A gentle introduction to Expectation Maximization

Mark Johnson
Brown University

November 2009
Outline

What is Expectation Maximization?

Mixture models and clustering

EM for sentence topic modeling
Why Expectation Maximization?

• **Expectation Maximization** (EM) is a general approach for solving problems involving *hidden* or *latent variables* $Y$

• Goal: learn the parameter vector $\theta$ of a model $P_\theta(X, Y)$ from training data $D = (x_1, \ldots, x_n)$ consisting of samples from $P_\theta(X)$, i.e., $Y$ is hidden

• Maximum likelihood estimate using $D$:

$$\hat{\theta} = \arg\max_\theta L_D(\theta) = \arg\max_\theta \prod_{i=1}^{n} \sum_{y \in Y} P_\theta(x_i, y)$$

• EM is useful when directly optimizing $L_D(\theta)$ is intractible, but computing MLE from fully-observed data $D' = ((x_1, y_1), \ldots, (x_n, y_n))$ is easy
Outline

What is Expectation Maximization?

Mixture models and clustering

EM for sentence topic modeling
Mixture models and clustering

- A **mixture model** is a linear combination of models

\[
P(X = x) = \sum_{y \in \mathcal{Y}} P(Y = y) \, P(X = x | Y = y),\ 
\]

where:

- \( y \in \mathcal{Y} \) identifies the **mixture component**, 
- \( P(y) \) is the probability of generating mixture component \( y \), and 
- \( P(x | y) \) is the distribution associated with mixture component \( y \)

- In clustering, \( \mathcal{Y} = \{1, \ldots, m\} \) are the **cluster labels**
  - After learning \( P(y) \) and \( P(x | y) \), compute cluster probabilities for data item \( x_i \) as follows:

\[
P(Y = y | X = x_i) = \frac{P(Y = y) \, P(X = x_i | Y = y)}{\sum_{y' \in \mathcal{Y}} P(Y = y') \, P(X = x_i | Y = y')}\]
Mixtures of multinomials (1)

- $\mathcal{Y} = \{1, \ldots, m\}$, i.e., $m$ different clusters
  - $Y$ is coin identity in coin-tossing game
  - $Y$ is sentence topic in sentence clustering application
- $\mathcal{X} = \mathcal{U}^\ell$, i.e., each observation is a sequence $x = (u_1, \ldots, u_\ell)$, where each $u_k \in \mathcal{U}$
  - $\mathcal{U} = \{H, T\}$, $x$ is one sequence of coin tosses from same (unknown) coin
  - $\mathcal{U}$ is the vocabulary, $x$ is a sentence (sequence of words)
- Assume each $u_k$ is generated i.i.d. given $y$, so models have parameters:
  - $P(Y = y) = \pi_y$, i.e., probability of picking cluster $y$
  - $P(U_k = u | Y = y) = \varphi_{u|y}$, i.e., probability of generating a $u$ in cluster $y$
Mixtures of multinomials (2)

\[
\begin{align*}
P(Y = y) &= \pi_y \\
P(U_k = u | Y = y) &= \varphi_{u|y} \\
P(X = x, Y = y) &= \pi_y \prod_{k=1}^{\ell} \varphi_{u_k|y} \\
&= \pi_y \prod_{u \in U} c_u(x) \varphi_{u_k|y}
\end{align*}
\]

where \( x = (u_1, \ldots, u_\ell) \), and 
\( c_u(x) \) is number of times \( u \) appears in \( x \).
Coin-tossing example

\[ \pi_1 = \pi_2 = 0.5 \]
\[ \varphi_{H|1} = 0.1; \quad \varphi_{T|1} = 0.9 \]
\[ \varphi_{H|2} = 0.8; \quad \varphi_{T|2} = 0.2 \]

\[ P(X = HTHH, Y = 1) = \pi_1 \varphi_{H|1}^3 \varphi_{T|1}^1 = 0.00045 \]
\[ P(X = HTHH, Y = 2) = \pi_2 \varphi_{H|2}^3 \varphi_{T|2}^1 = 0.0512 \]
\[ P(X = HTHH) = \pi_1 \varphi_{H|1}^3 \varphi_{T|1}^1 + \pi_2 \varphi_{H|2}^3 \varphi_{T|2}^1 \]
\[ = 0.05165, \text{ so:} \]
\[ P(Y = 1 \mid X = HTHH) = \frac{P(X = HTHH, Y = 1)}{P(X = HTHH)} \]
\[ = 0.008712 \]
\[ P(Y = 2 \mid X = HTHH) = 0.9912 \]
Estimation from visible data

- Given visible data how would we estimate $\pi$ and $\varphi$?
- Data $D' = ((x_1, y_1), \ldots, (x_n, y_n))$, where each $x_i = (u_{i,1}, \ldots, u_{i,\ell})$

**Sufficient statistics** for estimating multinomial mixture:
- $n_y = \sum_{i=1}^n I(y, y_i)$, i.e., number of times cluster $y$ is seen
- $n_{u,y} = \sum_{i=1}^n c_u(x_i)I(y, y_i)$, i.e., number of times $u$ is seen in cluster $y$, where $c_u(x)$ is the number of times $u$ appears in $x$

- Maximum likelihood estimates:

\[
\hat{\pi}_y = \frac{n_y}{n}, \quad \hat{\varphi}_{u|y} = \frac{n_{u,y}}{\sum_{u' \in \mathcal{U}} n_{u',y}}
\]
Estimation from *hidden* data (1)

- Data \( D = (x_1, \ldots, x_n) \), where each \( x_i = (u_{i,1}, \ldots, u_{i,\ell}) \)
- Log likelihood of hidden data:

\[
\log L_D(\pi, \varphi) = \sum_{i=1}^{n} \log \sum_{y \in Y} \pi_y \prod_{u \in U} \varphi_{u|y}^{c_u(x_i)}
\]

- Imposing Lagrange multipliers and setting the derivative to zero, we can show:

\[
\hat{\pi}_y = \frac{E[n_y]}{n}; \quad \hat{\varphi}_{u|y} = \frac{E[n_{u,y}]}{\sum_{u' \in U} E[n_{u',y}]}, \text{ where:}
\]

\[
E[n_y] = \sum_{i=1}^{n} P_{\hat{\pi},\hat{\varphi}}(Y = y \mid X = x_i)
\]

\[
E[n_{u,y}] = \sum_{i=1}^{n} c_u(x_i) P_{\hat{\pi},\hat{\varphi}}(Y = y \mid X = x_i)
\]
Estimation from \textit{hidden} data (2)

\[ \hat{\pi}_y = \frac{E[n_y]}{n}, \quad \hat{\phi}_{u|y} = \frac{E[n_{u,y}]}{\sum_{u' \in U} E[n_{u',y}]} \]

where:

\[ E[n_y] = \sum_{i=1}^{n} P_{\hat{\pi},\hat{\phi}}(Y = y \mid X = x_i) \]

\[ E[n_{u,y}] = \sum_{i=1}^{n} c_u(x_i) P_{\hat{\pi},\hat{\phi}}(Y = y \mid X = x_i) \]

- Unlike in the visible data case, these are not a \textit{closed-form} solution for \( \hat{\pi} \) or \( \hat{\phi} \), as \( E[n_y] \) and \( E[n_{u,y}] \) involve \( \hat{\pi} \) and \( \hat{\phi} \)
- But they do suggest a \textit{fixed-point calculation procedure}
**EM for multinomial mixtures**

- Guess initial values $\pi^{(0)}$ and $\phi^{(0)}$
- For iterations $t = 1, 2, 3, \ldots$ do:
  - **E-step:** calculate expected values of sufficient statistics
    
    $E[n_y] = \sum_{i=1}^{n} P_{\pi^{(t-1)},\phi^{(t-1)}}(Y = y \mid X = x_i)$
    
    $E[n_{u,y}] = \sum_{i=1}^{n} c_u(x_i) P_{\pi^{(t-1)},\phi^{(t-1)}}(Y = y \mid X = x_i)$
  
  - **M-step:** update model based on sufficient statistics
    
    $\pi_y^{(t)} = \frac{E[n_y]}{n}$
    
    $\phi_{u|y}^{(t)} = \frac{E[n_{u,y}]}{\sum_{u' \in U} E[n_{u',y}]}$
Summary of the model

\[ P(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{\sum_{y' \in \mathcal{Y}} P(Y = y', X = x)} \]

\[ P_{\pi, \phi}(X = x, Y = y) = \pi_y \prod_{u \in \mathcal{U}} \phi_{u|y}^{c_u(x)}, \text{ where:} \]

\[ c_u(x) = \text{the number of times } u \text{ appears in } x \]
Outline

What is Expectation Maximization?

Mixture models and clustering

EM for sentence topic modeling
Homework hints

• The fact that different sentences have different lengths doesn’t affect the calculation
• $c_u(x_i)$ is the number of times word $u$ appears in sentence $x_i$
• You can initialize $\pi$ with a uniform distribution, but you’ll need to initialize $\varphi^{(0)}$ to break symmetry, e.g., by adding a random number of about $10^{-4}$
• You should compute the log likelihood at each iteration (it’s easy to do this as a by-product of the expectation calculations)
  ▶ There is a theorem that says the log likelihood never decreases on each EM step
  ▶ If your log likelihood decreases, then you have a bug!