CS 4824/ECE 4424: Clustering

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
EM algorithm — recap

- EM is a general procedure for learning from partly observed data
- Given observed variables $X$, unobserved $Z$ (\(X=\{F,A,H,N\}, \ Z=\{S\}\))

Begin with arbitrary choice for parameters $\theta$

Iterate until convergence:

- **E Step**: estimate the values of unobserved $Z$ conditioned on $X$ using $\theta$
- **M Step**: use observed values plus E-step estimates to derive a better $\theta$

- Guaranteed to find local maximum. Each iteration increases

\[
E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]
\]
What if we have no labeled data at all?

un

semi-supervised learning
Unsupervised clustering

Just extreme case of EM with zero labeled examples…
From partially unlabeled data to no labeled data at all...

<table>
<thead>
<tr>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>?</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>?</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

semi-supervised learning

unsupervised learning

Learn $P(Y|X)$
Clustering

- Given set of data points, without class labels, group them
- Unsupervised learning
- Which news items are similar? (or which customers, faces, web pages, ...)
- Many practical applications...
Clustering
Mixture Distributions

- Model joint distribution $P(X_1 \ldots X_n)$ as mixture of multiple distributions.

- Use discrete-valued random variable $Z$ to indicate which distribution is being used for each random draw.

$$P(X_1 \ldots X_n) = \sum_i P(Z = i) \ P(X_1 \ldots X_n | Z)$$

- Mixture of Gaussians:
  - Assume each data point $X=\langle X_1, \ldots, X_n \rangle$ is generated by one of several Gaussians, as follows:
    - randomly choose Gaussian $i$, according to $P(Z=i)$
    - randomly generate a data point $\langle x_1, x_2, \ldots, x_n \rangle$ according to the parameters of the Gaussian distributions corresponding to $i$
Mixture of Gaussians
EM for Mixture of Gaussian Clustering

- Let’s simplify to make this easier:
  - Assume $X = \langle X_1 \ldots X_n \rangle$, and the $X_i$ are conditionally independent given $Z$. $P(X|Z = j) = \prod_{i} N(X_i|\mu_{ji}, \sigma_{ji})$
  - Assume only 2 clusters (values of $Z$), and $\forall i, j, \sigma_{ji} = \sigma$
  - $P(X) = \sum_{j=1}^{2} P(Z = j|\pi) \prod_{i} N(x_i|\mu_{ji}, \sigma)$
  - Assume $\sigma$ known, $\pi_1 \ldots \pi_K, \mu_{1i} \ldots \mu_{Ki}$

- Observed: $X = \langle X_1 \ldots X_n \rangle$
- Unobserved: $Z$
EM

- Given observed variables $X$, unobserved $Z$,
  - define $Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')]$ where $\theta = (\pi, \mu_{ji})$

- Iterate until convergence:
  - E Step:
    - Calculate $P(Z(n)|X(n),\theta)$ for each example $X(n)$.
    - Use this to construct $Q(\theta'|\theta)$
  - M Step:
    - Replace current $\theta$ by
    $$\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$$
EM — E Step

- Calculate $P(Z(n) \mid X(n), \theta)$ for each observed example $X(n) = \langle x_1(n), x_2(n), \ldots, x_T(n) \rangle$

\[
P(z(n) = k \mid x(n), \theta) = \frac{P(x(n) \mid z(n) = k, \theta) \cdot P(z(n) = k \mid \theta)}{\sum_{j=0}^{1} P(x(n) \mid z(n) = j, \theta) \cdot P(z(n) = j \mid \theta)}
\]

\[
P(z(n) = k \mid x(n), \theta) = \frac{\left[ \prod_i P(x_i(n) \mid z(n) = k, \theta) \right] \cdot P(z(n) = k \mid \theta)}{\sum_{j=0}^{1} \prod_i P(x_i(n) \mid z(n) = j, \theta) \cdot P(z(n) = j \mid \theta)}
\]

\[
P(z(n) = k \mid x(n), \theta) = \frac{\left[ \prod_i N(x_i(n) \mid \mu_{k,i}, \sigma) \right] \cdot \left(\pi^k(1 - \pi)^{(1-k)}\right)}{\sum_{j=0}^{1} \left[ \prod_i N(x_i(n) \mid \mu_{j,i}, \sigma) \right] \cdot \left(\pi^j(1 - \pi)^{(1-j)}\right)}
\]
EM — M Step

- First consider update for $\pi$

\[ Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')] = E[\log P(X|Z, \theta') + \log P(Z|\theta')] \]

\[ \pi \leftarrow \arg \max_{\pi'} E_{Z|X,\theta}[\log P(Z|\pi')] \]

\[ E_{Z|X,\theta}[\log P(Z|\pi')] = \]

\[ \frac{\partial E_{Z|X,\theta}[\log P(Z|\pi')]}{\partial \pi'} = \]
EM — M Step

- First consider update for $\pi$

\[
Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')] = E[\log P(X|Z, \theta') + \log P(Z|\theta')]
\]

\[
\pi \leftarrow \arg\max_{\pi'} E_{Z|X,\theta}[\log P(Z|\pi')]
\]

\[
E_{Z|X,\theta}[\log P(Z|\pi')] = E_{Z|X,\theta}\left[\log \left(\pi' \sum_n z(n) (1 - \pi') \sum_n (1 - z(n))\right)\right]
\]

\[
= E_{Z|X,\theta}\left[\left(\sum_n z(n)\right) \log \pi' + \left(\sum_n (1 - z(n))\right) \log (1 - \pi')\right]
\]

\[
= \left(\sum_n E_{Z|X,\theta}[z(n)]\right) \log \pi' + \left(\sum_n E_{Z|X,\theta}[(1 - z(n))]\right) \log (1 - \pi')
\]

\[
\frac{\partial E_{Z|X,\theta}[\log P(Z|\pi')]}{\partial \pi'} = \left(\sum_n E_{Z|X,\theta}[z(n)]\right) \frac{1}{\pi'} + \left(\sum_n E_{Z|X,\theta}[(1 - z(n))]\right) \frac{-1}{1 - \pi'}
\]

\[
\pi \leftarrow \frac{\sum_{n=1}^{N} E[z(n)]}{\left(\sum_{n=1}^{N} E[z(n)]\right) + \left(\sum_{n=1}^{N} (1 - E[z(n)])\right)} = \frac{1}{N} \sum_{n=1}^{N} E[z(n)]
\]
Now consider update for $\mu_{ji}$

$$Q(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')] = E[\log P(X|Z, \theta') + \log P(Z|\theta')]$$

$$\mu_{ji} \leftarrow \arg \max_{\mu'_{ji}} E_{Z|X,\theta}[\log P(X|Z, \theta')]$$

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^{N} P(z(n) = j|x(n), \theta)}{\sum_{n=1}^{N} P(z(n) = j|x(n), \theta)} \cdot x_{i}(n)$$

Compare above to MLE if $Z$ were observable:

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^{N} \delta(z(n) = j)}{\sum_{n=1}^{N} \delta(z(n) = j)} \cdot x_{i}(n)$$
EM — putting it together

- Given observed variables $X$, unobserved $Z$,
  - define $Q(\theta' | \theta) = E_Z|X,\theta[\log P(X, Z | \theta')]$ where $\theta = \langle \pi, \mu_{ji} \rangle$

- Iterate until convergence:
  - E Step:
    - For each observed example $X(n)$, calculate $P(Z(n)|X(n), \theta)$
      \[
P(z(n) = k \mid x(n), \theta) = \frac{[\prod_i N(x_i(n)|\mu_{k,i}, \sigma)] (\pi^k(1-\pi)^{(1-k)})}{\sum_j^1[\prod_i N(x_i(n)|\mu_{j,i}, \sigma)] (\pi^j(1-\pi)^{(1-j)})}
      \]
  - M Step:
    - Update current $\theta$ by $\theta' \leftarrow \arg \max_{\theta'} Q(\theta' | \theta)$
      \[
      \pi \leftarrow \frac{1}{N} \sum_{n=1}^N E[z(n)]
      \]
      \[
      \mu_{ji} \leftarrow \frac{\sum_{n=1}^N P(z(n) = j|x(n), \theta) \cdot x_i(n)}{\sum_{n=1}^N P(z(n) = j|x(n), \theta)}
      \]
Demo Time 😊

https://lukapopijac.github.io/gaussian-mixture-model/
What you should know

- For learning from partly observed data
- Instead of MLE:  \( \theta \leftarrow \arg \max_{\theta} \log P(X, Z|\theta) \)
- EM estimates:  \( \theta \leftarrow \arg \max_{\theta} E_{Z|X,\theta} [\log P(X, Z|\theta)] \)
  - where \( X \) is observed part of the data, and \( Z \) is (partly) unobserved
- EM for training Bayes Nets
- Can also develop MAP version instead of EM
  - Write out expression for  \( E_{Z|X,\theta} [\log P(X, Z|\theta)] \)
  - E step: for each training example \( X^k \), calculate \( P(Z^k|X^k,\theta) \)
  - M step: choose new to maximize  \( E_{Z|X,\theta} [\log P(X, Z|\theta)] \)
Bayes Net—summary

- Representation
  - Bayes Net represent joint distributions as a DAG + conditional distributions
  - Let’s us calibrate conditional independence assumptions
- Inference
  - NP-hard in general
  - For some graph, closed form inference possible
  - Approximate methods exists too, e.g., Monte Carlo methods,…
- Learning
  - Easy for known graph, fully observed data (MLE, MAP etc.)
  - EM for partly observed data
  - Can handle the extreme case of completely unlabeled data