Notes on Neal and Hinton's Generalized Expectation Maximization (GEM) Algorithm

Mark Johnson

Brown University

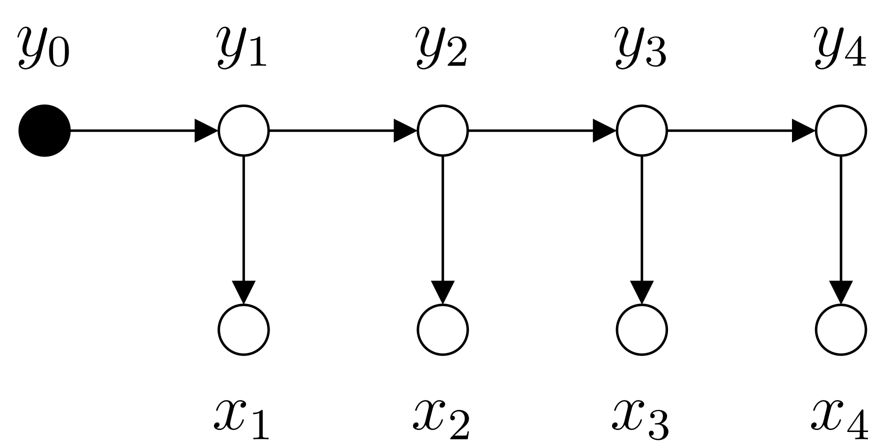
Febuary 2005, updated November 2008

Talk overview

- What kinds of problems does expectation maximization solve?
- An example of EM
- Relaxation, and proving that EM converges
- Sufficient statistics and EM
- The generalized EM algorithm

Hidden Markov Models

States (e.g., parts of speech)



Observations (e.g., words)

$$
\begin{aligned}
\mathrm{P}(\boldsymbol{Y}, \boldsymbol{X} | \boldsymbol{\theta}) &= \prod_{i=1}^{n} \mathrm{P}(Y_i | Y_{i-1}, \boldsymbol{\theta}) \mathrm{P}(X_i | Y_i, \boldsymbol{\theta}) \\
\mathrm{P}(y_i | y_{i-1}, \boldsymbol{\theta}) &= \theta_{y_i, y_{i-1}} \\
\mathrm{P}(x_i | y_i, \boldsymbol{\theta}) &= \theta_{x_i, y_i}
\end{aligned}
$$

Maximum likelihood estimation

- Given *visible data* $(\boldsymbol{y}, \boldsymbol{x})$, how can we estimate $\boldsymbol{\theta}$?
- Maximum likelihood principle:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \underset{\theta}{\operatorname{argmax}}\, L_{(y,x)}(\boldsymbol{\theta}), \text{ where:} \\
L_{(\boldsymbol{y},\boldsymbol{x})}(\boldsymbol{\theta}) &= \log \mathrm{P}_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x}) \;=\; \log \mathrm{P}(\boldsymbol{y}, \boldsymbol{x} | \boldsymbol{\theta})
\end{aligned}$$

- For a HMM, these are simple to calculate:

$$\begin{aligned}
\widehat{\theta}_{y_i,y_j} &= \frac{n_{y_i,y_j}(\boldsymbol{y}, \boldsymbol{x})}{\sum_{y_i'} n_{y_i',y_j}(\boldsymbol{y}, \boldsymbol{x})} \\
\widehat{\theta}_{x_i,y_i} &= \frac{n_{x_i,y_i}(\boldsymbol{y}, \boldsymbol{x})}{\sum_{x_i'} n_{x_i',y_i}(\boldsymbol{y}, \boldsymbol{x})}
\end{aligned}$$

ML estimation from hidden data

- Our model defines $\mathrm{P}(\boldsymbol{Y}, \boldsymbol{X})$, but our data only contains values for $\boldsymbol{X}$, i.e., the variable $\boldsymbol{Y}$ is *hidden*
  - HMM example: D only contains words $\boldsymbol{x}$ but not their labels $\boldsymbol{y}$
- Maximum likelihood principle still applies:

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\theta} L_{\boldsymbol{x}}(\boldsymbol{\theta}), \text{where:} \\ L_{\boldsymbol{x}}(\theta) &= \log \mathrm{P}(\boldsymbol{x}|\theta) = \log \sum_{\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}} \mathrm{P}(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) \end{aligned}$$

- But maximizing $L_{\boldsymbol{x}}(\boldsymbol{\theta})$ may now be a non-trivial problem!

What does Expectation Maximization do?

- Expectation Maximization (EM) is a *maximum likelihood estimation procedure* for problems with hidden variables
- EM is good for problems where:
  - our model $\mathrm{P}(Y, X|\theta)$ involves variables Y and X
  - our training data contains x but not y
  - maximizing $\mathrm{P}(x|\theta)$ is hard
  - maximizing $\mathrm{P}(y, x|\theta)$ is easy
- In HMM example: if training data consists of words x alone, and does not contain their labels

The EM algorithm

- The EM algorithm:
  - Guess an initial model $\theta^{(0)}$
  - For $t = 1, 2, \ldots,$ compute $Q^{(t)}(y)$ and $\theta^{(t)}$, where

$$\begin{aligned}
Q^{(t)}(y) &= \mathrm{P}(y|x, \theta^{(t-1)}) && \text{(E-step)} \\
\theta^{(t)} &= \operatorname*{argmax}_{\theta} E_{Y\sim Q^{(t)}}[\log \mathrm{P}(Y, x|\theta)] && \text{(M-step)} \\
&= \operatorname*{argmax}_{\theta} \sum_{y\in\mathcal{Y}} Q^{(t)}(y) \log \mathrm{P}(y, x|\theta) \\
&= \operatorname*{argmax}_{\theta} \prod_{y\in\mathcal{Y}} \mathrm{P}(y, x|\theta)^{Q^{(t)}(y)}
\end{aligned}$$

- $Q^{(t)}(y)$ is probability of "pseudo-data" y using model $\theta^{(t-1)}$
- $\theta^{(t)}$ is the MLE based on pseudo-data (y, x), where each (y, x) is weighted according to $Q^{(t)}(y)$

HMM example

- For a HMM, the EM formulae are:

$$
\begin{aligned}
Q^{(t)}(\boldsymbol{y}) &= \mathrm{P}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}^{(t-1)}) \\
&= \frac{\mathrm{P}(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}^{(t-1)})}{\sum_{\boldsymbol{y}\in\boldsymbol{\mathcal{Y}}} \mathrm{P}(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}^{(t-1)})} \\
\theta^{(t)}_{y_i,y_j} &= \frac{\sum_{\boldsymbol{y}\in\boldsymbol{\mathcal{Y}}} Q^{(t)}(\boldsymbol{y}) n_{y_i,y_j}(\boldsymbol{y}, \boldsymbol{x})}{\sum_{y'_i} \sum_{\boldsymbol{y}\in\boldsymbol{\mathcal{Y}}} Q^{(t)}(y) n_{y'_i,y_j}(\boldsymbol{y}, \boldsymbol{x})} \\
\theta^{(t)}_{x_i,y_i} &= \frac{\sum_{y\in\mathcal{Y}} Q^{(t)}(y) n_{x_i,y_i}(\boldsymbol{y}, \boldsymbol{x})}{\sum_{x'_i} \sum_{y\in\mathcal{Y}} Q^{(t)}(y) n_{x'_i,y_i}(\boldsymbol{y}, \boldsymbol{x})}
\end{aligned}
$$

EM converges — overview

- Neal and Hinton define a function $F(Q,\theta)$ where:
  - $Q(Y)$ is a probability distribution over the hidden variables
  - θ are the model parameters

$$\begin{aligned}
\operatorname*{argmax}_{\theta} \max_{Q} F(Q,\theta) &= \widehat{\theta}, \text{the MLE of } \theta \\
\max_{Q} F(Q,\theta) &= L_x(\theta), \text{the log likelihood of } \theta \\
\operatorname*{argmax}_{Q} F(Q,\theta) &= \mathrm{P}(Y|x,\theta) \text{ for all } \theta
\end{aligned}$$

- The EM algorithm is an *alternating maximization* of F

$$\begin{aligned}
Q^{(t)} &= \operatorname*{argmax}_{Q} F(Q,\theta^{(t-1)}) && \text{(E-step)} \\
\theta^{(t)} &= \operatorname*{argmax}_{\theta} F(Q^{(t)},\theta) && \text{(M-step)}
\end{aligned}$$

The EM algorithm converges

$$
\begin{aligned}
F(Q,\theta) &= E_{Y\sim Q}[\log \mathrm{P}(Y,x|\theta)] + H(Q) \\
&= L_x(\theta) - \mathrm{KL}(Q(Y)||\mathrm{P}(Y|x,\theta)) \\
H(Q) &= \text{entropy of } Q \\
L_x(\theta) &= \log \mathrm{P}(x|\theta) \;=\; \text{log likelihood of } \theta \\
\mathrm{KL}(Q||\mathrm{P}) &= \text{KL divergence between } Q \text{ and P} \\
Q^{(t)}(Y) &= \mathrm{P}(Y|x,\theta^{(t-1)}) \qquad = \operatorname*{argmax}_{Q} F(Q,\theta^{(t-1)}) \quad \text{(E-st} \\
\theta^{(t)} &= \operatorname*{argmax}_{\theta} E_{Y\sim Q^{(t)}}[\log \mathrm{P}(Y,x|\theta)] \;=\; \operatorname*{argmax}_{\theta} F(Q^{(t)},\theta) \quad \text{(M-s}
\end{aligned}
$$

- The maximum value of F is achieved at $\theta = \widehat{\theta}$ and $Q(Y) = \mathrm{P}(Y|x,\widehat{\theta})$.
- The sequence of F values produced by the EM algorithm is *non-decreasing* and *bounded above* by $L(\widehat{\theta})$.

Generalized EM

- Idea: anything that increases F gets you closer to $\widehat{\theta}$
- Idea: insert any additional operations you want into the EM algorithm so long as they don't decrease F
  - Update θ after each data item has been processed
  - Visit some data items more often than others
  - Only update some components of θ on some iterations

Incremental EM for factored models

- Data and model both factor: $Y = (Y_1, \ldots, Y_n), X = (X_1, \ldots, X_n)$

$$\mathrm{P}(Y, X|\theta) = \prod_{i=1}^{n} \mathrm{P}(Y_i, X_i|\theta)$$

- Incremental EM algorithm:
  - Initialize $\theta^{(0)}$ and $Q_i^{(0)}(Y_i)$ for $i = 1, \ldots, n$
  - E-step: Choose some data item i to be updated

$$\begin{aligned} Q_j^{(t)} &= Q_j^{(t-1)} \text{ for all } j \neq i \\ Q_i^{(t)}(Y_i) &= \mathrm{P}(Y_i|x_i, \theta^{(t-1)}) \end{aligned}$$

  - M-step:

$$\theta^{(t)} = \operatorname*{argmax}_{\theta} E_{Y \sim Q^{(t)}}[\log \mathrm{P}(Y, x|\theta)]$$

EM using sufficient statistics

- Model parameters θ estimated from *sufficient statistics* S:

$$(Y, X) \rightarrow S \rightarrow \theta$$

- In HMM example, pseudo-counts are sufficient statistics
- EM algorithm with sufficient statistics:

$$\begin{aligned} \tilde{s}^{(t)} &= E_{Y \sim \mathrm{P}(Y|x,\theta^{(t-1)})}[S] && \text{(E-step)} \\ \theta^{(t)} &= \text{maximum likelihood value for } \theta \text{ based on } \tilde{s}^{(t)} && \text{(M-step)} \end{aligned}$$

Incremental EM using sufficient statistics

- Incremental EM algorithm with sufficient statistics:

$$\boxed{(Y_i, X_i) \rightarrow S_i} \rightarrow S \rightarrow \theta \qquad\qquad S = \sum_i S_i$$

  - Initialize $\theta^{(0)}$ and $\tilde{s}_i^{(0)}$ for $i = 1, \ldots, n$
  - E-step: Choose some data item i to be updated

$$\begin{aligned} \tilde{s}_j^{(t)} &= \tilde{s}_j^{(t-1)} \text{ for all } j \neq i \\ \tilde{s}_i^{(t)} &= E_{Y_i \sim \mathrm{P}(Y_i|x_i,\theta^{(t-1)})}[S_i] \\ \tilde{s}^{(t)} &= \tilde{s}^{(t-1)} + (\tilde{s}_i^{(t)} - \tilde{s}_i^{(t-1)}) \end{aligned}$$

  - M-step:

$$\theta^{(t)} = \text{maximum likelihood value for } \theta \text{ based on } \tilde{s}^{(t)}$$

Conclusion

- The Expectation-Maximization algorithm is a general technique for using supervised maximum likelihood estimators to solve unsupervised estimation problems
- The E-step and the M-step can be viewed as steps of an *alternating maximization procedure*
  - The functional F is bounded above by the log likelihood
  - Each E and M step increases F

  $\Rightarrow$ Eventually the EM algorithm converges to a *local optimum* (not necessarily a global optimum)
- We can insert any steps we like into the EM algorithm so long as they do not decrease F

  $\Rightarrow$ Incremental versions of the EM algorithm