Introduction to Data Mining

Lecture8 Expectation-Maximization (EM) Algorithm

Jun Huang

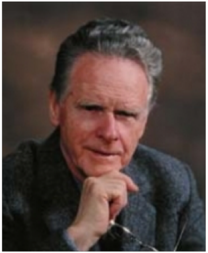
Anhui University of Technology

Spring 2018

huangjun_cs@163.com

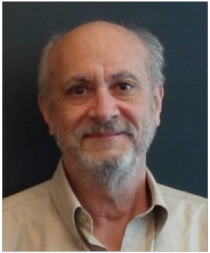
Expectation-Maximization

- EM Algorithm was named and explained by Arthur Dempster, Nan Laird and Donald Rubin in 1977
- It is pointed out in the paper that the method has been proposed many times in special circumstances
- EM is typically used to **compute maximum likelihood estimates given incomplete data samples**.



Arthur P. Dempster

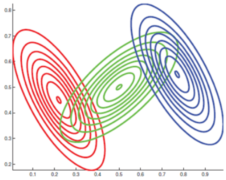
Nan M. Laird



Donald B. Rubin

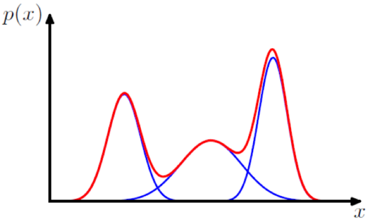
Motivation: Finite Mixture Models

- Convex combination of multiple density functions
- Capable of approximating any arbitrary distribution
- In many applications, their parameters are determined by ML, typically using EM Algorithm
- Widely used in:
  - Data Mining
  - Pattern Recognition
  - Machine Learning
  - Statistical Analysis



Mixture of Gaussians

- **Simple linear superposition of Gaussian components**:
- Gaussian distribution suffer from significant limitations when it comes to modelling real data sets



$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

$$\sum_{k=1}^{K} \pi_k = 1, \; 0 \leq \pi_k \leq 1$$

K-means Clustering

- Data set $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$
  - N observations of D-dimensional Euclidian variable $\mathbf{x}$
- Goal: Partition $\mathcal{D}$ into K clusters
  - Minimize within-cluster distance
  - Maximize between-cluster distance

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

- μ_k : Center of the k-th cluster
  - Mean of the points in the cluster k
- $r_{nk} \in \{0, 1\}$: Binary Indicator Variable
  - If $\mathbf{x}_n$ is assigned to cluster k: $r_{nk} = 1$
  - Else: $r_{nk} = 0$
- Find values for r_{nk} and μ_k minimizing J

K-means Clustering

- Select initial random points k for each cluster Iteratively do two successive optimizations until convergence:
- **E- Find r_{nk} using μ_k:**

$$r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$$

- **M- Update μ_k using r_{nk} calculated in the step E:**

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

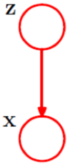
$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \mu_k) = 0 \Rightarrow \mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

Defining the Model: Mixture

- Let's introduce $\mathbf{z}$:
  - $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$
  - one of K representation

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

$$p(z_k = 1) = \pi_k$$

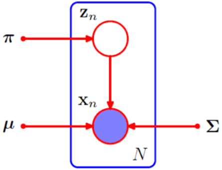


- Now define $p(\mathbf{x}, \mathbf{z})$:
  - $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$
- Reformulate the mixture distribution of $\mathbf{x}$ using $\mathbf{z}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

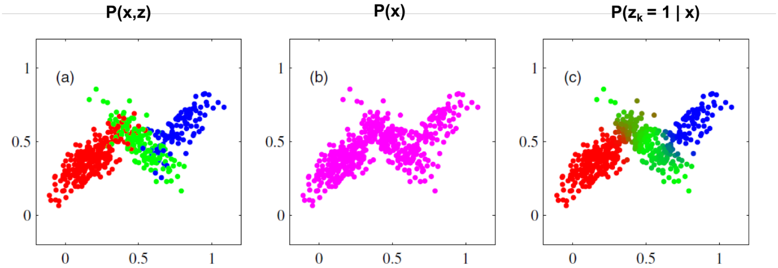
Defining the Model: Posterior

- Derive the posterior probability of $\mathbf{z}$, observing $\mathbf{x}$, in terms of the mixture distribution that we defined:



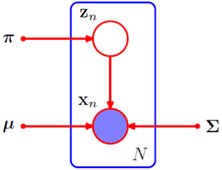
$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)} \\ p(z_k = 1) &= \pi_k\end{aligned}$$

Joint Distribution –Marginal Distribution – Responsibility



Defining the Model: Log Likelihood

- Having a data set of observations $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$
- Matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$
- i.i.d. data points $\mathbf{x}_n$ with corresponding latent points $\mathbf{z}_n$.



$$p(\mathbf{X}|\pi, \mu, \Sigma) = \prod_{n=1}^{N} \left\{ \sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right\}$$

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right\}$$

EM: Algorithm

1. In the **expectation** step we use the current values for the parameters to **valuate the posterior probabilities**
2. In the **maximization** step we use these posterior probabilities to **re-estimate** the **model parameters**, such as the means, covariance matrix and mixing coefficients.

EM for Gaussian Mixtures: Mean

- Set the derivative of the likelihood function with respect to μ_k vector to zero:

$$\ln p(\mathbf{X}|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k,\Sigma_k) \right\}$$

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_j,\Sigma_j)}$$

$$-\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_j,\Sigma_j)} \Sigma_k(\mathbf{x}_n - \mu_k) = 0$$

- By rearranging, we can obtain:$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$, where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

EM for Gaussian Mixtures: Covariance Matrix

- Set the derivative of the likelihood function with respect to Σ_k to zero:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

- Note:Single Gaussian fitted to the data set, but each data point weighted by the corresponding posterior probability and denominated by the effective number of points associated with that component

EM for Gaussian Mixtures: Mixing Coefficient

- Set the derivative of likelihood function with respect to π_k:
  - Constraint: Mixing coefficients need to sum to one.
  - Use a Lagrange multiplier and maximize:

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

- Take the derivative, multiply both sides by π and sum over k:

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)} + \lambda = 0$$

- Then, we can obtain:

$$\pi_k = \frac{N_k}{N}$$

EM for Gaussian Mixtures

1. Initialize the means μ_k, covariances Σ_k and mixing coefficients π_k, and evaluate the initial value of the log likelihood
2. **E-step**: Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}$$

3. **M-step**: Re-estimate the parameters using the current responsibilities

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

4. Evaluate the log likelihood $\ln p(\mathbf{X}|\mu, \Sigma, \pi)$, and check for convergence of either the parameters or the log likelihood

EM: A Broader View

- Finding maximum likelihood solutions for models with latent variables:

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$

- **Complete data** is in the form of X, Z
- Our observed data X is **incomplete**, **we can not directly use maximum likelihood**
- Because we can not use the complete-data likelihood, we **instead use its expected value under the posterior distribution of the latent variable**: **E step**

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

- Then we maximize this expectation function: **M step**

$$\theta^{\text{old}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}})$$

The General EM Algorithm

1. Choose an initial setting for the parameters θ^{old}
2. **E-step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$
3. **M-step** Evaluate θ^{new} given by

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}})$$

where

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$$

and return to step 2

Question 1

- **What is the difference between EM algorithm and ML (Maximum Likelihood) estimation?**
  - EM algorithm tries to find a ML estimation for the parameters of a model with latent variables.
  - In each iteration of the algorithm, latent variables are calculated and being used to maximize the likelihood, they are like 'side effects' of the maximization process.
  - EM is not guaranteed to converge to the global maximum, but it is guaranteed to converge to a maximum and improve the likelihood of the model at every step.

Question 2

- **What is a mixture model and what is the benefit of using it? What is the mathematical expression for a Gaussian mixture?**
  - Finite mixture models are convex combinations or weighted sums of multiple density functions.
  - With enough components, they are capable of approximating any arbitrary distribution.
  - PDF of a Gaussian mixture is in the form of

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

Question 3

- **How can we make use of EM to solve a clustering problem?**
  - EM is used for maximizing the likelihood for models with incomplete/ latent variables.
  - When we consider a clustering problem, we can model the problem by introducing cluster labels as latent variables:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$