Expectation-Maximization and Gaussian Mixture Models

CAP6676: Knowledge Representation
Instructor: Guo-Jun Qi
Hard VS. Soft Clustering

- K-Means
  - Hard assignment of examples into different groups
  - An algorithm alternating between assignment and update of cluster means.

- Generalization
  - Soft assignment: allowing each example to be assigned to multiple clusters
  - Alternating algorithm to EM algorithm
What is EM?

- An algorithm alternating between assignment and estimation

- A *parameter estimation* method: it falls into the general framework of *maximum-likelihood estimation* (MLE).

- The general form was given in (Dempster, Laird, and Rubin, 1977), although essence of the algorithm appeared previously in various forms.
Outline

- Maximum Likelihood Estimation (MLE)
- EM: basic concepts
- EM for Gaussian Mixture Model (GMM)
Outline

- Maximum Likelihood Estimation (MLE)
- EM: basic concepts
- EM for Gaussian Mixture Model (GMM)
What is MLE?

- Given
  - A sample $X = \{X_1, \ldots, X_n\}$
  - A vector of parameters $\theta$

- We define
  - Likelihood of the data: $P(X | \theta)$
  - Log-likelihood of the data: $L(\theta) = \log P(X|\theta)$

- Given $X$, find

$$\theta_{ML} = \arg \max_{\theta \in \Omega} L(\Theta)$$
Often we assume that $X_i$s are independently identically distributed (i.i.d.)

\[
\theta_{ML} = \arg \max_{\theta \in \Theta} L(\Theta) \\
= \arg \max_{\theta \in \Theta} \log P(X | \Theta) \\
= \arg \max_{\theta \in \Theta} \log P(X_1, ..., X_n | \Theta) \\
= \arg \max_{\theta \in \Theta} \log \prod_{i} P(X_i | \Theta) \\
= \arg \max_{\theta \in \Theta} \sum_{i} \log P(X_i | \Theta)
\]

Depending on the form of $p(x|\theta)$, solving optimization problem can be easy or hard.
Example 1:
Fit a line to observed data
Maximum likelihood estimation for the slope of a single line

Data: \((X_n, Y_n), n = 1, \ldots, N\)

Model: \(Y = aX + w\)

where \(w \sim N(0, \sigma^2)\)

Data likelihood for point \(n\): \[ p(Y_n | X_n, a) \sim N(aX_n, \sigma^2) \]

Maximum likelihood estimate:

\[ \hat{a} = \text{argmax}_a \log p(Y_n | X_n, a) = \text{argmax}_a \sum_{n=1}^{N}(Y_n - aX_n)^2 \]

\[ \hat{a} = \frac{\sum_{n} Y_n X_n}{\sum_{n} X_n^2}. \]
Example 2: Fitting two lines to observed data
Fitting two lines: on the one hand…

If we knew which points went with which lines, we’d be back at the single line-fitting problem, twice.
Fitting two lines, on the other hand…

We could figure out the probability that any point came from either line if we just knew the two equations for the two lines.
Expectation Maximization (EM): a solution to chicken-and-egg problems
EM example:

Step 1: randomly guess two lines as initialization.
EM example:

Step 2: Assign points to the closest lines.
EM example:

Step 3: Update the current fitting of lines to the assigned points.
EM example:

Step 4: Re-assign points to the closest lines.
EM example:

Step 5: Re-fit lines to the assigned points.

Converged!
EM Algorithm

EM with hard assignment:
- Randomly guess the initial models
  - The cluster means in K-means and the lines in multi-line fitting problem
- Repeat
  - Hard assignment: assign data points to the corresponding clusters based on fitness
    - Fitness: distance to cluster mean in K-means and distance to the line in multi-line fitting.
  - Re-fitting of the model for each cluster
    - In K-means: update cluster mean
    - In multi-line fitting: update the line
    - Both re-fitting rules can be formulated by maximizing the corresponding likelihood of generating assigned data points
- Until convergence
Outline

- Maximum Likelihood Estimation (MLE)
- EM: basic concepts (soft assignment)
- EM for Gaussian Mixture Model (GMM)
Basic setting in EM

- \( X \) is a set of data points: **observed** data
- \( Y \) is a cluster index denoting the assignment of data point to a cluster
  - \( Y \) is not a deterministic variable as in hard assignment; it is now a random variable for soft assignment.
- EM is a method to find \( \theta_{\text{ML}} \) where
  \[
  \theta_{\text{ML}} = \arg \max_{\theta \in \Omega} L(\Theta) = \arg \max_{\theta \in \Omega} \log P(X | \Theta)
  \]
- Calculating \( P(X | \theta) \) directly is hard.
  - Because we have to marginalize out \( Y \).
- Calculating \( P(X,Y|\theta) \) is much simpler, where \( Y \) is “hidden” data (or “missing” data).
The EM terminology

- $Z = (X, Y)$
  - $Z$: complete data ("augmented data")
  - $X$: observed data ("incomplete" data)
  - $Y$: hidden data ("missing" data)

\[
P(X, Y = k \mid \Theta) = P(X \mid Y = k, \Theta) P(Y = k \mid \Theta) = P(X \mid \theta_k) P_k
\]

Where

\[
P(X \mid \theta_k) \equiv P(X \mid Y = k, \Theta)
\]

is the distribution for cluster $k$, and

\[
P_k = P(Y = k \mid \Theta)
\]
EM Algorithm

EM with soft assignment:

- Randomly guess the initial models $\theta_k$ for K clusters.
- Repeat
  - Expectation (soft assignment): assign each data point $X$ according to the posterior distribution
    \[ P(Y = k | X, \Theta^t) = \frac{P(X | \theta_k^t)P(Y = k)}{\sum_{k'} P(X | \theta_{k'}^t)P(Y = k')} \]
  - Maximization: Re-fitting of the model $\theta_k$ for each cluster
    \[ \theta^{t+1} = \arg\max Q(\Theta, \Theta^t) = E_{P(Y | X, \Theta^t)}P(X, Y | \Theta) = \sum_k P(Y = k | X, \Theta^t)P(X, Y = k | \Theta) \]
    Which is weighted likelihood by the soft assignment.
- Until convergence
Outline

- Maximum Likelihood Estimation (MLE)
- EM: basic concepts
- EM for Gaussian Mixture Model (GMM)
  - Each cluster is modeled by Gaussian distribution
  - All the clusters form a mixture of Gaussian distributions
Gaussian Distribution

\[
f(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp \left( -\frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{2} \right)
\]

- \(d\) = feature dimensions
- \(x\) = random data point
- \(\mu\) = mean value
- \(\Sigma\) = covariance matrix
GMMs – Gaussian Mixture Models

- Suppose we have 1000 data points in 2D space (w,h)
GMMs – Gaussian Mixture Models

- Assume each data point is normally distributed
- Obviously, there are 5 sets of underlying gaussians
GMM

- Model the data distribution by a combination of Gaussian functions
- Given a set of sample points, how to estimate their assignments and the parameters of the Gaussian distribution for each cluster?
EM with GMM

- In the context of GMM
  - X: data points
  - Y: which Gaussian creates which data points
  - Distribution of complete data
    \[
    P(X|\Theta) = \sum_{k=1}^{K} p(X|Y = k, \Theta)p(Y = k|\Theta)
    \]
  - Gaussian distribution for each cluster
    \[
    P(X|Y = k, \Theta) = P(X|\mu_k, \Sigma_k) \left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)
    \]
  - Proportion of examples belonging to each cluster
    \[
    p_k = p(Y = k|\Theta)
    \]
    where \(p_k\)'s must sum up to 1.
Soft assignment

- **E-Step:** Assign each data $X_i$ to a cluster $k$

$$p_{ik}^t = P(Y_i = k|X_i, \Theta) = \frac{p_k^t p(x_i|\mu_k, \Sigma_k^t)}{\sum_{k'} p_{k'}^t p(x_i|\mu_{k'}, \Sigma_{k'}^t)}$$

- **Max-Step:** maximize the weighted likelihood

$$\max Q(\Theta, \Theta^t) = \sum_{i} \sum_{k=1}^{K} p(Y_i = k|X_i, \Theta^t) \log(p(X_i, Y_i = k|\Theta)) = \sum_{i} \sum_{k=1}^{K} p(Y_i = k|X_i, \Theta^t) (\log p_k + \log p(X_i|\mu_k^t, \Sigma_k^t))$$

$$p_{k}^{t+1} = \frac{1}{n} \sum_{i=1}^{n} p_{ik}^t \mu_k^{t+1} = \frac{\sum_{i=1}^{n} p_{ik}^t x_i}{\sum_{i=1}^{n} p_{ik}^t}, \Sigma_k^{t+1} = \frac{\sum_{i=1}^{n} p_{ik}^t (x_i - \mu_k^t)(x_i - \mu_k^t)^T}{\sum_{i=1}^{n} p_{ik}^t}$$
EM for GMM (summary)

- **Objective:**
  Given N data points, find maximum likelihood estimation of $\Theta$:
  \[
  \Theta = \arg \max \limits_\Theta f(x_1, \ldots, x_N | \Theta)
  \]

- **Algorithm:**
  1. Guess initial $\Theta$
  2. Perform E step (expectation)
     - Based on $\Theta$, assign each data point to a specific Gaussian component
  3. Perform M step (maximization)
     - Based on data points clustering, maximize Q function
  4. Repeat 2-3 until convergence (~tens iterations)
EM Example

After first iteration
EM Example

After 2nd iteration
EM Example

After 3rd iteration
EM Example

After 4th iteration
EM Example

After 5th iteration

Copyright © 2001, 2004, Andrew W. Moore
EM Example

After 6th iteration
EM Example

After 20th iteration

Copyright © 2001, 2004, Andrew W. Moore

Clustering with Gaussian Mixtures: Slide 47
EM is more than clustering

- Y need not necessarily be a variable of cluster membership
- It may just be any variable that is unknown to us.
  - Bad pixels in an image
  - Lost labels for an example
- There could be many possible Ys.
- EM is still applicable
  - As long as the posterior of Y is tractable to compute.
Summary

- EM is composed of two basic steps
  - E-Step: Estimate the posterior distribution of unobserved variables based on the current estimate of model parameters
  - M-Step: Update the model parameters by maximizing the expected likelihood weighted by the posterior distribution obtained in E-Step

- Application of EM into Gaussian Mixture Model
The steps of EM algorithm

- E-step: calculate

\[ Q(\theta, \theta^t) = E_{p(y|x, \theta^t)} \left[ \log p(x, y | \theta) \right] \]
\[ = \int \log p(x, y | \theta) p(y | x, \theta^t) dy \]

- M-step: find

\[ \theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^t) \]

See online notes for more derivations and analysis!
How to write the Q function under GMM setting

- Likelihood of a data set is the multiplication of all the sample likelihood, so

\[ Q(\Theta, \Theta^t) = \sum_{i=1}^{n} \sum_{k=1}^{K} p(y_i = k \mid x_i, \Theta^t) \log (p(x_i, y_i = k \mid \Theta)) \]

\[ p(y_i = k \mid x_i, \theta^t) = \frac{p_k^t p(x_i \mid \theta_k^t)}{\sum_{k'}^{} p_k^t p(x_i \mid \theta_k^t)} \]

\[ p(x_i, y_i = k \mid \theta^{t+1}) = p(y_i = k \mid \theta^{t+1}) p(x_i \mid y_i = k, \theta^{t+1}) = p_k^{t+1} p(x_i \mid \theta_k^{t+1}) \]
The Q function specific for GMM is

$$Q(\theta^t, \theta^{t+1}) = \sum_{i=1}^{n} \sum_k p_k^t p\left(x_i \mid \theta_k^t\right) \log \left(p_{k}^{t+1} p\left(x_i \mid \theta_k^{t+1}\right)\right)$$

Plug in the definition of $p(x \mid \Theta_k)$, compute derivative w.r.t. the parameters, we obtain the iteration procedures

- **E step**
  $$p_{ik}^t = p\left(y_i = k \mid x_i, \theta^t\right) = \frac{p_k^t p(x_i \mid \theta_k^t)}{\sum_{k'} p_{ik'}^t p(x_i \mid \theta_{k'}^t)},$$

- **M step**
  $$p_{k}^{t+1} = \frac{1}{n} \sum_{i=1}^{n} p_{ik}^t, \quad \mu_k^{t+1} = \frac{\sum_{i=1}^{n} p_{ik}^t x_i}{\sum_{i=1}^{n} p_{ik}^t}, \quad \Sigma_k^{t+1} = \frac{\sum_{i=1}^{n} p_{ik}^t (x_i - \mu_k^t) (x_i - \mu_k^t)^T}{\sum_{i=1}^{n} p_{ik}^t}$$