Machine Learning for OR & FE
The EM Algorithm

Martin Haugh
Department of Industrial Engineering and Operations Research
Columbia University
Email: martin.b.haugh@gmail.com
Outline

The EM Algorithm
  E.G. Missing Data in a Multinomial Model
  E.G. Normal Mixture Models Revisited

Detour: Kullback-Leibler Divergence
  The EM Algorithm Revisited
  E.G. Questionnaires and Missing Data
The EM Algorithm (for Computing ML Estimates)

Assume the complete data-set consists of \( \mathcal{Z} = (\mathcal{X}, \mathcal{Y}) \)
– but only \( \mathcal{X} \) is observed.

The complete-data log likelihood is denoted by \( l(\theta; \mathcal{X}, \mathcal{Y}) \) where \( \theta \) is the unknown parameter vector for which we wish to find the MLE.

**E-Step:** Compute the expected value of \( l(\theta; \mathcal{X}, \mathcal{Y}) \) given the observed data, \( \mathcal{X} \), and the current parameter estimate \( \theta_{old} \). In particular, we define

\[
Q(\theta; \theta_{old}) := \mathbb{E} [ l(\theta; \mathcal{X}, \mathcal{Y}) | \mathcal{X}, \theta_{old} ]
\]

\[
= \int l(\theta; \mathcal{X}, y) \ p(y | \mathcal{X}, \theta_{old}) \ dy
\]  

– \( p(\cdot | \mathcal{X}, \theta_{old}) \equiv \) conditional density of \( \mathcal{Y} \) given observed data, \( \mathcal{X} \), and \( \theta_{old} \)
– \( Q(\theta; \theta_{old}) \) is the expected complete-data log-likelihood.

**M-Step:** Compute \( \theta_{new} := \max_{\theta} Q(\theta; \theta_{old}) \).
The EM Algorithm

Now set $\theta_{old} = \theta_{new}$ and iterate E- and M-steps until sequence of $\theta_{new}$’s converges.

Convergence to a local maximum can be guaranteed under very general conditions

– will see why below.

If suspected that log-likelihood function has multiple local maximaums then the EM algorithm should be run many times

– using a different starting value of $\theta_{old}$ on each occasion.

The ML estimate of $\theta$ is then taken to be the best of the set of local maximaums obtained from the various runs of the EM algorithm.
Why Does the EM Algorithm Work?

Will use $p(\cdot | \cdot)$ to denote a generic conditional PDF. Now observe that

$$l(\theta; X) = \ln p(X | \theta)$$
$$= \ln \int p(X, y | \theta) \, dy$$
$$= \ln \int \frac{p(X, y | \theta)}{p(y | X, \theta_{old})} p(y | X, \theta_{old}) \, dy$$
$$= \ln \mathbb{E} \left[ \frac{p(X, Y | \theta)}{p(Y | X, \theta_{old})} \right] | X, \theta_{old}$$
$$\geq \mathbb{E} \left[ \ln \left( \frac{p(X, Y | \theta)}{p(Y | X, \theta_{old})} \right) \right] | X, \theta_{old} \tag{2} \text{ by Jensen's inequality}$$
$$= \mathbb{E} \left[ \ln p(X, Y | \theta) \right] | X, \theta_{old} - \mathbb{E} \left[ \ln p(Y | X, \theta_{old}) \right] | X, \theta_{old}$$
$$= Q(\theta; \theta_{old}) - \mathbb{E} \left[ \ln p(Y | X, \theta_{old}) \right] | X, \theta_{old} \tag{3}$$

Also clear (why?) that inequality in (2) is an equality if we take $\theta = \theta_{old}$.
Why Does the EM Algorithm Work?

Let \( g(\theta \mid \theta_{old}) \) denote the right-hand-side of (3).

Therefore have

\[
l(\theta; \mathcal{X}) \geq g(\theta \mid \theta_{old})
\]

for all \( \theta \) with equality when \( \theta = \theta_{old} \).

So any value of \( \theta \) that increases \( g(\theta \mid \theta_{old}) \) beyond \( g(\theta_{old} \mid \theta_{old}) \) must also increase \( l(\theta; \mathcal{X}) \) beyond \( l(\theta_{old}; \mathcal{X}) \).

The M-step finds such a \( \theta \) by maximizing \( Q(\theta; \theta_{old}) \) over \( \theta \)

- this is equivalent (why?) to maximizing \( g(\theta \mid \theta_{old}) \) over \( \theta \).

Also worth noting that in many applications the function \( Q(\theta; \theta_{old}) \) will be a convex function of \( \theta \)

- and therefore easy to optimize.
Schematic for general E-M algorithm

\[
\text{ln } p(X|\theta)
\]

**Figure 9.14 from Bishop** (where \( \mathcal{L}(q, \theta) \) is \( g(\theta | \theta_{old}) \) in our notation)
E.G. Missing Data in a Multinomial Model

Suppose \( x := (x_1, x_2, x_3, x_4) \) is a sample from a \( \text{Mult}(n, \pi_\theta) \) distribution where

\[
\pi_\theta = \left( \frac{1}{2} + \frac{1}{4} \theta, \frac{1}{4} (1 - \theta), \frac{1}{4} (1 - \theta), \frac{1}{4} \theta \right).
\]

The likelihood, \( L(\theta; x) \), is then given by

\[
L(\theta; x) = \frac{n!}{x_1!x_2!x_3!x_4!} \left( \frac{1}{2} + \frac{1}{4} \theta \right)^{x_1} \left( \frac{1}{4} (1 - \theta) \right)^{x_2} \left( \frac{1}{4} (1 - \theta) \right)^{x_3} \left( \frac{1}{4} \theta \right)^{x_4}
\]

so that the log-likelihood \( l(\theta; x) \) is

\[
l(\theta; x) = C + x_1 \ln \left( \frac{1}{2} + \frac{1}{4} \theta \right) + (x_2 + x_3) \ln (1 - \theta) + x_4 \ln (\theta)
\]

– where \( C \) is a constant that does not depend on \( \theta \).

Could try to maximize \( l(\theta; x) \) over \( \theta \) directly using standard non-linear optimization algorithms

– but we will use the EM algorithm instead.
E.G. Missing Data in a Multinomial Model

To do this we assume the complete data is given by $y := (y_1, y_2, y_3, y_4, y_5)$ and that $y$ has a Mult$(n, \pi_\theta^*)$ distribution where

$$\pi_\theta^* = \left( \frac{1}{2}, \frac{1}{4} \theta, \frac{1}{4} (1 - \theta), \frac{1}{4} (1 - \theta), \frac{1}{4} \theta \right).$$

However, instead of observing $y$ we only observe $(y_1 + y_2, y_3, y_4, y_5)$, i.e, $x$.

Therefore take $X = (y_1 + y_2, y_3, y_4, y_5)$ and take $Y = y_2$.

Log-likelihood of complete data then given by

$$l(\theta; X, Y) = C + y_2 \ln(\theta) + (y_3 + y_4) \ln(1 - \theta) + y_5 \ln(\theta)$$

where again $C$ is a constant containing all terms that do not depend on $\theta$.

Also “clear” that conditional “density” of $Y$ satisfies

$$f(Y \mid X, \theta) = \text{Bin} \left( y_1 + y_2, \frac{\theta/4}{1/2 + \theta/4} \right).$$

Can now implement the E-step and M-step.
Recall that \( Q(\theta; \theta_{old}) := E[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old}] \).

**E-Step:** Therefore have

\[
Q(\theta; \theta_{old}) := C + E[y_2 \ln(\theta) \mid \mathcal{X}, \theta_{old}] + (y_3 + y_4) \ln(1 - \theta) + y_5 \ln(\theta)
\]

\[
= C + (y_1 + y_2)p_{old} \ln(\theta) + (y_3 + y_4) \ln(1 - \theta) + y_5 \ln(\theta)
\]

where

\[
p_{old} := \frac{\theta_{old}/4}{1/2 + \theta_{old}/4}.
\]
**E.G. Missing Data in a Multinomial Model**

**M-Step:** Must now maximize $Q(\theta; \theta_{old})$ to find $\theta_{new}$.

Taking the derivative we obtain

$$\frac{dQ}{d\theta} = \frac{(y_1 + y_2)}{\theta} p_{old} - \frac{(y_3 + y_4)}{1 - \theta} + \frac{y_5}{\theta}$$

$$= 0 \quad \text{when} \quad \theta = \theta_{new}$$

where

$$\theta_{new} := \frac{y_5 + p_{old}(y_1 + y_2)}{y_3 + y_4 + y_5 + p_{old}(y_1 + y_2)}.$$  \hfill (5)

Equations (4) and (5) now define the EM iteration

– which begins with some judiciously chosen value of $\theta_{old}$. 
Clustering via normal mixture models is an example of probabilistic clustering
– we assume the data are IID draws
– will consider only the scalar case but note the vector case is similar.

So suppose $X = (X_1, \ldots, X_n)$ are IID random variables each with PDF

$$f_x(x) = \sum_{j=1}^{m} p_j \frac{e^{-(x-\mu_j)^2/2\sigma_j^2}}{\sqrt{2\pi\sigma_j^2}}$$

where $p_j \geq 0$ for all $j$ and where $\sum_j p_j = 1$

– parameters are the $p_j$’s, $\mu_j$’s and $\sigma_j$’s
– typically estimated via MLE
– which we can do via the EM algorithm.
Normal Mixture Models Revisited

We assume the presence of an additional or latent random variable, $Y$, where

$$P(Y = j) = p_j, \quad j = 1, \ldots, m.$$ 

Realized value of $Y$ then determines which of the $m$ normals generates the corresponding value of $X$

− so there are $n$ such random variables, $(Y_1, \ldots, Y_n) := \mathcal{Y}$.

Note that

$$f_{x|y}(x_i \mid y_i = j, \theta) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}} \quad (6)$$

where $\theta := (p_1, \ldots, p_m, \mu_1, \ldots, \mu_m, \sigma_1, \ldots, \sigma_m)$ is the unknown parameter vector.

The complete data likelihood is given by

$$L(\theta; X, \mathcal{Y}) = \prod_{i=1}^{n} p_{y_i} \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} e^{-\frac{(x_i - \mu_{y_i})^2}{2\sigma_{y_i}^2}}.$$
Normal Mixture Models

The EM algorithm starts with an initial guess, $\theta_{old}$, and then iterates the E-step and M-step until convergence.

**E-Step:** Need to compute $Q(\theta; \theta_{old}) := E[l(\theta; X, Y) \mid X, \theta_{old}]$.

Straightforward to show that

$$Q(\theta; \theta_{old}) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(Y_i = j \mid x_i, \theta_{old}) \ln \left( f_{x \mid y}(x_i \mid y_i = j, \theta) P(Y_i = j \mid \theta) \right).$$

(7)

Note that $f_{x \mid y}(x_i \mid y_i = j, \theta)$ is given by (6) and that $P(Y_i = j \mid \theta_{old}) = p_{j, old}$.

Finally, can compute (7) analytically since

$$P(Y_i = j \mid x_i, \theta_{old}) = \frac{P(Y_i = j, X_i = x_i \mid \theta_{old})}{P(X_i = x_i \mid \theta_{old})} = \frac{f_{x \mid y}(x_i \mid y_i = j, \theta_{old}) P(Y_i = j \mid \theta_{old})}{\sum_{k=1}^{m} f_{x \mid y}(x_i \mid y_i = k, \theta_{old}) P(Y_i = k \mid \theta_{old})}. \quad (8)$$
**Normal Mixture Models**

**M-Step:** Can now maximize $Q(\theta; \theta_{old})$ by setting the vector of partial derivatives, $\partial Q/\partial \theta$, equal to 0 and solving for $\theta_{new}$.

After some algebra, we obtain

\[
\mu_{j,new} = \frac{\sum_{i=1}^{n} x_i P(Y_i = j \mid x_i, \theta_{old})}{\sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old})}
\]  
\[
\sigma_{j,new}^2 = \frac{\sum_{i=1}^{n} (x_i - \mu_j)^2 P(Y_i = j \mid x_i, \theta_{old})}{\sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old})}
\]  
\[
p_{j,new} = \frac{1}{n} \sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old}).
\]

Given an initial estimate, $\theta_{old}$, the EM algorithm cycles through (9) to (11) repeatedly, setting $\theta_{old} = \theta_{new}$ after each cycle, until the estimates converge.
Kullback-Leibler Divergence

Let $P$ and $Q$ be two probability distributions such that if $Q(x) = 0$ then $P(x) = 0$.

The Kullback-Leibler (KL) divergence or relative entropy of $Q$ from $P$ is defined to be

$$KL(P || Q) = \int_x P(x) \ln \left( \frac{P(x)}{Q(x)} \right)$$

(12)

with the understanding that $0 \log 0 = 0$.

The KL divergence is a fundamental concept in information theory and machine learning.

Can imagine $P$ representing some true but unknown distribution that we approximate with $Q$

- $KL(P || Q)$ then measures the “distance” between $P$ and $Q$.

This interpretation is valid because we will see below that $KL(P || Q) \geq 0$

- with equality if and only if $P = Q$. 


Kullback-Leibler Divergence

The KL divergence is not a true measure of distance since it is:

1. **Asymmetric** in that $\text{KL}(P \| Q) \neq \text{KL}(Q \| P)$
2. And does not satisfy the **triangle inequality**.

In order to see that $\text{KL}(P \| Q) \geq 0$ we first recall that a function $f(\cdot)$ is **convex** on $\mathbb{R}$ if

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \quad \text{for all } \alpha \in [0, 1].$$

We also recall Jensen’s inequality:

**Jensen’s Inequality:** Let $f(\cdot)$ be a convex function on $\mathbb{R}$ and suppose $E[X] < \infty$ and $E[f(X)] < \infty$. Then $f(E[X]) \leq E[f(X)]$. 
Kullback-Leibler Divergence

Noting that $-\ln(x)$ is a convex function we have

\[
\text{KL}(P \parallel Q) = -\int_x P(x) \ln \left( \frac{Q(x)}{P(x)} \right) \\
\geq -\ln \left( \int_x P(x) \frac{Q(x)}{P(x)} \right) \quad \text{by Jensen’s inequality} \\
= 0.
\]

Moreover it is clear from (12) that $\text{KL}(P \parallel Q) = 0$ if $P = Q$.

In fact because $-\ln(x)$ is strictly convex it is easy to see that $\text{KL}(P \parallel Q) = 0$ only if $P = Q$. 
A Nice Optimization “Trick”

Suppose $c \in \mathbb{R}_+^n$ and we wish to maximize $\sum_{i=1}^n c_i \ln(q_i)$ over pmf’s, $q = \{q_1, \ldots, q_n\}$.

Let $p = \{p_1, \ldots, p_n\}$ where $p_i := c_i / \sum_j c_j$ so that $p$ is a (known) pmf.

We then have:

$$\max_q \sum_{i=1}^n c_i \ln(q_i) = \left( \sum_{i=1}^n c_i \right) \max_q \left\{ \sum_{i=1}^n p_i \ln(q_i) \right\}$$

$$= \left( \sum_{i=1}^n c_i \right) \max_q \left\{ \sum_{i=1}^n p_i \ln(p_i) - \sum_{i=1}^n p_i \ln \left( \frac{p_i}{q_i} \right) \right\},$$

$$= \left( \sum_{i=1}^n c_i \right) \left( \sum_{i=1}^n p_i \ln(p_i) - \min_{q} KL(p \mid\mid q) \right),$$

from which it follows (why?) that the optimal $q^*$ satisfies $q^* = p$.

Could have saved some time using this trick in earlier multinomial model example — in particular obtaining (5)
The EM Algorithm Revisited

As before, goal is to maximize the likelihood function \( L(\theta; \mathcal{X}) \) which is given by

\[
L(\theta; \mathcal{X}) = p(\mathcal{X} | \theta) = \int_y p(\mathcal{X}, y | \theta) \, dy.
\]  

(13)

Implicit assumption underlying EM algorithm: it is difficult to optimize \( p(\mathcal{X} | \theta) \) with respect to \( \theta \) directly

– but much easier to optimize \( p(\mathcal{X}, \mathcal{Y} | \theta). \)

First introduce an arbitrary distribution, \( q(\mathcal{Y}) \), over the latent variables, \( \mathcal{Y} \).

Note we can decompose log-likelihood, \( l(\theta; \mathcal{X}) \), into two terms according to

\[
l(\theta; \mathcal{X}) := \ln p(\mathcal{X} | \theta) = \mathcal{L}(q, \theta) + \text{KL}(q \| p_{\mathcal{Y}|\mathcal{X}})
\]

(14)

“energy”
The EM Algorithm Revisited

$L(q, \theta)$ and $\text{KL}(q \| p_{Y|X})$ are the likelihood and KL divergence and are given by

$$L(q, \theta) = \int_Y q(Y) \ln \left( \frac{p(X, Y|\theta)}{q(Y)} \right)$$

$$\text{KL}(q \| p_{Y|X}) = -\int_Y q(Y) \ln \left( \frac{p(Y|X, \theta)}{q(Y)} \right).$$

(15)

It therefore follows (why?) that $L(q, \theta) \leq l(\theta; X)$ for all distributions, $q(\cdot)$.

Can now use the decomposition of (14) to define the EM algorithm, beginning with an initial parameter estimate, $\theta_{old}$. 
The EM Algorithm Revisited

**E-Step:** Maximize the lower bound, $\mathcal{L}(q, \theta_{old})$, with respect to $q(\cdot)$ while keeping $\theta_{old}$ fixed.

In principle this is a variational problem since we are optimizing a functional, but the solution is easily found. First note that $l(\theta_{old}; \mathcal{X})$ does not depend on $q(\cdot)$.

Then follows from (14) with $\theta = \theta_{old}$ that maximizing $\mathcal{L}(q, \theta_{old})$ is equivalent to minimizing $\text{KL}(q \| p_Y|X)$.

Since this latter term is always non-negative we see that $\mathcal{L}(q, \theta_{old})$ is optimized when $\text{KL}(q \| p_Y|X) = 0$

- by earlier observation, this is the case when we take $q(Y) = p(Y|X, \theta_{old})$.

At this point we see that the lower bound, $\mathcal{L}(q, \theta_{old})$, now equals current value of log-likelihood, $l(\theta_{old}; \mathcal{X})$. 

The EM Algorithm Revisited

**M-Step:** Keep $q(Y)$ fixed and maximize $\mathcal{L}(q, \theta)$ over $\theta$ to obtain $\theta_{new}$.

This will therefore cause the lower bound to increase (if it is not already at a maximum)

– which in turn means the log-likelihood must also increase.

Moreover, at this new value $\theta_{new}$ it will no longer be the case that $\text{KL}(q \parallel p_Y|x) = 0$

– so by (14) the increase in the log-likelihood will be greater than the increase in the lower bound.
Comparing Classical EM With General EM

It is instructive to compare the E-step and M-step of the general EM algorithm with the corresponding steps of the original EM algorithm.

To do this, first substitute \( q(Y) = p(Y | X, \theta_{old}) \) into (15) to obtain

\[
L(q, \theta) = Q(\theta; \theta_{old}) + \text{constant} \quad (16)
\]

where \( Q(\theta; \theta_{old}) \) is the expected complete-date log-likelihood as defined in (1).

The M-step of the general EM algorithm is therefore identical to the M-step of original algorithm since the constant term in (16) does not depend on \( \theta \).

The E-step in general EM algorithm takes \( q(Y) = p(Y | X, \theta_{old}) \) which, at first glance, appears to be different to original E-step.

But there is no practical difference: original E-step simply uses \( p(Y | X, \theta_{old}) \) to compute \( Q(\theta; \theta_{old}) \) and, while not explicitly stated, the general E-step must also do this since it is required for the M-step.
E.G. Imputing Missing Data (Again)

\( N \) respondents were asked to answer \( m \) questions each. The observed data are:

\[
v_{iq} = \begin{cases} 
1 & \text{if respondent } i \text{ answered yes to question } q \\
0 & \text{if respondent } i \text{ answered no to question } q \\
- & \text{if respondent } i \text{ did not answer question } q 
\end{cases}
\]

\[
y_{iq} = \begin{cases} 
1 & v_{iq} \in \{0, 1\} \\
0 & \text{otherwise}
\end{cases}
\]

We assume the following model:

- \( K \) classes of respondents: \( \pi = (\pi_1, \ldots, \pi_K) \) with \( \pi_k = P(\text{respondent in class } k) \)
- Latent variables \( z_i \in \{1, \ldots, K\} \) for \( i = 1, \ldots, N \)
- Class dependent probability of answers: \( \sigma_{kq} = \mathbb{P}(v_{iq} = 1 \mid z_i = k) \)
- Parameters \( \theta = (\pi, \sigma) \)

Log-likelihood with \( \mathcal{X} := \{v_{iq} \mid i = 1, \ldots, N, q = 1, \ldots, m\} \):

\[
l(\theta; \mathcal{X}) = \sum_{i=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k \prod_{q:y_{iq}=1} \sigma_{kq} v_{iq} (1 - \sigma_{kq})^{1-v_{iq}} \right)
\]

**Question:** What implicit assumptions are we making here?
EM for Imputing Missing Data

Take $Y := (z_1, \ldots, z_N)$.

Complete-data log-likelihood then given by

$$l(\theta; \mathcal{X}, Y) = \sum_{i=1}^{N} \sum_{k=1}^{K} 1_{\{z_i = k\}} \ln \left( \pi_k \prod_{q: y_{iq} = 1} \sigma_{kq}^{v_{iq}} (1 - \sigma_{kq})^{(1 - v_{iq})} \right)$$

**E-Step:** Need to compute $Q(\theta; \theta_{old})$. We have

$$Q(\theta; \theta_{old}) = \mathbb{E} [l(\theta; \mathcal{X}, Y) \mid \mathcal{X}, \theta_{old}]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}^{old} \ln \left( \pi_k \prod_{q: y_{iq} = 1} \sigma_{kq}^{v_{iq}} (1 - \sigma_{kq})^{(1 - v_{iq})} \right)$$

where

$$\gamma_{ik}^{old} := \mathbb{P}(z_i = k \mid v_i, \theta_{old}) \propto \pi_k^{old} \mathbb{P}(v_i \mid z_i = k)$$

$$= \pi_k^{old} \prod_{q: y_{iq} = 1} (\sigma_{kq}^{old})^{v_{iq}} (1 - \sigma_{kq}^{old})^{(1 - v_{iq})}$$
EM for Imputing Missing Data

**M-Step:** Now solve for $\theta_{new} = \max_\theta Q(\theta; \theta_{old})$:

We have

$$Q(\theta; \theta_{old}) = \sum_{k=1}^{K} \left( \sum_{i=1}^{N} \gamma_{ik}^{old} \right) \ln(\pi_k) + \sum_{k=1}^{K} \sum_{q=1}^{m} \left( \sum_{i:y_{iq}=1} \gamma_{ik}^{old} v_{iq} \right) \ln(\sigma_{kq})$$

$$+ \left( \sum_{i:y_{iq}=1} \gamma_{ik}^{old} (1 - v_{iq}) \right) \ln(1 - \sigma_{kq})$$

Solving $\max_\theta Q(\theta; \theta_{old})$ yields

$$\pi_{k}^{new} = \frac{\sum_{i=1}^{N} \gamma_{ik}^{old}}{\sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ij}^{old}}$$

$$\sigma_{kq}^{new} = \frac{\sum_{i:y_{iq}=1} \gamma_{ik}^{old} v_{iq}}{\sum_{i:y_{iq}=1} \gamma_{ik}^{old}}.$$ 

Now iterate E- and M-steps until convergence.