Machine Learning

Lecture 15: Clustering

Sibei Yang

SIST

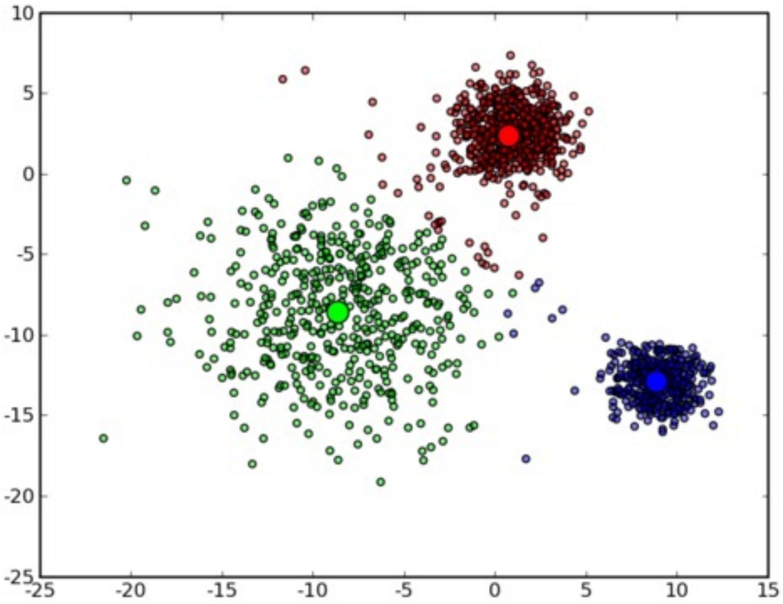
Email: yangsb@shanghaitech.edu.cn

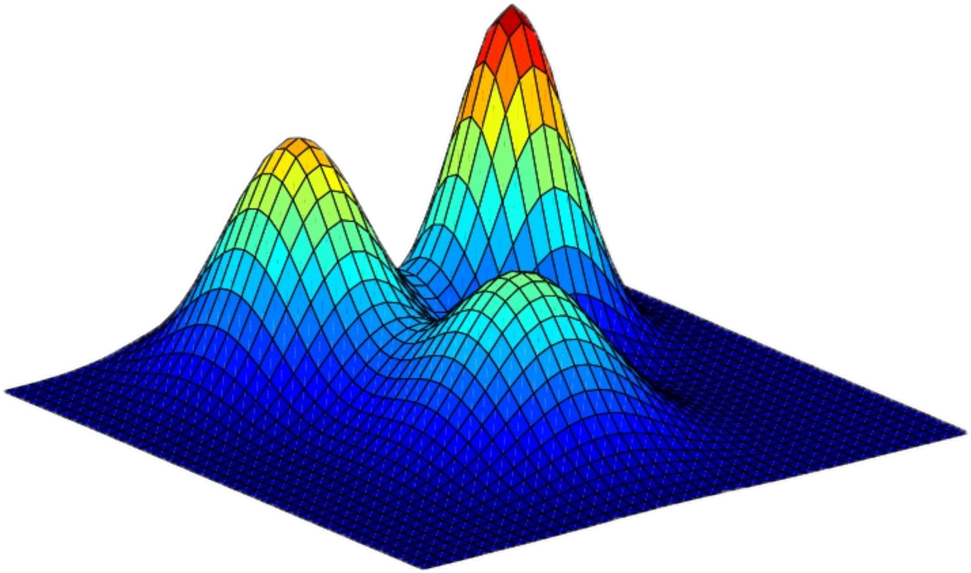
Algorithms

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: **k-means**, **k-medoids**
- Model-based:
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - Typical methods: **GMM**
- Dimensionality reduction approach
  - First dimensionality reduction, then clustering
  - Typical methods: **Spectral clustering**, Ncut

Gaussian Mixture Model

- Gaussian Mixture Model (GMM) is one of the most popular clustering methods which can be viewed as a linear combination of different Gaussian components.





Gaussian Mixture Model

- Multivariate Gaussian
  - $\boldsymbol{\mu}$: mean of the distribution
  - Σ: covariance of the distribution

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

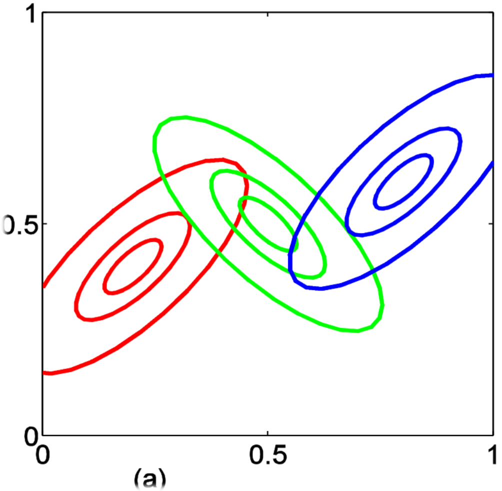
  - Maximum likelihood estimation

$$\begin{cases} \widehat{\boldsymbol{\mu}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \boldsymbol{x_i} \\ \hat{\Sigma} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (\boldsymbol{x_i}-\widehat{\boldsymbol{\mu}})(\boldsymbol{x_i}-\widehat{\boldsymbol{\mu}})^{\boldsymbol{T}} \end{cases}$$

Gaussian Mixture Model

- Linear combination of Gaussians
  - Assumption: K Gaussians, each has a contribution of π_k to the data points

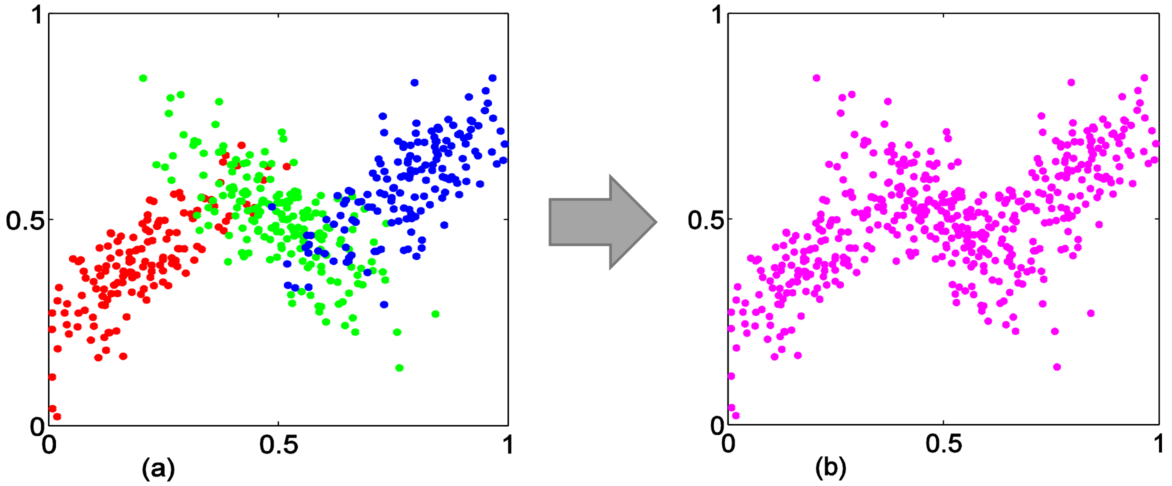
$$\begin{cases} p(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x};\boldsymbol{\theta}_k) \\ \boldsymbol{\Theta} = \{\pi_1,\cdots,\pi_K,\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_K\}, \sum_{k=1}^{k} \pi_k = 1, \pi_k \in [0,1] \\ p_k(\boldsymbol{x};\boldsymbol{\theta}_k) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \end{cases}$$



  - Parameters to be estimated: $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$

Gaussian Mixture Model

- The process of generating a data point
  - first pick one of the components with probability π_k
  - then draw a sample $\boldsymbol{x}_i$ from that component distribution
- Each data point is generated by one of k components



Gaussian Mixture Model

- The log-likelihood function:

$$\log \prod_{i=1}^{N} p\left(\boldsymbol{x}^{(i)}; \boldsymbol{\Theta}\right) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)$$

  is difficult to find solutions.

- Using **Expectation Maximization** (EM) algorithm:

- E-step:

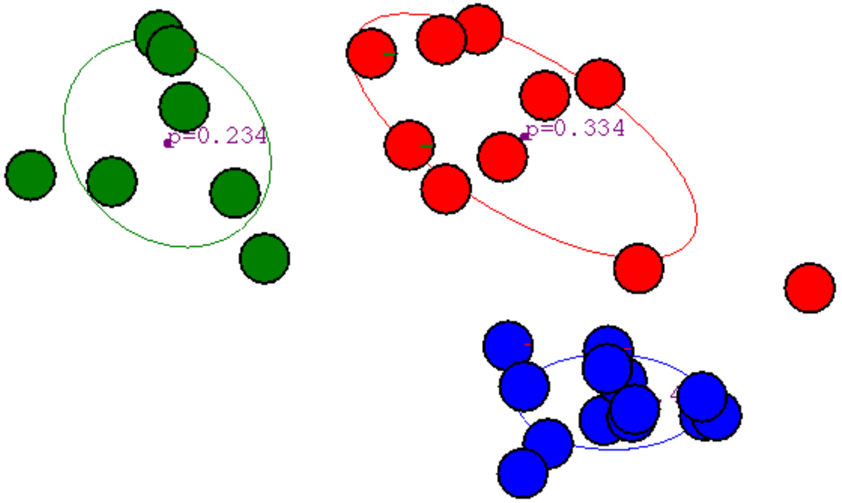
$$
\begin{aligned}
Q^i\left(\boldsymbol{z}_k^{(i)}\right) &= p\left(\boldsymbol{z}_k^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\Theta}\right) \\
&= \frac{\pi_k \mathcal{N}\left(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}
\end{aligned}
$$

- M-step:
  - Take the derivative of the log likelihood to obtain estimates for π_k, μ_k, Σ_k directly

$$
\begin{aligned}
\pi_k &= \frac{\sum_{i=1}^{M} Q^i\left(\boldsymbol{z}_k^{(i)}\right)}{M} \\
\boldsymbol{\mu}_k &= \frac{\sum_{i=1}^{M} \boldsymbol{x}^{(i)} Q^i\left(\boldsymbol{z}_k^{(i)}\right)}{\sum_{i=1}^{M} Q^i\left(\boldsymbol{z}_k^{(i)}\right)} \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{i=1}^{M}\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k\right)\left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k\right)^T Q^i\left(\boldsymbol{z}_k^{(i)}\right)}{\sum_{i=1}^{M} Q^i\left(\boldsymbol{z}_k^{(i)}\right)}
\end{aligned}
$$

- Do the iterations until convergence, then $Q^i\left(\boldsymbol{z}_k^{(i)}\right)$ can be used for clustering

Gaussian Mixture Model: An example



K-Means vs. GMM

- Objective function:
  - Minimize the TSD
- Can be optimized by an EM algorithm.
  - E-step: assign points to clusters.
  - M-step: optimize clusters.
  - Performs hard assignment during E-step.
- Assumes spherical clusters with equal probability of a cluster.

- Objective function
  - Maximize the log-likelihood.
- EM algorithm
  - E-step: Compute posterior probability of membership.
  - M-step: Optimize parameters.
  - Perform soft assignment during E-step.
- Can be used for non-spherical clusters. Can generate clusters with different probabilities.