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 θ of a model $\mathrm{P}_\theta(X, Y)$ from training data $D = (x_1, \ldots, x_n)$ consisting of samples from $\mathrm{P}_\theta(X)$, i.e., *Y is hidden*
- Maximum likelihood estimate using D:

$$\hat{\theta} \;=\; \underset{\theta}{\operatorname{argmax}}\, L_D(\theta) \;=\; \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} \sum_{y \in \mathcal{Y}} \mathrm{P}_\theta(x_i, y)$$

- 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

$$\mathrm{P}(X = x) \quad = \quad \sum_{y \in \mathcal{Y}} \mathrm{P}(Y = y)\, \mathrm{P}(X = x | Y = y), \text{ where:}$$

  $y \in \mathcal{Y}$ identifies the *mixture component*,
  $\mathrm{P}(y)$ is probability of generating mixture component y, and
  $\mathrm{P}(x|y)$ is distribution associated with mixture component y
- In clustering, $\mathcal{Y} = \{1, \ldots, m\}$ are the *cluster labels*
  - After learning $\mathrm{P}(y)$ and $\mathrm{P}(x|y)$, compute cluster probabilities for data item x_i as follows:

$$\mathrm{P}(Y = y | X = x_i) \quad = \quad \frac{\mathrm{P}(Y = y)\, \mathrm{P}(X = x_i | Y = y)}{\sum_{y' \in \mathcal{Y}} \mathrm{P}(Y = y')\, \mathrm{P}(X = x_i | Y = y')}$$

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 $u_k \in \mathcal{U}$
  - $\mathcal{U} = \{\mathsf{H}, \mathsf{T}\}$, 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:
  - $\mathrm{P}(Y = y) = \pi_y$, i.e., probability of picking cluster y
  - $\mathrm{P}(U_k = u | Y = y) = \varphi_{u|y}$, i.e., probability of generating a u in cluster y

Mixtures of multinomials (2)

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

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

Coin-tossing example

$$\begin{aligned}
\pi_1 &= \pi_2 = 0.5 \\
\varphi_{\mathsf{H}|1} &= 0.1; \qquad \varphi_{\mathsf{T}|1} = 0.9 \\
\varphi_{\mathsf{H}|2} &= 0.8; \qquad \varphi_{\mathsf{T}|2} = 0.2
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}(X = \mathsf{HTHH}, Y = 1) &= \pi_1\, \varphi^3_{\mathsf{H}|1}\, \varphi^1_{\mathsf{T}|1} = 0.00045 \\
\mathrm{P}(X = \mathsf{HTHH}, Y = 2) &= \pi_2\, \varphi^3_{\mathsf{H}|2}\, \varphi^1_{\mathsf{T}|2} = 0.0512 \\
\mathrm{P}(X = \mathsf{HTHH}) &= \pi_1\, \varphi^3_{\mathsf{H}|1}\, \varphi^1_{\mathsf{T}|1} + \pi_2\, \varphi^3_{\mathsf{H}|2}\, \varphi^1_{\mathsf{T}|2} \\
&= 0.05165, \text{ so:} \\
\mathrm{P}(Y = 1 \mid X = \mathsf{HTHH}) &= \frac{\mathrm{P}(X = \mathsf{HTHH}, Y = 1)}{\mathrm{P}(X = \mathsf{HTHH})} \\
&= 0.008712 \\
\mathrm{P}(Y = 2 \mid X = \mathsf{HTHH}) &= 0.9912
\end{aligned}$$

Estimation from *visible* data

- Given visible data how would we estimate π and φ?
- 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:
  - $n_y = \sum_{i=1}^{n} \mathbb{I}(y, y_i)$, i.e., number of times cluster y is seen
  - $n_{u,y} = \sum_{i=1}^{n} c_u(x_i)\mathbb{I}(y, y_i)$, 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:

$$\begin{aligned} \widehat{\pi}_y &= \frac{n_y}{n} \\ \widehat{\varphi}_{u|y} &= \frac{n_{u,y}}{\sum_{u' \in \mathcal{U}} n_{u',y}} \end{aligned}$$

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:

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

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

$$\widehat{\pi}_y = \frac{\mathrm{E}[n_y]}{n}; \qquad \widehat{\varphi}_{u|y} = \frac{\mathrm{E}[n_{u,y}]}{\sum_{u' \in \mathcal{U}} \mathrm{E}[n_{u',y}]}, \text{ where:}$$

$$\mathrm{E}[n_y] = \sum_{i=1}^{n} \mathrm{P}_{\widehat{\pi},\widehat{\varphi}}(Y = y \mid X = x_i)$$

$$\mathrm{E}[n_{u,y}] = \sum_{i=1}^{n} c_u(x_i)\, \mathrm{P}_{\widehat{\pi},\widehat{\varphi}}(Y = y \mid X = x_i)$$

Estimation from *hidden* data (2)

$$
\begin{aligned}
\widehat{\pi}_y &= \frac{\mathrm{E}[n_y]}{n}; \qquad \widehat{\varphi}_{u|y} = \frac{\mathrm{E}[n_{u,y}]}{\sum_{u' \in \mathcal{U}} \mathrm{E}[n_{u',y}]}, \text{ where:} \\
\mathrm{E}[n_y] &= \sum_{i=1}^{n} \mathrm{P}_{\widehat{\pi},\widehat{\varphi}}(Y = y \mid X = x_i) \\
\mathrm{E}[n_{u,y}] &= \sum_{i=1}^{n} c_u(x_i)\, \mathrm{P}_{\widehat{\pi},\widehat{\varphi}}(Y = y \mid X = x_i)
\end{aligned}
$$

- Unlike in the visible data case, these are not a *closed-form* solution for $\widehat{\pi}$ or $\widehat{\varphi}$, as $\mathrm{E}[n_y]$ and $\mathrm{E}[n_{u,y}]$ involve $\widehat{\pi}$ and $\widehat{\varphi}$
- But they do suggest a *fixed-point calculation procedure*

EM for multinomial mixtures

- Guess initial values $\pi^{(0)}$ and $\varphi^{(0)}$
- For iterations $t = 1, 2, 3, \ldots$ do:
  - *E-step:* calculate expected values of sufficient statistics

$$
\begin{aligned}
\mathrm{E}[n_y] &= \sum_{i=1}^{n} \mathrm{P}_{\pi^{(t-1)},\varphi^{(t-1)}}(Y = y \mid X = x_i) \\
\mathrm{E}[n_{u,y}] &= \sum_{i=1}^{n} c_u(x_i)\, \mathrm{P}_{\pi^{(t-1)},\varphi^{(t-1)}}(Y = y \mid X = x_i)
\end{aligned}
$$

  - *M-step:* update model based on sufficient statistics

$$
\begin{aligned}
\pi_y^{(t)} &= \frac{\mathrm{E}[n_y]}{n} \\
\varphi_{u|y}^{(t)} &= \frac{\mathrm{E}[n_{u,y}]}{\sum_{u' \in \mathcal{U}} \mathrm{E}[n_{u',y}]}
\end{aligned}
$$

Summary of the model

$$\begin{aligned}
P(Y = y \mid X = x) &= \frac{P(Y = y, X = x)}{\sum_{y' \in \mathcal{Y}} P(Y = y', X = x)} \\
P_{\pi,\varphi}(X = x, Y = y) &= \pi_y \prod_{u \in \mathcal{U}} \varphi_{u|y}^{c_u(x)}, \text{ where:} \\
c_u(x) &= \text{the number of times } u \text{ appears in } x
\end{aligned}$$

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 π 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!