Maximum likelihood
Gaussian mixtures using expectation maximization

Jarno Seppänen
2008-02-27
**Bayes vs. maximum likelihood**

- **Model**: what we are observing (data generating process)
- **Likelihood**: model applied to specific data
- **Prior distribution**: before-the-fact beliefs of different outcomes
- **Posterior distribution**: after-the-fact distribution of model parameters
- **Maximum likelihood**: point estimate of parameters, ignoring prior and posterior
- **Maximum likelihood ≠ Bayesian**

\[
\hat{\theta} = \arg \max_{\theta} p(x | \theta)
\]

\[
p(\theta | x) = \frac{p(x | \theta) p(\theta)}{p(x)}
\]

Scaling term Z
Example: flipping two coins

- Flipped coin is picked randomly with probability $p(z_k = 1) = \pi_k$

<table>
<thead>
<tr>
<th>$z_1$</th>
<th>$z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Which coin was flipped (unknown)
Example: flipping two coins

- Two coins with unknown probabilities $\theta_1$ and $\theta_2$ of landing heads
  
  \[ p(x|z_1, z_2, \theta_1, \theta_2) = B(x|\theta_1)^{z_1} B(x|\theta_2)^{z_2} \]
  
  \[ B(x|\theta) = \theta^x (1 - \theta)^{1-x} \]

- Flipped coin is picked randomly with probability $p(z_k = 1) = \pi_k$

<table>
<thead>
<tr>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Example: flipping two coins

Problem: since each flipped coin $z_{ik}$ is hidden, how can we learn anything from just the observations $x_i$?

Solution: estimate the responsibility $\gamma(z_{ik})$ of each coin pick $z_{ik}$ for each observation $x_i$:

$$\gamma(z_k) \equiv p(z_k = 1|x, \pi_1, \pi_2, \theta_1, \theta_2)$$
$$= \frac{p(x|z_k = 1, \theta_k)p(z_k = 1|\pi_k)}{p(x|\pi, \theta)}$$
$$= \frac{\pi_k B(x|\theta_k)}{\sum_j \pi_j B(x|\theta_j)}$$
Example: flipping two coins

- Two coins with unknown probabilities $\theta_1$ and $\theta_2$ of landing heads
  
  \[ p(x|z_1, z_2, \theta_1, \theta_2) = B(x|\theta_1)^{z_1} B(x|\theta_2)^{z_2} \]
  
  \[ B(x|\theta) = \theta^x (1 - \theta)^{1-x} \]

- Flipped coin is picked randomly with probability $p(z_k = 1) = \pi_k$

- Unknowns:
  - Which coin was flipped for each $i$?

  \[ \gamma(z_k) \equiv p(z_k = 1|x, \pi_1, \pi_2, \theta_1, \theta_2) = \frac{\pi_k B(x|\theta_k)}{\sum_j \pi_j B(x|\theta_j)} \]

  - What were the parameters?

  \[ \hat{\pi}_k = \frac{1}{N} \sum_i \gamma(z_{ik}), \quad \hat{\theta}_k = \frac{\sum_i \gamma(z_{ik}) x_i}{\sum_i \gamma(z_{ik})} \]

<table>
<thead>
<tr>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$x$</th>
<th>$\gamma(z_1)$</th>
<th>$\gamma(z_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Example: flipping two coins

- Two coins with unknown probabilities $\theta_1$ and $\theta_2$ of landing heads
  \[ p(x|z_1, z_2, \theta_1, \theta_2) = B(x|\theta_1)^{z_1} B(x|\theta_2)^{z_2} \]
  \[ B(x|\theta) = \theta^x (1 - \theta)^{1-x} \]
- Flipped coin is picked randomly with probability $p(z_i = 1) = \pi_i$
- Unknowns:
  - Which coin was flipped for each $i$?
  - What were the parameters?

<table>
<thead>
<tr>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$x$</th>
<th>$\gamma(z_1)$</th>
<th>$\gamma(z_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

E step

M step
Expectation maximization

• The EM algorithm is used to solve maximum likelihood problems in general
• Applies when the likelihood has unknown *hidden (latent) variables*
  • Latent variables are model assumptions
  • Mixture problems can be stated using discrete latent variables
  • Hidden Markov model state is a discrete latent variable
  • Kalman filter state is a continuous latent variable
• EM not tied to mixture problems however
  • Any kind of probabilistic model with unknown variables
  • EM for Hidden Markov models is also known as *Baum-Welch* or *forward-backward recursion*
  • Very widely used
Expectation maximization

• EM finds local likelihood maxima by iterative optimization
  • Alternating between $E$ step and $M$ step until convergence
  • $E$ step: compute expected values of hidden variables, given parameters
  • $M$ step: re-estimate parameters, given values for hidden variables
  • Analogous to $k$-means
  • Not guaranteed to find global likelihood maximum
Mixtures of Gaussians

- Multivariate Gaussian (normal) distribution $\mathcal{N}(x | \mu, \Sigma)$ parametrized by mean vector $\mu$ and covariance matrix $\Sigma$
- Gaussian mixture model is a linear combination of $K$ Gaussian densities

$$p(x | \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)$$

- $\pi_k$ Component weights or mixing coefficients
- $\mu_k$ Component means
- $\Sigma_k$ Component covariances
- $\theta = \{ \pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K \}$ Model parameters
Covariance structures

• Covariance matrix structure can be controlled
• Full covariance matrices
  • Unrestricted covariance (symmetric)
  • Rotated ellipsoid shape
  • \((D + 1)D/2\) parameters
• Diagonal covariance \(\Sigma_{ij} = 0, \ i \neq j\)
  • Zeros except on diagonal
  • Ellipsoid shape
  • \(D\) parameters
• Scaled identity covariance \(\Sigma = \lambda I\)
  • Zeros except on diagonal
  • Diagonal values are equal
  • Spherical shape
  • One parameter
Gaussian mixture likelihood

• Log-likelihood of Gaussian mixture

\[
\log p(x|\theta) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}
\]

• Maximum likelihood

\[
\hat{\theta} = \arg\max_{\theta} \log p(x|\theta)
\]

• Has no closed-form solution
Latent variable formulation of Gaussian mixtures

- To apply EM, we introduce latent $K$-dimensional indicator variable $z$ to Gaussian mixtures
  - $z_k = 1$ iff $k^{th}$ component is active
  - $z_k = 0$ otherwise

- Then $p(z|\theta) = \prod_{k=1}^{K} \pi_k^{z_k}$ and $p(x|z, \theta) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}$

- Marginalizing over $z$,
  $p(x|\theta) = \sum_z p(z|\theta)p(x|z, \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$

- Therefore, we obtain the same Gaussian mixture density as earlier
EM for Gaussian mixtures

- Complete-data log likelihood

\[
\log p(x, z|\theta) = \log \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k z_{nk} \mathcal{N}(x_n|\mu_k, \Sigma_k) z_{nk}
\]

\[
= \sum_n \sum_k z_{nk} [\log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k)]
\]

- Responsibilities

\[
\gamma(z_k) \equiv p(z_k = 1|x, \theta) = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}
\]
**EM algorithm for Gaussian mixtures**

**E step**
- Evaluate responsibilities, given current parameters

\[
\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_j^K \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \Sigma_j)}
\]

**M step**
- Re-estimate parameters, given current responsibilities

\[
\begin{align*}
\pi_k &= \frac{1}{N} \sum_{n=1}^N \gamma(z_{nk}) \\
\mu_k &= \frac{\sum_n \gamma(z_{nk}) \mathbf{x}_n}{\sum_n \gamma(z_{nk})} \\
\Sigma_k &= \frac{\sum_n \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_n \gamma(z_{nk})}
\end{align*}
\]
EM algorithm for Gaussian mixtures

E step

• Evaluate responsibilities, given current parameters

\[ \gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)} \]

M step

• Re-estimate parameters, given current responsibilities

\[ \pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) \]

\[ \mu_k = \frac{\sum_n \gamma(z_{nk}) x_n}{\sum_n \gamma(z_{nk})} \]

\[ \Sigma_k = \frac{\sum_n \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_n \gamma(z_{nk})} \]
EM algorithm for Gaussian mixtures

E step

- Evaluate responsibilities, given current parameters

\[ \gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)} \]

M step

- Re-estimate parameters, given current responsibilities

\[ \pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) \]

\[ \mu_k = \frac{\sum_{n} \gamma(z_{nk}) x_n}{\sum_{n} \gamma(z_{nk})} \]

\[ \Sigma_k = \frac{\sum_{n} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n} \gamma(z_{nk})} \]
EM algorithm for Gaussian mixtures

E step
- Evaluate responsibilities, given current parameters

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}$$

M step
- Re-estimate parameters, given current responsibilities

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\mu_k = \frac{\sum_{n} \gamma(z_{nk}) x_n}{\sum_{n} \gamma(z_{nk})}$$

$$\Sigma_k = \frac{\sum_{n} \gamma(z_{nk}) (x_n - \mu_k) (x_n - \mu_k)^T}{\sum_{n} \gamma(z_{nk})}$$

Mean responsibility

Weighted mean of observations

Weighted mean of variances

Mean (new) responsibility

Weighted mean of observations

Weighted mean of variances
Demo: mixture of three Gaussians

- Two-dimensional data
- Find max likelihood estimate for means of a mixture of $k=3$ Gaussians
  - $\Sigma = I$
- E step
  - Show responsibilities in color
  - Red, green, blue color corresponding to each of three components
- M step
  - Show Gaussian mixture density in color
E step #1, log p(X|θ)=−Inf
M step #1, log p(X|θ) = -3034.917653
E step #2, log p(X|\theta) = -3034.917653
M step #2, log p(X|\theta)=-1730.289138
E step #3, log p(X|θ)=-1730.289138
M step #3, $\log p(X|\theta) = -1217.593757$
E step #4, log p(X|θ)=-1217.593757
M step #4, log p(X|θ)=-1173.170092
E step #5, \( \log p(X|\theta) = -1173.170092 \)
M step #5, log p(X|θ)=-1171.878985
E step #6, $\log p(X|\theta) = -1171.878985$
General EM problem

- How to find maximum likelihood estimates of latent variables in a probabilistic model?

- Assumptions
  - Latent variables are random variables
  - Complete-data likelihood is known
  - Complete-data likelihood expectation can be maximized

- Requirements
  - Observed data
  - Complete-data likelihood
  - Initial parameters

- EM is not a single algorithm, but a recipe of an algorithm
General EM solution

• Given
  • Observed data $x_1, x_2, \ldots, x_N$
  • Complete-data likelihood $p(x, z|\theta)$
  • Initial parameters $\theta^{\text{old}}$

• E step
  • Compute expected value of complete-data log likelihood $\log p(x, z|\theta)$ over hidden variables $z$, given parameters $\theta^{\text{old}}$
  \[
  Q(\theta, \theta^{\text{old}}) = \mathbb{E}_z[\log p(x, z|\theta) | x, \theta^{\text{old}}] = \sum_z p(z|x, \theta^{\text{old}}) \log p(x, z|\theta)
  \]
  • Assuming $z$ is a random variable, $z \sim p(z|x, \theta^{\text{old}})$

• M step
  • Re-estimate parameters $\theta^{\text{new}}$, given values for hidden variables $Q(\theta, \theta^{\text{old}})$
  \[
  \theta^{\text{new}} = \arg\max_{\theta} Q(\theta, \theta^{\text{old}})
  \]
  • Repeat E and M steps until convergence (and restart with random parameters)
General EM algorithm properties

- Convergence to local maximum guaranteed
  - Each iteration can only increase the observed data likelihood
  - Not necessarily global maximum likelihood
- Benefits
  - Applicable to a wide variety of problems (latent probabilistic models)
  - Faster than Monte Carlo simulations
- Drawbacks
  - Maximum likelihood point estimates are often misleading, e.g., Gaussian singularities
  - Not strictly maximum likelihood (local maxima)
  - Slow convergence for big data
- Bayesian alternatives
  - Monte Carlo simulation
  - Variational approximation
Summary

- Expectation maximization is an algorithm recipe for solving maximum likelihood problems having latent variables
  - Algorithm recipe: family of related algorithms, not a single algorithm
  - Maximum likelihood: point estimate, not fully Bayesian
  - Latent variables: estimate arbitrary hidden parameters, not only Gaussian mixtures
- EM algorithm is iterative and converges to local likelihood maxima
- Full Bayesian alternatives are often computationally heavier
  - Monte Carlo, etc.
- EM for Gaussian mixtures
  - Estimate mixture weights, means, and covariances, given observations
  - Soft version of $k$-means
  - How to choose $k$?
References
