Expectation-Maximization (EM) Algorithm

Soma Biswas
Department of Electrical Engineering,
Indian Institute of Science, Bangalore.
MLE Overview

- Say I give you a coin with $P(\text{heads}) = \theta, P(\text{tails}) = 1 - \theta$
- But I don’t tell you the value of $\theta$
- Now say I let you flip the coin $n$ times
  - You get $h$ heads and $n-h$ tails
- What is the natural estimate of $\theta$?
  - This is $\hat{\theta} = h/n$
- More formally, the likelihood of $\theta$ is governed by a binomial distribution: $P(\theta) = \binom{n}{h} \theta^h (1 - \theta)^{n-h}$
  - Can prove $\hat{\theta}$ is the maximum-likelihood estimate of $\theta$
  - Differentiate with respect to $\theta$, set equal to 0
Maximum-Likelihood

- Density Function $p(x|\Theta) : \Theta$ set of parameters (eg. Means and co-variances for a set of Gaussians)

- Data Set of size N drawn from this distribution $\mathcal{X} = \{x_1, \ldots, x_N\}$.

- Assume: data iid: independent and identically distributed with distribution $p$

- Density for the samples is: $p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} p(x_i|\Theta) = \mathcal{L}(\Theta|\mathcal{X})$.

  Likelihood of the parameters given the data (function of the parameters when the data $X$ is fixed)
EM Motivation

- To solve any ML-type problem, we analytically maximize the likelihood function
  - Works for 1D Gaussian (find $\mu$, $\sigma^2$)

- Problems:
  - Distribution may not be well-behaved, or have too many parameters
  - Say your likelihood function is a mixture of 1000 1000-dimensional Gaussians (1M parameters)
  - Direct maximization is not feasible

- Solution: introduce hidden variables to
  - Simplify the likelihood function (more common)
  - Account for actual missing data

- EM algorithm used to find the maximum likelihood parameters in cases where the equations cannot be solved directly.
- Involve latent variables + unknown parameters + known data observations.
- Either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points.
Hidden and Observed Variables

- **Observed** variables: directly measurable from the data, e.g.
  - The waveform values of a speech recording
  - Is it raining today?
  - Did the smoke alarm go off?

- **Hidden** variables: influence the data, but not trivial to measure
  - The phonemes that produce a given speech recording
  - \( P( \text{rain today} | \text{rain yesterday}) \)
  - Is the smoke alarm malfunctioning?
Review

- Let $X$, $Y$ be r.v.’s drawn from the distributions $P(x)$ and $P(y)$
- Conditional distribution given by: $P(y|x) = \frac{P(y, x)}{P(x)}$
- Then $E[Y] = \sum_y P(y)y$
- For function $h(Y)$: $E[h(Y)] = \sum_y P(y)h(y)$
- Given a particular value of $X$ ($X=x$):
  $E[h(Y)|x] = \sum_y P(y|x)h(y)$
EM

- **Assume:** Data $X$ is observed and is generated by some distribution.

- Incomplete data

- Complete data set exists $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$

- Joint Density Function: $p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta)$

- Complete Data Likelihood Function: $\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) = p(\mathcal{X}, \mathcal{Y}|\Theta)$

  \begin{itemize}
  \item Random Variable since the missing information $Y$ is unknown and random
  \end{itemize}

- Incomplete data likelihood function: $\mathcal{L}(\Theta|\mathcal{X})$

- **E Step:** expected value of the complete-data log-likelihood with respect to the unknown data $Y$ given the observed data $X$ and the current parameter estimates.

  $$Q(\Theta, \Theta^{(i-1)}) = E \left[ \log p(X, Y|\Theta)|\mathcal{X}, \Theta^{(i-1)} \right]$$
EM

- \( \Theta^{(i-1)} \): current parameters estimates that we used to evaluate the expectation
- \( \Theta \): new parameters that we optimize to increase \( Q \)

\( \mathcal{X} \) and \( \Theta^{(i-1)} \) are constants, \( \Theta \) is a normal variable that we wish to adjust, and \( \mathcal{Y} \) is a random variable governed by the distribution \( f(\mathcal{Y} | \mathcal{X}, \Theta^{(i-1)}) \). The right side of Equation 1 can therefore be

\[
E \left[ \log p(\mathcal{X}, \mathcal{Y} | \Theta) | \mathcal{X}, \Theta^{(i-1)} \right] = \int_{\mathcal{Y} \in \mathcal{Y}} \log p(\mathcal{X}, \mathcal{Y} | \Theta) f(\mathcal{Y} | \mathcal{X}, \Theta^{(i-1)}) d\mathcal{Y}
\]

The marginal distribution of the unobserved data and is dependent on both the observed data and on the current parameters.

- Easy case: the marginal distribution is a simple analytical expression
- Difficult Case: this density might be very hard to obtain.
EM-Steps

- **E Step:** expected value of the complete-data log-likelihood with respect to the unknown data $Y$ given the observed data $X$ and the current parameter estimates.

\[
Q(\Theta, \Theta^{(i-1)}) = E \left[ \log p(X, Y|\Theta) | X, \Theta^{(i-1)} \right]
\]

- **M Step:**
  - Maximize the expectation we computed in the first step
  \[
  \Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)})
  \]
  - Each iteration is guaranteed to increase the log likelihood
  - Algorithm is guaranteed to converge to a local maximum of the likelihood function.
EM Summary

Repeat until convergence
  - E-step: Compute expectation of \( \log P_\theta(y, x) \)
    \[
    Q(\theta, \theta') = \sum_y P_\theta(y|x) \log P_\theta(y, x)
    \]
    (\( \theta', \theta \): old, new distribution parameters)
  - M-step: Find \( \theta \) that maximizes \( Q \)

EM Theorem:
  - If \( \sum_y \log P_\theta(y, x) P_{\theta'}(y|x) > \sum_y \log P_{\theta'}(y, x) P_{\theta'}(y|x) \)
  - then \( P_\theta(x) > P_{\theta'}(x) \)

Interpretation
  - As long as we can improve the expectation of the log-likelihood, EM improves our model of observed variable \( x \)
  - Actually, it’s not necessary to maximize the expectation, just need to make sure that it increases – this is called “Generalized EM”

\[
Q(\Theta^{(i)}, \Theta^{(i-1)}) > Q(\Theta, \Theta^{(i-1)})
\]
Application: Gaussian Mixture Model (GMM)

- Gaussian/normal distribution
  - Parameters: mean $\mu$ and variance $\sigma^2$
    \[ G_{\mu,\sigma}(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \]
  - In the multi-dimensional case, assume isotropic Gaussian: same variance in all dimensions
  - We can model arbitrary distributions with density mixtures

For $d$ dimensions, the Gaussian distribution of a vector $x = (x^1, x^2, \ldots, x^d)^T$ is defined by:

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

where $\mu$ is the mean and $\Sigma$ is the covariance matrix of the Gaussian.

**Example:**

\[ \mu = (0,0)^T \quad \Sigma = \begin{pmatrix} 0.25 & 0.30 \\ 0.30 & 1.00 \end{pmatrix} \]
Density Mixtures

- Combine $m$ elementary densities to model a complex data distribution
  \[ P_\Theta(x) = \sum_{i=1}^{m} \alpha_i P_{\theta_i}(x) \quad \Theta = (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_m) \]
- $k$th Gaussian parametrized by $\theta_k = \{\mu_k, \sigma_k\}$
What is a GMM?

• **Problem:**

  Given a set of data \( X = \{x_1, x_2, \ldots, x_N \} \) drawn from an unknown distribution (probably a GMM), estimate the parameters \( \theta \) of the GMM model that fits the data.

• **Solution:**

  Maximize the likelihood \( p(X \mid \theta) \) of the data with regard to the model parameters?

  \[
  \theta^* = \arg \max_{\theta} p(X \mid \theta) = \arg \max_{\theta} \prod_{i=1}^{N} p(x_i \mid \theta)
  \]
EM for GMM

- Combine $m$ elementary densities to model a complex data distribution

$$P_{\Theta}(x) = \sum_{i=1}^{m} \alpha_i P_{\theta_i}(x) \quad \Theta = (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_m)$$

- Log-likelihood function of the data $x$ given $\Theta$:

$$\log P_{\Theta}(x) = \log \left( \sum_{i=1}^{m} \alpha_i P_{\theta_i}(x) \right)$$

- Log of sum – hard to optimize analytically!

- Instead, introduce hidden variable $y$
  - $y = k : x$ generated by Gaussian $k$
  - $\log P_{\Theta}(x, y) = \log \left( \alpha_y P_{\theta_y}(x) \right)$

- EM formulation: maximize

$$Q(\Theta, \Theta') = \sum_{y} P_{\theta'_y}(y|x) \log \left( \alpha_y P_{\theta_y}(x) \right)$$
EM for GMM Contd.

- **Goal**: maximize $Q(\Theta, \Theta') = \sum_y P_{\Theta'}(y|x) \log (\alpha_y P_{\Theta_y}(x))$
- $n$ (observed) data points: $x_1, \ldots, x_n$
- $n$ (hidden) labels: $y_1, \ldots, y_n$
  - $y_i = k : x_i$ generated by Gaussian $k$
- Several pages of math later, we get:
- **E step**: compute likelihood of $y_i = k$

$$\Lambda_{i,k} = P_{\Theta'}(y_i = k|x_i) = \frac{\alpha'_k P_{\Theta'_k}(x_i)}{P_{\Theta'}(x_i)} = \frac{\alpha'_k P_{\Theta'_k}(x_i)}{\sum_{j=1}^m \alpha_j P_{\Theta'_j}(x_i)}$$

- **M step**: update $\alpha_k$, $\mu_k$, $\sigma_k$ for each Gaussian $k=1..m$

$$\alpha_k = \frac{1}{n} \sum_{i=1}^n \Lambda_{i,k}$$
$$\mu_k = \frac{\sum_{i=1}^n x_i \Lambda_{i,k}}{\sum_{i=1}^n \Lambda_{i,k}}$$
$$\sigma_k^2 = \frac{\sum_{i=1}^n \Lambda_{i,k} ||x_i - \mu_k||^2}{\sum_{i=1}^n \Lambda_{i,k}}$$
EM for GMM – Graphical Representation

Hidden variable: for each point, which Gaussian generated it?

E-Step: for each point, estimate the probability that each Gaussian generated it.

M-Step: modify the parameters according to the hidden variable to maximize the likelihood of the data (and the hidden variable).
EM for HMM

\[ Q(\Theta, \Theta^{(i-1)}) = E \left[ \log P(\mathcal{X}, \mathcal{Y}|\Theta) | \mathcal{X}, \Theta^{(i-1)} \right] = \int_{\mathcal{Y} \in \mathcal{Y}} \log P(\mathcal{X}, \mathcal{Y}|\Theta) f(\mathcal{Y} | \mathcal{X}, \Theta^{(i-1)}) d\mathcal{Y} \]

We consider \( O = (o_1, \ldots, o_T) \) to be the observed data and the underlying state sequence \( q = (q_1, \ldots, q_T) \) to be hidden or unobserved. The incomplete-data likelihood function is given by \( P(O|\lambda) \) whereas the complete-data likelihood function is \( P(O, q|\lambda) \). The \( Q \) function therefore is:

\[ Q(\lambda, \lambda') = \sum_{q \in \mathcal{Q}} \log P(O, q|\lambda) P(O, q|\lambda') \]

where \( \lambda' \) are our initial (or guessed, previous, etc.) estimates of the parameters

\[ P(O, q|\lambda) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \]

\[ Q(\lambda, \lambda') = \sum_{q \in \mathcal{Q}} \log \pi_{q_0} P(O, q|\lambda') + \sum_{q \in \mathcal{Q}} \left( \sum_{t=1}^{T} \log a_{q_{t-1}q_t} \right) p(O, q|\lambda') + \sum_{q \in \mathcal{Q}} \left( \sum_{t+1}^{T} \log b_{q_t}(o_t) \right) P(O, q|\lambda') \]

Since the parameters we wish to optimize are now independently split into the three terms in the sum, we can optimize each term individually.
EM for HMM

- **First Term**  
  \[ \sum_{q \in Q} \log \pi_{q_0} P(O, q|\lambda') = \sum_{i=1}^{N} \log \pi_i p(O, q_0 = i|\lambda') \]

- **Using Lagrange Multiplier**  
  \[ \frac{\partial}{\partial \pi_i} \left( \sum_{i=1}^{N} \log \pi_i p(O, q_0 = i|\lambda') + \gamma \left( \sum_{i=1}^{N} \pi_i - 1 \right) \right) = 0 \]
  \[ \pi_i = \frac{P(O, q_0 = i|\lambda')}{P(O|\lambda')} \]

- **Second Term**  
  \[ \sum_{q \in Q} \left( \sum_{t=1}^{T} \log a_{q_{t-1}q_t} \right) p(O, q|\lambda') = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \log a_{ij} P(O, q_{t-1} = i, q_t = j|\lambda') \]

- **Using Lagrange multiplier**  
  \[ \sum_{j=1}^{T} a_{ij} = 1 \]

  \[ a_{ij} = \frac{\sum_{t=1}^{T} P(O, q_{t-1} = i, q_t = j|\lambda')} {\sum_{t=1}^{T} P(O, q_{t-1} = i|\lambda')} } \]

  \[ b_i(k) = \frac{\sum_{t=1}^{T} P(O, q_t = i|\lambda') \delta_{q_t, v_k}} {\sum_{t=1}^{T} P(O, q_t = i|\lambda')} } \quad \sum_{j=1}^{L} b_i(j) = 1 \]