelihood function is not bounded. Global optimum is meaningless. (Don't overdo it!)
- How to initialize? Example:
  - Apply k-means first.
  - Initialize μ_k and Σ_k using all the samples classified to cluster k.
  - Initialize a_k by the proportion of data assigned to cluster k by k-means.
- In practice, we may want to reduce model complexity by putting constraints on the parameters. For instance, assume equal priors, identical covariance matrices for all the components.

Examples

- The heart disease data set is taken from the UCI machine learning database repository.
- There are 297 cases (samples) in the data set, of which 137 have heart diseases. Each sample contains 13 quantitative variables, including cholesterol, max heart rate, etc.
- We remove the mean of each variable and normalize it to yield unit variance.
- data are projected onto the plane spanned by the two most dominant principal component directions.
- A two-component Gaussian mixture is fit.



The heart disease data set and the estimated cluster densities. Left: The scatter plot of the data. Right: The contour plot of the pdf estimated using a single-layer mixture of two normals. The thick lines are the boundaries between the two clusters based on the estimated pdfs of individual clusters.

Classification Likelihood

- The likelihood:

$$L(\mathbf{x}|\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} a_k \phi(x_i|\mu_k, \Sigma_k) \right)$$

  maximized by the EM algorithm is sometimes called *mixture likelihood*.

- Maximization can also be applied to the *classification likelihood*. Denote the collection of cluster identities of all the samples by $\mathbf{y} = \{y_1, ..., y_n\}$.

$$\tilde{L}(\mathbf{x}|\theta, \mathbf{y}) = \sum_{i=1}^{n} \log \left(a_{y_i} \phi(x_i|\mu_{y_i}, \Sigma_{y_i}) \right)$$

- The cluster identities y_i, $i = 1, ..., n$ are treated as parameters together with θ and are part of the estimation.
- To maximize $\tilde{L}$, EM algorithm can be modified to yield an ascending algorithm. This modified version is called *Classification EM* (CEM).

Classification EM

A classification step is inserted between the E-step and the M-step.

- Initialize parameters
- E-step: Compute the posterior probabilities for all $i = 1, ..., n$, $k = 1, ..., K$,

$$p_{i,k} = \frac{a_k^{(p)} \phi(x_i \mid \mu_k^{(p)}, \Sigma_k^{(p)})}{\sum_{k=1}^K a_k^{(p)} \phi(x_i \mid \mu_k^{(p)}, \Sigma_k^{(p)})} .$$

- Classification:

$$y_i^{(p+1)} = \arg\max_k p_{i,k} .$$

Or equivalently, let $\hat{p}_{i,k'} = 1$ if $k' = \arg\max_k p_{i,k}$ and 0 otherwise.

- M-step:

$$a_k^{(p+1)} = \frac{\sum_{i=1}^n \hat{p}_{i,k}}{n} = \frac{\sum_{i=1}^n I(y_i^{(p+1)} = k)}{n}$$

$$\mu_k^{(p+1)} = \frac{\sum_{i=1}^n \hat{p}_{i,k} x_i}{\sum_{i=1}^n \hat{p}_{i,k}} = \frac{\sum_{i=1}^n I(y_i^{(p+1)} = k) x_i}{\sum_{i=1}^n I(y_i^{(p+1)} = k)}$$

$$\begin{aligned} \Sigma_k^{(p+1)} &= \frac{\sum_{i=1}^n \hat{p}_{i,k} (x_i - \mu_k^{(p+1)})(x_i - \mu_k^{(p+1)})^t}{\sum_{i=1}^n \hat{p}_{i,k}} \\ &= \frac{\sum_{i=1}^n I(y_i^{(p+1)} = k)(x_i - \mu_k^{(p+1)})(x_i - \mu_k^{(p+1)})^t}{\sum_{i=1}^n I(y_i^{(p+1)} = k)} \end{aligned}$$

- Repeat the above three steps until converge.

Comment:

- CEM tends to underestimate the variances. It usually converges much faster than EM. For the purpose of clustering, it is generally believed that it performs similarly as EM.
- If we assume equal priors a_k and also the covariance matrices Σ_k are identical and are a scalar matrix, CEM is exactly k-means. (Exercise)