

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

Mark Johnson

Brown University

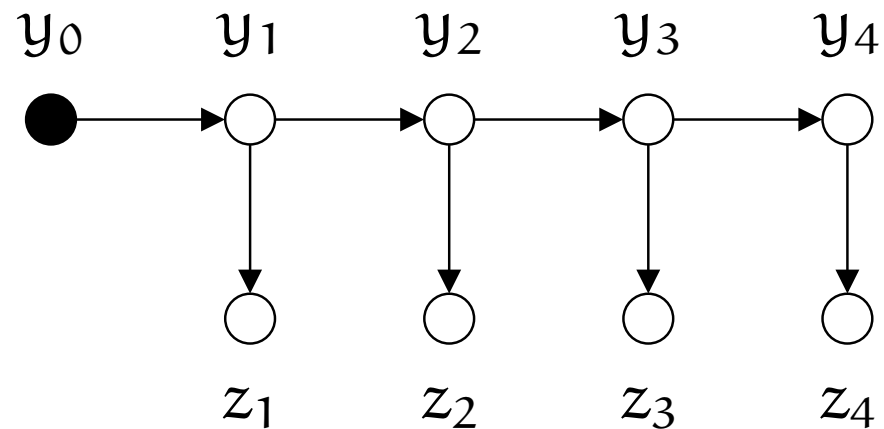
February 2005

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)

$$P(Y, Z|\theta) = \prod_{i=1}^n P(Y_i|Y_{i-1}, \theta)P(Z_i|Y_i, \theta)$$

$$P(y_i|y_{i-1}, \theta) = \theta_{y_i, y_{i-1}}$$

$$P(z_i|y_i, \theta) = \theta_{z_i, y_i}$$

Maximum likelihood estimation

- Given *visible data* (\mathbf{y}, \mathbf{z}) , how can we estimate θ ?
- Maximum likelihood principle:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta), \text{ where:}$$

$$L(\theta) = \log P_{\theta}(\mathbf{y}, \mathbf{z}) = P(\mathbf{y}, \mathbf{z}|\theta)$$

- For a HMM, these are simple to calculate:

$$\hat{\theta}_{\mathbf{y}_i, \mathbf{y}_{i-1}} = \frac{\#(\mathbf{y}_i, \mathbf{y}_{i-1})}{\#(\mathbf{y}_{i-1})}$$

$$\hat{\theta}_{\mathbf{z}_i, \mathbf{y}_i} = \frac{\#(\mathbf{z}_i, \mathbf{y}_i)}{\#(\mathbf{y}_i)}$$

ML estimation from hidden data

- Our model defines $P(Y, Z)$, but our data only contains values for Z , i.e., the variable Y is *hidden*
 - HMM example: D only contains words z but not parts of speech y

- Maximum likelihood principle still applies:

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta), \text{ where:}$$

$$L(\theta) = \log P(z|\theta) = \log \sum_{y \in \mathcal{Y}} P(y, z|\theta)$$

- But maximizing $L(\theta)$ 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 $P(Y, Z|\theta)$ involves variables Y and Z
 - our training data contains z but not y
 - maximizing $P(z|\theta)$ is hard
 - maximizing $P(y, z|\theta)$ is easy
- In HMM example: if training data consists of words z alone, and does not contain parts-of-speech

The EM algorithm

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

$$\tilde{P}^{(t)}(\mathbf{y}) = P(\mathbf{y}|\mathbf{z}, \theta^{(t-1)}) \quad (\text{E-step})$$

$$\theta^{(t)} = \operatorname{argmax}_{\theta} E_{Y \sim \tilde{P}^{(t)}} [\log P(Y, z|\theta)] \quad (\text{M-step})$$

$$= \operatorname{argmax}_{\theta} \sum_{\mathbf{y} \in \mathcal{Y}} \tilde{P}^{(t)}(\mathbf{y}) \log P(\mathbf{y}, z|\theta)$$

$$= \operatorname{argmax}_{\theta} \prod_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{y}, z|\theta)^{\tilde{P}^{(t)}(\mathbf{y})}$$

- $\tilde{P}^{(t)}(\mathbf{y})$ is probability of “pseudo-data” \mathbf{y} using model $\theta^{(t-1)}$
- $\theta^{(t)}$ is the MLE based on pseudo-data (\mathbf{y}, z) , where each (\mathbf{y}, z) is weighted according to $\tilde{P}^{(t)}(\mathbf{y})$

HMM example

- For a HMM, the EM formulae are:

$$\begin{aligned}\tilde{P}^{(t)}(\mathbf{y}) &= P(\mathbf{y}|\mathbf{z}, \theta^{(t-1)}) \\ &= \frac{P(\mathbf{y}, \mathbf{z}|\theta^{(t-1)})}{\sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{y}, \mathbf{z}|\theta^{(t-1)})} \\ \theta_{\mathbf{y}_i, \mathbf{y}_{i-1}}^{(t)} &= \frac{\sum_{\mathbf{y} \in \mathcal{Y}} \#(\mathbf{y}_i, \mathbf{y}_{i-1}) \tilde{P}^{(t)}(\mathbf{y})}{\sum_{\mathbf{y} \in \mathcal{Y}} \#(\mathbf{y}_{i-1}) \tilde{P}^{(t)}(\mathbf{y})} \\ \theta_{\mathbf{z}_i, \mathbf{y}_i}^{(t)} &= \frac{\sum_{\mathbf{y} \in \mathcal{Y}} \#(\mathbf{z}_i, \mathbf{y}_i) \tilde{P}^{(t)}(\mathbf{y})}{\sum_{\mathbf{y} \in \mathcal{Y}} \#(\mathbf{y}_i) \tilde{P}^{(t)}(\mathbf{y})}\end{aligned}$$

EM converges — overview

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

$$\operatorname{argmax}_{\theta} \max_{\tilde{P}} F(\tilde{P}, \theta) = \hat{\theta}, \text{ the MLE of } \theta$$

$$\max_{\tilde{P}} F(\tilde{P}, \theta) = L(\theta), \text{ the log likelihood of } \theta$$

$$\operatorname{argmax}_{\tilde{P}} F(\tilde{P}, \theta) = P(Y|z, \theta) \text{ for all } \theta$$

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

$$\tilde{P}^{(t)} = \operatorname{argmax}_{\tilde{P}} F(\tilde{P}, \theta^{(t-1)}) \quad (\text{E-step})$$

$$\theta^{(t)} = \operatorname{argmax}_{\theta} F(\tilde{P}^{(t)}, \theta) \quad (\text{M-step})$$

The EM algorithm converges

$$\begin{aligned} F(\tilde{P}, \theta) &= E_{Y \sim \tilde{P}}[\log P(Y, z|\theta)] + H(\tilde{P}) \\ &= L(\theta) - D_Y(\tilde{P}(Y) \| P(Y|z, \theta)) \end{aligned}$$

$$H(\tilde{P}) = \text{entropy of } \tilde{P}$$

$$L(\theta) = \log P(z|\theta) = \text{log likelihood of } \theta$$

$$D(\tilde{P} \| P) = \text{KL divergence between } \tilde{P} \text{ and } P$$

$$\tilde{P}^{(t)}(Y) = P(Y|z, \theta^{(t-1)}) = \underset{\tilde{P}}{\operatorname{argmax}} F(\tilde{P}, \theta^{(t-1)}) \quad (\text{E-step})$$

$$\theta^{(t)} = \underset{\theta}{\operatorname{argmax}} E_{Y \sim \tilde{P}^{(t)}}[\log P(Y, z|\theta)] = \underset{\theta}{\operatorname{argmax}} F(\tilde{P}^{(t)}, \theta) \quad (\text{M-step})$$

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

Generalized EM

- Idea: anything that increases F gets you closer to $\hat{\theta}$
- 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: $Y = (Y_1, \dots, Y_n), Z = (Z_1, \dots, Z_n)$

$$P(Y, Z|\theta) = \prod_{i=1}^n P(Y_i, Z_i|\theta)$$

- Incremental EM algorithm:

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

$$\begin{aligned}\tilde{P}_j^{(t)} &= \tilde{P}_j^{(t-1)} \text{ for all } j \neq i \\ \tilde{P}_i^{(t)}(Y_i) &= P(Y_i|z_i, \theta^{(t-1)})\end{aligned}$$

- M-step:

$$\theta^{(t)} = \operatorname{argmax}_{\theta} E_{Y \sim \tilde{p}^{(t)}} [\log P(Y, z|\theta)]$$

EM using sufficient statistics

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

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

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

$$\tilde{s}^{(t)} = \mathbb{E}_{Y \sim \text{Pr}(Y|Z, \theta^{(t-1)})}[S] \quad (\text{E-step})$$

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

Incremental EM using sufficient statistics

- Incremental EM algorithm with sufficient statistics:

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

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

$$\begin{aligned}\tilde{s}_j^{(t)} &= \tilde{s}_j^{(t-1)} \text{ for all } j \neq i \\ \tilde{s}_i^{(t)} &= \mathbb{E}_{Y_i \sim \Pr(Y_i | z_i, \theta^{(t-1)})} [S_i] \\ \tilde{s}^{(t)} &= \tilde{s}^{(t-1)} + (\tilde{s}_i^{(t)} - \tilde{s}_i^{(t-1)})\end{aligned}$$

- M-step:

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