Expectation Maximization

Brandon Caie and Jonny Coutinho
EM algorithm provides a general approach to learning in presence of unobserved variables.

In many practical learning settings, only a subset of relevant features or variables might be observable.
– Eg: Hidden Markov, Bayesian Belief Networks
Simple Example: Coin Flipping

• Suppose you have 2 coins, A and B, each with a certain bias of landing heads, $\theta_A, \theta_B$.

• Given data sets $X_A = \{x_{1,A}, \ldots, x_{m_A,A}\}$ and $X_B = \{x_{1,B}, \ldots, x_{m_B,B}\}$

  Where $x_{i,j} = \begin{cases} 
  1 & \text{if heads} \\
  0 & \text{otherwise}
\end{cases}$

• No hidden variables – easy solution. $\theta_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_{i,j}$; sample mean
Simplified MLE

Goal: determine coin parameters without knowing the identity of each data set’s coin.

Solution: Expectation-maximization
What if you were given the same dataset of coin flip results, but no coin identities defining the datasets?

Here: \( X = \{x_1, \ldots, x_m\} \); the observed variable

\[
Z = \begin{pmatrix}
Z_{1,1} & \ldots & Z_{m,1} \\
\ldots & Z_{i,j} & \ldots \\
Z_{1,k} & \ldots & Z_{m,k}
\end{pmatrix}
\]

where \( z_{i,j} = \begin{cases} 
1; & \text{if } x_i \text{ is from } j^{th} \text{ coin} \\
0; & \text{otherwise}
\end{cases} \)

But \( Z \) is not known. (Ie: ‘hidden’ / ‘latent’ variable)
EM Algorithm

0) Initialize some arbitrary hypothesis of parameter values ($\theta$):
   $\theta = \{ \theta_1, ..., \theta_k \}$
   coin flip example: $\theta = \{\theta_A, \theta_B\} = \{0.6, 0.5\}$

1) Expectation (E-step)
   
   $E[z_{i,j}] = \frac{p(x = x_i | \theta = \theta_j)}{\sum_{n=1}^{k} p(x = x_i | \theta = \theta_n)}$

2) Maximization (M-step)

   $\theta_j = \frac{\sum_{i=1}^{m} E[z_{i,j}] x_i}{\sum_{i=1}^{m} E[z_{i,j}]}$

   If $z_{i,j}$ is known:

   $\theta_j = \frac{\sum_{j=1}^{m_j} x_i}{m_j}$
EM- Coin Flip example

- Initialize $\theta_A$ and $\theta_B$ to chosen value
  - Ex: $\theta_A = 0.6, \theta_B = 0.5$
- Compute a probability distribution of possible completions of the data using current parameters
EM- Coin Flip example

Set 1  H T T T H H T T H H

• What is the probability that I observe 5 heads and 5 tails in coin A and B given the initializing parameters $\theta_A=0.6$, $\theta_B=0.5$?
• Compute likelihood of set 1 coming from coin A or B using the binomial distribution with mean probability $\theta$ on $n$ trials with $k$ successes

$$p(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

• Likelihood of “A”=0.00079
• Likelihood of “B”=0.00097
• Normalize to get probabilities $\Rightarrow$ A=0.45, B=0.55
The E-step

- $P(Coin=A) = 0.45$; $P(Coin=B) = 0.55$
- Estimate how these probabilities can account for the number of observed heads and tails in the coin flip set
- Repeat for each data set

<table>
<thead>
<tr>
<th>Coin A</th>
<th>Coin B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\approx 2.2$ H, $2.2$ T</td>
<td>$\approx 2.8$ H, $2.8$ T</td>
</tr>
<tr>
<td>$\approx 7.2$ H, $0.8$ T</td>
<td>$\approx 1.8$ H, $0.2$ T</td>
</tr>
<tr>
<td>$\approx 5.9$ H, $1.5$ T</td>
<td>$\approx 2.1$ H, $0.5$ T</td>
</tr>
<tr>
<td>$\approx 1.4$ H, $2.1$ T</td>
<td>$\approx 2.6$ H, $3.9$ T</td>
</tr>
<tr>
<td>$\approx 4.5$ H, $1.9$ T</td>
<td>$\approx 2.5$ H, $1.1$ T</td>
</tr>
<tr>
<td>$\approx 21.3$ H, $8.6$ T</td>
<td>$\approx 11.7$ H, $8.4$ T</td>
</tr>
</tbody>
</table>
### The M-step

<table>
<thead>
<tr>
<th>Coin A</th>
<th>Coin B</th>
</tr>
</thead>
<tbody>
<tr>
<td>≈ 2.2 H, 2.2 T</td>
<td>≈ 2.8 H, 2.8 T</td>
</tr>
<tr>
<td>≈ 7.2 H, 0.8 T</td>
<td>≈ 1.8 H, 0.2 T</td>
</tr>
<tr>
<td>≈ 5.9 H, 1.5 T</td>
<td>≈ 2.1 H, 0.5 T</td>
</tr>
<tr>
<td>≈ 1.4 H, 2.1 T</td>
<td>≈ 2.6 H, 3.9 T</td>
</tr>
<tr>
<td>≈ 4.5 H, 1.9 T</td>
<td>≈ 2.5 H, 1.1 T</td>
</tr>
<tr>
<td>≈ 21.3 H, 8.6 T</td>
<td>≈ 11.7 H, 8.4 T</td>
</tr>
</tbody>
</table>

\[
\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71
\]

\[
\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58
\]
1. Choose starting parameters
2. Estimate probability using these parameters that each data set \( (x_i) \) came from \( j^{th} \) coin \( (E[z_{i,j}]) \)
3. Use these probability values \( (E[z_{i,j}]) \) as weights on each data point when computing a new \( \theta_j \) to describe each distribution
4. Summate these expected values, use maximum likelihood estimation to derive new parameter values to repeat process
Gaussian Mixture Models

- When data is continuous, can be described by Normal Distributions
Gaussian Mixture Models

- Cluster data as Gaussians, with parameters: \((\mu_j, \sigma_j^2, \pi_j)\)

\[
p(z = j) = \pi_j
\]

\[
p(x|z = j) = N(x; \mu_i, \sigma_i^2)
\]
EM algorithm in Gaussian Mixtures

Step 0) Initialize $\theta = \{\mu_1, \ldots, \mu_k\} \ \{\sigma_1^2, \ldots, \sigma_k^2\} \ \{\pi_1, \ldots, \pi_k\}$ (assuming $k$ clusters)

Step 1) Expectation: compute $r_{i,j}$ for each $x_i$

$$r_{i,j} = \frac{\pi_{i,j} \ p(x|z = j)}{\sum_{n=1}^{k} \pi_{i,n} \ p(x|z = n)}$$
Step 2) Maximization:

\[ m_j = \sum_i r_{i,j} \]
\[ \pi_j = \frac{m_j}{m} \]
\[ \mu_j = \frac{1}{m_j} \sum_i r_{i,j} x_i \]
\[ \sigma_j^2 = \frac{1}{m_j} \sum_i r_{i,j} (x_i - \mu_j)^2 \]
Example of EM in Gaussian Mixtures

From P. Smyth
ICML 2001

Alexander Ihler, https://www.youtube.com/watch?v=qMTuMa86NzU
EM ITERATION 1

From P. Smyth
ICML 2001
Overfitting through convergence
Initializing Parameters

- Hidden variables and incomplete data lead to more complex likelihood functions with many local optima
- Since EM only solves for a single local optima, choosing a good initial parameter estimation is critical
- Strategies to improve initialization
  - Multiple random restarts
  - Use prior knowledge
  - Output of a simpler, though less robust algorithm
Resources

- Matlab EM Algorithm
- Tom Mitchell- Machine Learning: Chapter 6 (on lab wiki)
- EM Algorithm Derivation, Convergence, Hidden Markov and GMM Applications
- Nature Review Article