CS 2750: Machine Learning
Expectation Maximization

Prof. Adriana Kovashka
University of Pittsburgh
April 11, 2017
Plan for this lecture

- EM for Hidden Markov Models (last lecture)
- EM for Gaussian Mixture Models
- EM in general
- K-means ~ GMM
- EM for Naïve Bayes
The Gaussian Distribution

\[ \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \]

\[ \mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\} \]
Mixtures of Gaussians

Old Faithful data set

Single Gaussian

Mixture of two Gaussians
Mixtures of Gaussians

Combine simple models into a complex model:

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

Component
Mixing coefficient

\( \forall k : \pi_k \geq 0 \quad \sum_{k=1}^{K} \pi_k = 1 \)
Mixtures of Gaussians

• Introducing latent variables $z$, one for each $x$
  • 1-of-K representation:

$$p(z_k = 1) = \pi_k$$

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

• Also:

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$

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

• Then $p(x) = \Sigma_z p(x, z) = \Sigma_z p(z) p(x|z) =$

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

Figures from Chris Bishop
Mixtures of Gaussians

- Responsibility of component $k$ for explaining $\mathbf{x}$:

$$
\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mathbf{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mathbf{\mu}_j, \Sigma_j)}.
$$
Generating samples

- Sample a value $z^*$ from $p(z)$ then sample a value for $x$ from $p(x|z^*)$
- Color generated samples using $z^*$ (left)
- Color samples using responsibilities (right)
Finding parameters of mixture

- Want to maximize:

\[
\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right)
\]

- Set derivative with respect to means to 0:

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

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

Figures from Chris Bishop
Finding parameters of mixture

• Set derivative wrt covariances to 0:

\[ \Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \]

• Set derivative wrt mixing coefficients to 0:

\[ \pi_k = \frac{N_k}{N} \]
Reminder

• Responsibilities:

\[ \gamma(z_k) \equiv p(z_k = 1 | x) = \frac{\pi_k \mathcal{N}(x | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x | \mu_j, \Sigma_j)} \]

• So parameters of GMM depend on responsibilities and vice versa…

• What can we do?

Figures from Chris Bishop
Reminder: K-means clustering

Basic idea: randomly initialize the $k$ cluster centers, and iterate:

1. *Randomly* initialize the cluster centers, $c_1, \ldots, c_K$
2. *Given cluster centers*, determine points in each cluster
   - For each point $p$, find the closest $c_i$. Put $p$ into cluster $i$
3. *Given points in each cluster*, solve for $c_i$
   - Set $c_i$ to be the mean of points in cluster $i$
4. If $c_i$ have changed, repeat Step 2
Iterative algorithm

EM for Gaussian Mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\mu_k$, covariances $\Sigma_k$ and mixing coefficients $\pi_k$, and evaluate the initial value of the log likelihood.

2. E step. Evaluate the responsibilities using the current parameter values

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

(9.23)
3. **M step.** Re-estimate the parameters using the current responsibilities

\[ \mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \]  
\[ \Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k^{\text{new}})(x_n - \mu_k^{\text{new}})^T \]  
\[ \pi_k^{\text{new}} = \frac{N_k}{N} \] 

where

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

4. Evaluate the log likelihood

\[ \ln p(X | \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right\} \]

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.
Figures from Chris Bishop
EM in general

- For models with hidden variables, the probability of the data $\mathbf{X} = \{x_n\}$ depends on the hidden variables $\mathbf{Z} = \{z_n\}$
- Complete data set $\{\mathbf{X}, \mathbf{Z}\}$
- Incomplete data set (observed) $\mathbf{X}$, have $p(\mathbf{Z}|\mathbf{X}, \theta)$
- We want to maximize:

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$
EM in general

- We don’t know the values for the hidden variables $Z$
- We also don’t know the model parameters $\theta$
- Set one, infer the other
  - Keep model parameters $\theta$ fixed, infer the hidden variables $Z$ (compute conditional probability)
  - Keep $Z$ fixed, infer the model parameters $\theta$ (maximize expectation)
EM in general

The General EM Algorithm

Given a joint distribution $p(X, Z|\theta)$ over observed variables $X$ and latent variables $Z$, governed by parameters $\theta$, the goal is to maximize the likelihood function $p(X|\theta)$ with respect to $\theta$.

1. Choose an initial setting for the parameters $\theta^{\text{old}}$.
2. E step Evaluate $p(Z|X, \theta^{\text{old}})$.
3. M step Evaluate $\theta^{\text{new}}$ given by

$$\theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}})$$

(9.32)

where

$$Q(\theta, \theta^{\text{old}}) = \sum_{Z} p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta).$$

(9.33)

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$$

(9.34)

and return to step 2.
GMM in the context of general EM

• Complete data log likelihood:

$$\ln p(X, Z|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{ \ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}$$

• Expectation of this likelihood:

$$\mathbb{E}_Z[\ln p(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}[z_{nk}] \gamma(z_{nk}) \left\{ \ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}$$

• Choose initial values $\mu^{\text{old}}, \Sigma^{\text{old}}, \pi^{\text{old}}$, evaluate responsibilities $\gamma(z_{nk})$

• Fix responsibilities, maximize expectation wrt $\mu, \Sigma, \pi$

Figures from Chris Bishop
K-means \sim GMM

- Consider GMM where all components have covariance matrix $\epsilon I$ ($I =$ the identity matrix):

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp \left\{ -\frac{1}{2\epsilon} \left\| x - \mu_k \right\|^2 \right\}$$

- Responsibilities are then:

$$\gamma(z_{nk}) = \frac{\pi_k \exp \left\{ -\left\| x_n - \mu_k \right\|^2 / 2\epsilon \right\}}{\sum_j \pi_j \exp \left\{ -\left\| x_n - \mu_j \right\|^2 / 2\epsilon \right\}}$$

- As $\epsilon \to 0$, most responsibilities $\to 0$, except $\gamma(z_{nj}) \to 1$ for whichever $j$ has the smallest $\left\| x_n - \mu_j \right\|^2$
K-means $\sim$ GMM

- Expectation of the complete-data log likelihood, which we want to maximize:

$$
E_Z[\ln p(X, Z|\mu, \Sigma, \pi)] \rightarrow -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 + \text{const.}
$$

- Recall in K-means we wanted to minimize:

$$
J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2
$$

where $r_{nk} = 1$ if instance $n$ belongs to cluster $k$, 0 otherwise
Example: EM for Naïve Bayes

- We have a graphical model:

  ![Graphical Model]

  - Flavor = class (hidden)
  - Shape + color = features (observed)

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(round, brown)</td>
<td>0.282</td>
</tr>
<tr>
<td>P(round, red)</td>
<td>0.139</td>
</tr>
<tr>
<td>P(square, brown)</td>
<td>0.141</td>
</tr>
<tr>
<td>P(square, red)</td>
<td>0.438</td>
</tr>
<tr>
<td>Total candies</td>
<td>1000</td>
</tr>
</tbody>
</table>

Log likelihood $P(X|\Theta)$

$$= 282 \times \log(0.282) + \ldots = -1827.17$$

Example from Rebecca Hwa
Example: EM for Naïve Bayes

- Model parameters = conditional probability table, initialize it randomly

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P(straw)</td>
<td>0.59</td>
</tr>
<tr>
<td>P(choc)</td>
<td>0.41</td>
</tr>
<tr>
<td>P(round</td>
<td>straw)</td>
</tr>
<tr>
<td>P(square</td>
<td>straw)</td>
</tr>
<tr>
<td>P(brown</td>
<td>straw)</td>
</tr>
<tr>
<td>P(red</td>
<td>straw)</td>
</tr>
<tr>
<td>P(round</td>
<td>choc)</td>
</tr>
<tr>
<td>P(square</td>
<td>choc)</td>
</tr>
<tr>
<td>P(brown</td>
<td>choc)</td>
</tr>
<tr>
<td>P(red</td>
<td>choc)</td>
</tr>
</tbody>
</table>
Example: EM for Naïve Bayes

• Expectation step
  • Keeping model (CPT) fixed, estimate values for joint probability distribution by multiplying $P(\text{flavor}) \times P(\text{shape} | \text{flavor}) \times P(\text{color} | \text{flavor})$
  • Then use these to compute conditional probability of the hidden variable flavor: $P(\text{straw} | \text{shape}, \text{color})$
    $= \frac{P(\text{straw}, \text{shape}, \text{color})}{P(\text{shape}, \text{color})}$
    $= \frac{P(\text{straw}, \text{shape}, \text{color})}{\sum_{\text{flavor}} P(\text{flavor}, \text{shape}, \text{color})}$

Example from Rebecca Hwa
Example: EM for Naïve Bayes

- Expectation step

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>p(straw, round, brown)</td>
<td>0.05</td>
<td>p(straw</td>
</tr>
<tr>
<td>p(straw, square, brown)</td>
<td>0.00</td>
<td>p(straw</td>
</tr>
<tr>
<td>p(straw, round, red)</td>
<td>0.52</td>
<td>p(straw</td>
</tr>
<tr>
<td>p(straw, square, red)</td>
<td>0.02</td>
<td>p(straw</td>
</tr>
<tr>
<td>p(choc, round, brown)</td>
<td>0.19</td>
<td>p(choc</td>
</tr>
<tr>
<td>p(choc, square, brown)</td>
<td>0.10</td>
<td>p(choc</td>
</tr>
<tr>
<td>p(choc, round, red)</td>
<td>0.08</td>
<td>p(choc</td>
</tr>
<tr>
<td>p(choc, square, red)</td>
<td>0.04</td>
<td>p(choc</td>
</tr>
<tr>
<td>P(round, brown)</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>P(round, red)</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>P(square, brown)</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>P(square, red)</td>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>

Log likelihood $P(X|\Theta)$

$= 282*\log(0.24) + ... = -2912.63$

Example from Rebecca Hwa
Example: EM for Naïve Bayes

• Maximization step
  • Expected # “strawberry, round, brown” candies
    = $P(\text{strawberry} \mid \text{round, brown})$ \(\text{// from E step}\)
    * # (round, brown) = 0.19 * 282 = 53.6
  • Can find other expected counts needed for CPT

Example from Rebecca Hwa
Example: EM for Naïve Bayes

- Maximization step

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>exp # straw, round</td>
<td>174.56</td>
<td>0.33</td>
<td></td>
<td>0.52</td>
<td></td>
<td>0.48</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp. # straw, square</td>
<td>159.16</td>
<td>0.33</td>
<td></td>
<td>0.52</td>
<td></td>
<td>0.48</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # straw, red</td>
<td>277.91</td>
<td>0.67</td>
<td></td>
<td>0.37</td>
<td></td>
<td>0.63</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # straw, brown</td>
<td>55.81</td>
<td>0.67</td>
<td></td>
<td>0.37</td>
<td></td>
<td>0.63</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # straw</td>
<td>333.72</td>
<td>0.67</td>
<td></td>
<td>0.37</td>
<td></td>
<td>0.63</td>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # choc, round</td>
<td>246.44</td>
<td>0.33</td>
<td></td>
<td>0.52</td>
<td></td>
<td>0.48</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # choc, square</td>
<td>419.84</td>
<td>0.33</td>
<td></td>
<td>0.52</td>
<td></td>
<td>0.48</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # choc, red</td>
<td>299.09</td>
<td>0.33</td>
<td></td>
<td>0.52</td>
<td></td>
<td>0.48</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # choc, brown</td>
<td>367.19</td>
<td>0.33</td>
<td></td>
<td>0.52</td>
<td></td>
<td>0.48</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp # choc</td>
<td>666.28</td>
<td>0.33</td>
<td></td>
<td>0.52</td>
<td></td>
<td>0.48</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: EM for Naïve Bayes

• Maximization step

• Compute with new parameters (CPT)
  • $P(\text{round, brown})$
  • $P(\text{round, red})$
  • $P(\text{square, brown})$
  • $P(\text{square, red})$

• Compute likelihood $P(\mathbf{X}|\Theta)$
• Repeat until convergence

Example from Rebecca Hwa