Expectation-Maximization

Nuno Vasconcelos

ECE Department, UCSD
Expectation-maximization

we have seen that EM is a framework for ML estimation with missing data

i.e. problems where we have, two types of random variables

• $X$ observed random variable
• $Z$ hidden random variable

goal:

• given iid sample $D = \{x_1, \ldots, x_n\}$
• find parameters $\Psi^*$ that maximize likelihood with respect to $D$

\[
\Psi^* = \arg \max_{\Psi} P_X(\mathcal{D}; \Psi) \\
= \arg \max_{\Psi} \int P_{X \mid Z}(\mathcal{D} \mid z; \Psi) P_Z(z; \Psi) \, dz
\]
Expectation-maximization

- the set
  \[ D = \{x_1, \ldots, x_n\} \]
  is called the incomplete data
- the set
  \[ D_c = \{(x_1, z_1), \ldots, (x_n, z_n)\} \]
  is called the complete data
- we never get to see it, otherwise the problem would be trivial (standard ML)
- EM solves the problem by iterating between two steps
Expectation-maximization

the basic idea is quite simple

1. start with an initial parameter estimate $Ψ^{(0)}$

2. **E-step**: given current parameters $Ψ^{(i)}$ and observations in $D$, “guess” what the values of the $z_i$ are

3. **M-step**: with the new $z_i$, we have a complete data problem, solve this problem for the parameters, i.e. compute $Ψ^{(i+1)}$

4. go to 2.

this can be summarized as

- **E-step**: estimate parameters
- **M-step**: fill in class assignments $z_i$
The Q function

**main idea:** don’t know what complete data likelihood is, but can compute its expected value given observed data

this is the Q function

$$Q(\psi; \psi^{(n)}) = E_{Z|X;\psi^{(i)}} \left[ \log P_{X,Z}(D, \{z_1, \ldots, z_N\}; \psi) | D \right]$$

and is a bit tricky:

- it is the expected value of likelihood with respect to complete data (joint X and Z)
- given that we observed incomplete data (X)
- note that the likelihood is a function of $\Psi$ (the parameters that we want to determine)
- but to compute the expected value we need to use the parameter values from the previous iteration (because we need a distribution for $Z|X$)
Expectation-maximization

**E-step:**
- given estimates $\Psi^{(n)} = \{\Psi^{(n)}_1, \ldots, \Psi^{(n)}_C\}$
- compute expected log-likelihood of complete data

$$Q(\Psi; \Psi^{(n)}) = E_{Z|X;\Psi^{(n)}} \left[ \log P_{X,Z}(\mathcal{D}, \{z_1, \ldots, z_N\}; \Psi) | \mathcal{D} \right]$$

**M-step:**
- find parameter set that maximizes this expected log-likelihood

$$\Psi^{(n+1)} = \arg \max_{\Psi} Q(\Psi; \Psi^{(n)})$$

let’s make this more concrete by looking at a toy example
Example

Toy model: \( X \) iid, \( Z \) iid, \( X_i \sim N(\mu, 1) \), \( Z_i \sim \lambda e^{-\lambda z} \), \( X \) independent of \( Z \)

\[
Q(\psi; \psi^{(n)}) = E_{Z|X;\psi^{(n)}} \left[ \log P_{X,Z}(D, \{z_1, \ldots, z_N\}; \psi) | D \right]
\]

\[
= E_{Z|X;\psi^{(n)}} \left[ -\sum_k \frac{(x_k - \mu)^2}{2} - \frac{N}{2} \log 2\pi - \lambda \sum_k z_k + N \log \lambda | D \right]
\]

\[
= -\sum_k \frac{(x_k - \mu)^2}{2} - \frac{N}{2} \log 2\pi - \lambda \sum_k E_{Z|X;\psi^{(n)}}[z_k|x_k] + N \log \lambda
\]

\[
= -\sum_k \frac{(x_k - \mu)^2}{2} - \frac{N}{2} \log 2\pi - \lambda \sum_k E_{Z_k;\psi^{(n)}}[z_k] + N \log \lambda
\]

\[
= -\sum_k \frac{(x_k - \mu)^2}{2} - \frac{N}{2} \log 2\pi - N\lambda E_{Z;\psi^{(n)}}[z] + N \log \lambda
\]

\[
= -\sum_k \frac{(x_k - \mu)^2}{2} - \frac{N}{2} \log 2\pi - N \frac{\lambda}{\lambda^{(n)}} + N \log \lambda
\]
Example

\[ \psi^{(n+1)} = \arg \max_{\psi} Q(\psi; \psi^{(n)}) \]

\[ Q(\psi; \psi^{(n)}) = -\sum_{k} \frac{(x_k - \mu)^2}{2} - \frac{N}{2} \log 2\pi - N \frac{\lambda}{\lambda^{(n)}} + N \log \lambda \]

\[ \frac{\partial Q}{\partial \mu} = 0 \Leftrightarrow \mu^{(n+1)} = \frac{1}{n} \sum_{k} x_k \]

\[ \frac{\partial Q}{\partial \lambda} = 0 \Leftrightarrow \lambda^{(n+1)} = \lambda^{(n)} \]

- this makes sense:
  - since hidden variables \( Z \) are independent of observed \( X \)
  - ML estimate of \( \mu \) is always the same: the sample mean, no dependence on \( z_i \)
  - ML estimate of \( \lambda \) is always the initial estimate \( \lambda^{(0)} \): since the observations are independent of the \( z_i \), we have no information on what \( \lambda \) should be, other than initial guess.

- note that model does not make sense, not EM solution
EM for mixtures

- we have also seen a more serious example

- ML estimation of the parameters of a mixture

\[ P_X(x; \psi) = \sum_{c=1}^{C} P_{X|Z}(x|c; \psi_c) \pi_c \]

- we noted that the right way to represent Z is to use a binary vector of size equal to the # of classes

\[ z \in \{e_1, \ldots, e_C\} \]

\[ e_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad \text{(jth position)} \]

- in which case complete data log-likelihood is linear on \( z_{ij} \)

\[ \log P_{X,Z}(D, \{z_1, \ldots, z_n\}; \psi) = \sum_{i,j} z_{ij} \log \left[ P_{X|Z}(x_i|e_j, \psi) \pi_j \right] \]
EM for mixtures

- the Q function becomes

\[ Q(\psi; \psi^{(n)}) = E_{Z|X; \psi^{(n)}}[\log P_{X,Z}(D, \{z_1, \ldots, z_N\}; \psi)|D] \]

\[ = \sum_{i,j} E_{Z|X; \psi^{(n)}}[z_{ij}|D] \log \left[ P_{X|Z}(x_i|e_j, \psi) \pi_j \right] \]

- i.e. to compute it we only need to find

\[ E_{Z|X; \psi^{(n)}}[z_{ij}|D], \ \forall i, j \]

- and since \( z_{ij} \) is binary and only depends on \( x_i \)

\[ E_{Z|X; \psi^{(n)}}[z_{ij}|D] = P_{Z|X}(z_{ij} = 1|x_i; \psi^{(n)}) = P_{Z|X}(e_j|x_i; \psi^{(n)}) \]

- the E-step reduces to computing the posterior probability of each point under each class!
Expectation-maximization

and the EM algorithm reduces to

1. E-step: Q function

\[ h_{ij} = P_{Z|X}(e_j|x_i; \psi^{(n)}) \]

\[ Q(\psi; \psi^{(n)}) = \sum_{i,j} h_{ij} \log \left[ P_{X|Z}(x_i|e_j, \psi) \pi_j \right] \]

2. M-step: solve the maximization, deriving a closed-form solution if there is one

\[ \psi^{(n+1)} = \arg \max_{\psi} \sum_{ij} h_{ij} \log \left[ P_{X|Z}(x_i|e_j, \psi) \pi_j \right] \]

under whatever constraints need to be considered, e.g.

\[ \sum_j \pi_j = 1 \]
Convergence of EM

- so far we have shown that EM
  - makes intuitive sense
  - leads to intuitive update equations
- the obvious question is: “how do we know that it converges to something useful?”
- it turns out that the proof is frustratingly simple
  - “it takes longer to understand what each term means than to do the proof itself”
- the only tool that we really need is Jensen’s inequality
- since this is such a useful inequality, let’s go over it in some detail
Concave functions

A function $f(x)$ is concave in $(a,b)$ if for all $x_1, x_2$ in $(a,b)$ and $\lambda$ in $[0,1]$

$$f[\lambda x_1 + (1 - \lambda)x_2] \geq \lambda f(x_1) + (1 - \lambda)f(x_2)$$
Jensen’s inequality

- if $f(x)$ is concave and $X$ a random variable then

$$E[f(x)] \leq f(E[x])$$

- the proof is easy for discrete distributions, where it can be done by induction

1. assume $X$ has two states with probability $p_1, p_2$. If $f$ is concave, by definition

$$E[f(x)] = p_1 f(x_1) + p_2 f(x_2)$$

$$\leq f[p_1 x_1 + p_2 x_2] = f(E[x])$$

2. assume that the inequality holds for all random variables of $n$ states, i.e.

$$\sum_{i=1}^{n} p_i f(x_i) \leq f \left( \sum_{i=1}^{n} p_i x_i \right)$$
Jensen’s inequality

- Assume \[ \sum_{i=1}^{n} p_i f(x_i) \leq f \left( \sum_{i=1}^{n} p_i x_i \right) \]

- Then for a r.v. with \( n+1 \) states

\[
E[f(x)] = \sum_{i=1}^{n+1} p_i f(x_i) = \sum_{i=1}^{n} p_i f(x_i) + p_{n+1} f(x_{n+1})
\]

\[
= (1 - p_{n+1}) \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} f(x_i) + p_{n+1} f(x_{n+1})
\]

\[
\leq (1 - p_{n+1}) f \left( \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} x_i \right) + p_{n+1} f(x_{n+1})
\]

and from the definition of concavity

\[
E[f(x)] \leq f \left( (1 - p_{n+1}) \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} x_i + p_{n+1} x_{n+1} \right)
\]
Jensen’s inequality

\[ E[f(x)] \leq f\left( \left(1 - p_{n+1}\right) \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} x_i + p_{n+1} x_{n+1} \right) \]

\[ = f\left( \sum_{i=1}^{n+1} p_i x_i \right) = f(E[x]) \]

in summary:

- inequality holds for r.v. with two states
- given that it holds for n states it also holds for n+1 states
- hence, by induction, it follows that for all discrete distributions and concave f(.)

\[ E[f(x)] \leq f(E[x]) \]

the result generalizes for the continuous case, but the proof is more complicated
EM convergence

we are now ready to show that EM converges

recall: the goal is to maximize \( \log P_X(\mathcal{D}; \psi) \)

using

\[
P_{X,Z}(\mathcal{D}, z; \psi) = P_{Z|X}(z|\mathcal{D}; \psi)P_X(\mathcal{D}; \psi)
\]

this can be written as

\[
\log P_X(\mathcal{D}; \psi) = \log P_{X,Z}(\mathcal{D}, z; \psi) - \log P_{Z|X}(z|\mathcal{D}; \psi)
\]

taking expectations on both sides and using the fact that the LHS does not depend on \( Z \)

\[
\log P_X(\mathcal{D}; \psi) = E_{Z|X; \psi(n)}[\log P_{X,Z}(\mathcal{D}, z; \psi)|\mathcal{D}] - E_{Z|X; \psi(n)}[\log P_{Z|X}(z|\mathcal{D}; \psi)|\mathcal{D}]
\]
EM convergence

and plugging in the definition of the Q function

\[
\log P_X(\mathcal{D}; \psi) = E_{Z|X; \psi(n)}[\log P_{X,Z}(\mathcal{D}, z; \psi)|\mathcal{D}] \\
- E_{Z|X; \psi(n)}[\log P_{Z|X}(z|\mathcal{D}; \psi)|\mathcal{D}] \\
= Q(\psi|\psi^{(n)}) + H(\psi|\psi^{(n)})
\]

where we have also introduced

\[
H(\psi|\psi^{(n)}) = -E_{Z|X; \psi(n)}[\log P_{Z|X}(z|\mathcal{D}; \psi)|\mathcal{D}] \\
= - \int P_{Z|X; \psi(n)}(z|\mathcal{D}; \psi^{(n)}) \log P_{Z|X}(z|\mathcal{D}; \psi) dz
\]
EM convergence

the key to proving convergence is this equation

\[ \log P_X(D; \psi) = Q(\psi|\psi^{(i)}) + H(\psi|\psi^{(i)}) \]

note, in particular, that

\[
\begin{align*}
\log P_X(D; \psi^{(n+1)}) - \log P_X(D; \psi^{(n)}) &= \\
&= Q(\psi^{(n+1)}|\psi^{(n)}) + H(\psi^{(n+1)}|\psi^{(n)}) \\
&\quad - [Q(\psi^{(n)}|\psi^{(n)}) + H(\psi^{(n)}|\psi^{(n)})] \\
&= Q(\psi^{(n+1)}|\psi^{(n)}) - Q(\psi^{(n)}|\psi^{(n)}) + H(\psi^{(n+1)}|\psi^{(n)}) - H(\psi^{(n)}|\psi^{(n)})
\end{align*}
\]
EM convergence

but, by definition of the M-step

\[ \psi^{(n+1)} = \underset{\psi}{\arg \max} Q(\psi | \psi^{(n)}) \]

it follows that

\[ Q(\psi^{(n+1)} | \psi^{(n)}) \geq Q(\psi^{(n)} | \psi^{(n)}) \]

and since

\[
\log P_X(\mathcal{D}; \psi^{(n+1)}) - \log P_X(\mathcal{D}; \psi^{(n)}) =
\]

\[
= \quad Q(\psi^{(n+1)} | \psi^{(n)}) - Q(\psi^{(n)} | \psi^{(n)})
\]

\[
+ H(\psi^{(n+1)} | \psi^{(n)}) - H(\psi^{(n)} | \psi^{(n)})
\]

we have

\[ \log P_X(\mathcal{D}; \psi^{(n+1)}) \geq \log P_X(\mathcal{D}; \psi^{(n)}) \]
EM convergence

we have

$$\log P_X(\mathcal{D}; \psi^{(n+1)}) \geq \log P_X(\mathcal{D}; \psi^{(n)})$$

if

$$H(\psi^{(n+1)}|\psi^{(n)}) \geq H(\psi^{(n)}|\psi^{(n)})$$

but, from

$$H(\psi|\psi^{(n)}) = -E_{Z|X;\psi^{(n)}}[\log P_{Z|X}(z|\mathcal{D}; \psi)|\mathcal{D}]$$

we have

$$H(\psi^{(n+1)}|\psi^{(n)}) - H(\psi^{(n)}|\psi^{(n)}) \leq -E_{Z|X;\psi^{(n)}}\left[\log \frac{P_{Z|X}(z|\mathcal{D}; \psi^{(n+1)})}{P_{Z|X}(z|\mathcal{D}; \psi^{(n)})}\right]_{\mathcal{D}}$$
EM convergence

and, since the log is a concave function, by Jensen’s
\[ E[f(x)] \leq f(E[x]) \]

\[
H(\psi^{(n+1)}|\psi^{(n)}) - H(\psi^{(n)}|\psi^{(n)}) \\
= -E_{Z|X;\psi^{(n)}} \left[ \log \frac{P_{Z|X}(z|D; \psi^{(n+1)})}{P_{Z|X}(z|D; \psi^{(n)})} \right]_D \\
\geq - \log E_{Z|X;\psi^{(n)}} \left[ \frac{P_{Z|X}(z|D; \psi^{(n+1)})}{P_{Z|X}(z|D; \psi^{(n)})} \right]_D \\
= - \log \int P_{Z|X;\psi^{(n)}}(z|D; \psi^{(n)}) \frac{P_{Z|X}(z|D; \psi^{(n+1)})}{P_{Z|X}(z|D; \psi^{(n)})} dz \\
= - \log 1 = 0
EM convergence

- this shows that

\[
\log P_X(\mathcal{D}; \psi^{(n+1)}) \geq \log P_X(\mathcal{D}; \psi^{(n)})
\]

- i.e. the log-likelihood of the incomplete data can only increase from iteration to iteration

- hence the algorithm converges

- note that there is no guarantee of convergence to a global minimum, only local
Geometric interpretation

one can also derive a geometric interpretation from

$$\log P_X(\mathcal{D}; \Psi) = Q(\Psi|\Psi^{(n)}) + H(\Psi|\Psi^{(n)})$$

by noting that

$$H(\Psi|\Psi^{(n)}) = -E_{Z|X;\Psi^{(n)}}[\log P_{Z|X}(z|\mathcal{D}; \Psi)|\mathcal{D}] = -\int P_{Z|X;\Psi^{(n)}}(z|\mathcal{D}; \Psi^{(n)}) \log P_{Z|X}(z|\mathcal{D}; \Psi) dz$$

is of the form

$$H(\Psi|\Psi^{(n)}) = -\int p_n(z) \log p(z) dz = \int p_n(z) \log \frac{p_n(z)}{p(z)} dz - \int p_n(z) \log p_n(z) dz$$
Geometric interpretation

is of the form

\[ H(\Psi | \Psi^{(n)}) = - \int p_n(z) \log p(z) dz \]

\[ = \int p_n(z) \log \frac{p_n(z)}{p(z)} dz - \int p_n(z) \log p_n(z) dz \]

\[ = KL[p_n||p] + H[p_n] \]

where \( KL[p||q] \) is the Kullback-Leibler divergence between \( p \) and \( q \), and \( H[p] \) the entropy of \( p \)

it can be shown that these two quantities are never negative, from which \( H(\Psi | \Psi^{(n)}) \geq 0 \) and

since

\[ \log P_X(D; \Psi) = Q(\Psi | \Psi^{(n)}) + H(\Psi | \Psi^{(n)}) \]
Geometric interpretation

we have

$$\log P_X(\mathcal{D}; \Psi) \geq Q(\psi|\psi^{(n)})$$

which means that the Q function is a lower bound to the log-likelihood of the observed data.

this allows an interpretation of the EM steps as

- E-step: lower-bound the observed log-likelihood
- M-step: maximize the lower bound
Geometric interpretation

consider next the difference between cost and bound

\[ \log P_X(D; \Psi) - Q(\Psi | \Psi^{(n)}) = H(\Psi | \Psi^{(n)}) \]

which can be written as

\[ H(\Psi | \Psi^{(n)}) = KL[p_n || p] + H[p_n] \]

with

\[ p_n(z) = P_{Z|X}(z|D; \Psi^{(n)}) \quad p(z) = P_{Z|X}(z|D; \Psi) \]

hence

\[ H(\Psi^{(n+1)} | \Psi^{(n)}) - H(\Psi^{(n)} | \Psi^{(n)}) = \]

\[ = KL[p_n || p_{n+1}] + H[p_n] - KL[p_n || p_n] - H[p_n] \]

\[ = KL[p_n || p_{n+1}] \geq 0 \]
Geometric interpretation

- note that since
  - by definition of M-step: \( Q(\psi^{(n+1)}|\psi^{(n)}) \geq Q(\psi^{(n)}|\psi^{(n)}) \)
  - by non-negativity of KL: \( H(\psi^{(n+1)}|\psi^{(n)}) \geq H(\psi^{(n)}|\psi^{(n)}) \)
- it follows that \( \log P_X(\mathcal{D}; \psi^{(n+1)}) \geq \log P_X(\mathcal{D}; \psi^{(n)}) \)
- EM converges without need for step sizes
- this is not the case for gradient ascent which uses the linear approximation
- if we move too far, there will be overshoot
Extensions

- note that in the proof we have really only used the fact that
  \[ Q(\psi^{(n+1)}|\psi^{(n)}) \geq Q(\psi^{(n)}|\psi^{(n)}) \]
- this means that
  - in M-step we do not necessarily need to maximize the Q-function
  - any step that increases it is sufficient

Generalized EM-algorithm

- E-step: compute
  \[ Q(\psi|\psi^{(n)}) = E_{\mathbf{Z}|\mathbf{X};\psi^{(n)}}[\log P_{\mathbf{X}|\mathbf{Z}(\mathbf{D}, \mathbf{z}; \psi)|\mathbf{D}] \]
- M-step: pick \( \psi^{(n+1)} \) such that
  \[ Q(\psi^{(n+1)}|\psi^{(n)}) \geq Q(\psi^{(n)}|\psi^{(n)}) \]
Extensions

Generalized EM-algorithm

- **E-step:** compute

\[ Q(\Psi | \Psi^{(n)}) = E_{Z|X;\Psi^{(n)}}[\log P_{X|Z}(\mathcal{D}, z; \Psi) | \mathcal{D}] \]

- **M-step:** pick \( \Psi^{(n+1)} \) such that

\[ Q(\Psi^{(n+1)} | \Psi^{(n)}) \geq Q(\Psi^{(n)} | \Psi^{(n)}) \]

very useful when M-step is itself non-trivial:

- e.g. if there is no closed-form solution one has to resort to numerical methods, like gradient ascent
- can be computationally intensive, lots of iterations per M-step
- in these cases, it is usually better to just perform a few iterations and move on to the next E-step
- no point in precisely optimizing M-step if everything is going to change when we compute the new E-step
MAP parameter estimates

- so far we have concentrated on ML estimation
- EM can be equally applied to obtain MAP estimates, with a straightforward extension
- recall that for MAP the goal is
  \[ \psi^* = \arg \max_{\psi} P_{\psi|X}(\psi | D) \]
  \[ = \arg \max_{\psi} P_{X|\psi}(D | \psi) P_{\psi}(\psi) \]
- this is not very different from ML, we just multiply by \( P_{\psi}(\psi) \)
- still a problem of estimation from incomplete data, with
  \[ P_{X|\psi}(D | \psi) = \int P_{X|Z,\psi}(D | z, \psi) P_{Z|\psi}(z | \psi) dz \]
MAP parameter estimates

- and there is a complete data posterior
  \[ P_{\Psi|X,Z}(\Psi|D, z) \]
- the E step is now to compute
  \[
  E_{Z|X,\Psi}[\log P_{\Psi|X,Z}(\Psi|D, z)|D, \Psi^{(n)}] = \\
  = E_{Z|X,\Psi}[\log P_{X,Z|\Psi}(D, z|\Psi)|D, \Psi^{(n)}] + \\
  + E_{Z|X,\Psi}[\log P_{\Psi}(\Psi)|D, \Psi^{(n)}] - \\
  - E_{Z|X,\Psi}[\log P_{X,Z}(D, z)|D, \Psi^{(n)}] \\
  = Q(\Psi|\Psi^{(n)}) + \log P_{\Psi}(\Psi) - \\
  - E_{Z|X,\Psi}[\log P_{X,Z}(D, z)|D, \Psi^{(n)}]
  \]
- note that the last term does not depend on \( \Psi \)
- does not affect M-step, we can drop it
MAP parameter estimates

- hence the E-step does not really change

E step: compute

\[ Q(\psi|\psi^{(n)}) = E_{Z|X,\psi}[\log P_{X,Z|\psi}(D, z|\psi)|D, \psi^{(n)}] \]

- and the M-step becomes

\[ \psi^{(n+1)} = \arg\max_{\psi} \left\{ Q(\psi|\psi^{(n)}) + \log P_\psi(\psi) \right\} \]

- this is the MAP-EM algorithm

- note that M-step looks like a standard Bayesian estimate procedure, and typically is

- e.g. for mixtures, it is equivalent to computing Bayesian estimates for each component, under “soft-assignments”
MAP parameter estimates

- in result, the estimates are similar to standard Bayesian estimates, but with
  - each point contributing to the parameters of all components
  - contribution weighted by the assignment probability

- but the important fact is that all the properties of Bayesian estimates still apply
  - conjugate priors
  - interpretation as additional, properly biased data, etc.

- this is a reason why our study of Bayesian estimation with simple models was so important
  - while a Gaussian is a fairly weak model
  - most densities can be approximated by a mixture of Gaussians
  - with EM we can generalize all we did quite easily
Any Questions?