The Expectation-Maximization (EM) Algorithm
(Ch 4.4.2)

Torbjørn Svendsen

NTNU

Department of Electronics & Telecommunications
1 Introduction

2 Theory
   - Problem formulation
   - Object function
   - The EM algorithm

3 Example: Gaussian Mixture distribution
The EM algorithm is a general estimation method based on incomplete data.

Examples of use:
- Estimate model parameters of distributions based on unlabeled training data.
- \( y \) - observation vector (known).
- \( x \) - class membership of \( y \) (unknown).
- \((y, x)\) - incomplete data.
- E.g., estimate parameters of a Gaussian mixture distribution,
  \[
p(y|\Phi) = \sum_k c_k N(y; \mu_k, \Sigma_k)
\]
- Wish to find the class dependent parameter set that maximizes the probability of the observations.
The EM algorithm is a general estimation method based on incomplete data.

Examples of use:
- Estimate model parameters of distributions based on unlabeled training data.
- \( y \) - observation vector (known)
- \( x \) - class membership of \( y \) (unknown)
- \((y, x)\) - incomplete data
- e.g. estimate parameters of a Gaussian mixture distribution,
  \[ p(y|\Phi) = \sum_k c_k N(y; \mu_k, \Sigma_k) \]
- Wish to find the class dependent parameter set that maximizes the probability of the observations.
The EM algorithm is a general estimation method based on incomplete data.

**Examples of use**

- Estimate model parameters of distributions based on unlabeled training data.
- \( y \) - observation vector (known)
- \( x \) - class membership of \( y \) (unknown)
- \((y, x)\) - incomplete data
- e.g. estimate parameters of a Gaussian mixture distribution,
\[
p(y|\Phi) = \sum_{k} c_k \mathcal{N}(y; \mu_k, \Sigma_k)
\]
- Wish to find the class dependent parameter set that maximizes the probability of the observations.
Problem formulation

- We need to find
  \[ \hat{\Phi} = \arg\max_{\Phi} P(Y = y | \Phi) = \arg\max_{\Phi} \log P(Y = y | \Phi) \]

  - This often difficult or impossible to solve

- \[ \arg\max_{\Phi} P(Y = y, X = x | \Phi) \]
  - is easier to calculate, but we do not know \( x \)

- \[ \arg\max_{\Phi} E_{\Phi} \{ \log P(Y = y, X = x | \Phi) \} _{X|Y} \]
  - can be calculated (independently of \( X \)).
  - \( \Phi \) are the new parameters (unknown)
  - \( \Phi \) are "old" parameters (assumed known)
  - \( E_{\Phi} \{ f(x) \} _{X|Y} \) is the expectation of \( f(x) \) evaluated over the distribution \( P(X|Y, \Phi) \)

- Need to show that solving (3) is equivalent to solving (1)
• Bayes’ rule: \( P(X = x, Y = y | \Phi) = P(X = x | Y = y, \Phi) \cdot P(Y = y | \Phi) \)

• i.e. we can write
\[
\log P(Y = y | \Phi) = \\
\log P(X = x, Y = y | \Phi) - \log P(X = x | Y = y, \Phi)
\]

Taking the expectation:
\[
E_{\Phi}\{\log P(Y = y | \bar{\Phi})\}_{X|Y} = \\
\sum_x \log P(Y = y | \bar{\Phi}) \cdot P(X = x | Y = y, \Phi) = \\
\log P(Y = y | \bar{\Phi}) \cdot \sum_x P(X = x | Y = y, \Phi) = \log P(Y = y | \bar{\Phi})
\]
\[
= 1
\]
Proof of object function optimality (2)

\[
\log P(Y = y | \Phi) = E_{\Phi} \{ \log P(Y = y | \Phi) \}_{X|Y} \\
= E_{\Phi} \{ \log [P(X = x, Y = y | \Phi) \cdot P(X = x | Y = y, \Phi)] \}_{X|Y} \\
= E_{\Phi} \{ \log P(X = x, Y = y | \Phi) \}_{X|Y} \\
- E_{\Phi} \{ \log P(X = x, Y = y | \Phi) \}_{X|Y} \\
= Q(\Phi, \Phi) - H(\Phi, \Phi)
\]

- It can be shown that \( H(\Phi, \Phi) \leq H(\Phi, \Phi) \) (Gibb’s inequality)
- If \( \Phi \) is chosen such that \( Q(\Phi, \Phi) \geq Q(\Phi, \Phi) \), we are assured that \( \log P(Y = y | \Phi) \geq \log P(Y = y | \Phi) \)
- \( \Rightarrow \hat{\Phi} = \arg\max_{\Phi} P(Y = y | \Phi) = \arg\max_{\Phi} Q(\Phi, \Phi) \)  
  \( \Phi \) 
  (only local optimum guaranteed)
- \( Q(\Phi, \Phi) \) is called the auxilliary function
The EM algorithm

1. Choose initial parameters, $\Phi$
2. Expectation/estimation: Calculate the auxiliary function $Q(\Phi, \bar{\Phi})$
3. Maximization: $\hat{\Phi} = \arg\max_{\Phi} Q(\Phi, \bar{\Phi})$
4. Set $\Phi = \bar{\Phi}$ and repeat from 2)

- The algorithm is guaranteed to converge to a local maximum
- The result will depend upon choice of initial parameters, $\Phi$
  - Good choice of $\Phi$ important
  - Usual: Initialize using results from K-means clustering
- Interpretation: Estimate the probability that $Y$ was generated by class $X$ using assumed parameters, $\Phi$. This provides complete data that can be used for estimating new parameters.
Example: Gaussian Mixture distribution

\[
p(y|\Phi) = \sum_{k=1}^{K} p(y|\Phi_k) = \sum_{k=1}^{K} c_k \mathcal{N}(y|\mu_k, \Sigma_k)
\]

\[
= \sum_{k=1}^{K} \frac{P(X = x) \cdot p(y|x, \Phi)}{p(y,x|\Phi_k)}
\]

**TASK:** Given a set of observations, \([y_1, y_2, ..., y_N]\), where \(y\) is an \(M\)-dimensional vector randomly generated by one of \(K\) multivariate Gaussian sources, find the parameters of all sources, and the class probabilities, i.e. \((c_k, \mu_k, \Sigma_k), \forall k\)
Example: GMM - Expectation calculation

Expectation calculation Observation $i$:

$$Q_i(\Phi, \hat{\Phi}) = E_{\Phi} \{ \log p(x_i, y_i | \Phi) \} x_i | y_i$$

$N$ observations:

$$Q(\Phi, \hat{\Phi}) = \sum_{i=1}^{N} Q_i(\Phi, \hat{\Phi}) = \sum_{i=1}^{N} \sum_{x} \frac{p(x_i | y_i, \Phi)}{p(y_i | \Phi)} \log p(x_i, y_i | \hat{\Phi})$$

Define: $\gamma^i_k = \frac{p(y_i, x_i | \Phi)}{p(y_i | \Phi)} = \frac{c_k N(y_i | \mu_k, \Sigma_k)}{\sum_{k=1}^{K} c_k N(y_i | \mu_k, \Sigma_k)}$

Then

$$Q(\Phi, \hat{\Phi}) = \sum_{i=1}^{N} \sum_{k} \gamma^i_k [\log \hat{c}_k - \frac{M}{2} \log 2\pi - \log |\hat{\Sigma}_k| - (y_i - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (y_i - \hat{\mu}_k)]$$
Example: GMM - Maximization

\[ Q(\Phi, \hat{\Phi}) = \sum_{i=1}^{N} \sum_{k} \gamma_{k}^{i} [\log \hat{c}_{k} - \frac{M}{2} \log 2\pi - \log |\hat{\Sigma}_{k}| - (\mathbf{y}_{i} - \hat{\mu}_{k})^{T}\hat{\Sigma}_{k}^{-1}(\mathbf{y}_{i} - \hat{\mu}_{k})] \]

is separable in \((c_{k}, \mu_{k}, \Sigma_{k})\), i.e. we can solve by partial derivation with respect to each parameter, with the condition that \(\sum_{k} c_{k} = 1\)

1. Solve prior probabilities - Lagrange: \(\frac{\partial}{\partial \hat{c}_{k}}[Q(\Phi, \hat{\Phi}) - \lambda(\sum_{k} c_{k} - 1)] = 0;\) and \(\sum_{k} c_{k} = 1\)

2. Solve for means: \(\frac{\partial Q(\Phi, \hat{\Phi})}{\partial \hat{\mu}_{k}} = 0\)

3. Solve for covariances: \(\frac{\partial Q(\Phi, \hat{\Phi})}{\partial \hat{\Sigma}_{k}} = 0\)
Example: GMM - Maximization of priors

Note that \( \sum_{k=1}^{K} \gamma_k^i = 1 \Rightarrow \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_k^i = N \) \( (*) \)

\[
\frac{\partial}{\partial \hat{c}_j} [Q(\Phi, \hat{\Phi}) - \lambda(\sum_k \hat{c}_k - 1)] = \frac{1}{\hat{c}_j} \sum_{i=1}^{N} \gamma_j^i + \lambda = 0
\]

\[\iff \hat{c}_j = \frac{1}{\lambda} \sum_{i=1}^{N} \gamma_j^i \quad (= \frac{1}{\lambda} \gamma_j)\]

Solve for \( \lambda \) by inserting condition and observing \( (*) \)

\[
\sum_{k=1}^{K} c_k = \frac{1}{\lambda} \sum_{k} \gamma_k = \frac{N}{\lambda} \iff \lambda = \frac{1}{N}
\]

Solution: \( \hat{c}_j = \frac{1}{N} \gamma_j, \quad j = 1, \ldots, K \), \( \gamma_k = \sum_{i=1}^{N} \gamma_k^i \)
Example: GMM - Maximization solution

Similar procedure to find means and covariances. The solutions are:

\[
\hat{c}_k = \frac{1}{N} \gamma_k \\
\hat{\mu}_k = \left( \sum_{i=1}^{N} \gamma_k^i y_i \right) / \gamma_k \\
\hat{\Sigma}_k = \left[ \sum_{i=1}^{N} \gamma_k^i (y_i - \mu_k)(y_i - \mu_k)^T \right] / \gamma_k
\]

where \( \gamma_k = \sum_{i=1}^{N} \gamma_k^i \)
Example: GMM - EM iteration

1. Choose initial parameters, \( \{c_k, \mu_k, \Sigma_k \mid k = 1, \ldots, K\} = \Phi \)

2. Estimate \( \{\gamma_i^k\} \) using \( \Phi \) and the observations \( [y_1, \ldots, y_N] \)

3. Maximize \( Q(\Phi, \hat{\Phi}) \) to find the new parameter set \( \hat{\Phi} \).

4. Set \( \Phi = \hat{\Phi} \) and iterate from 2. until a termination criterion is fulfilled.