Today:

• Learning graphical models
  1. EM: learning from partially observed data
  2. Mixture models, clustering
  3. Structure learning

Readings:

• Bishop chapter 9-9.2 mixture models
• Kevin Murphy chapter 11.4 (optional)

EM : Learning from Partially Observed Training Data
EM algorithm

- Let X be all *observed* variable values (over all examples)
- Let Z be all *unobserved* variable values

EM algorithm:

- Iterate until convergence:
  - **E step:** use current Bayes net parameters $\theta$ to estimate unobserved Z values

\[
\theta_{K=1|S=i,A=j} = P(K = 1|S = i, A = j) = \frac{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} P_\theta(s_m = i, a_m = j)}
\]

- **M step:** use estimated values of Z to retrain Bayes net params $\theta$
EM algorithm

- Let \( X \) be all observed variable values (over all examples)
- Let \( Z \) be all unobserved variable values

**EM algorithm:**

- Iterate until convergence:
  - **E step:** use current Bayes net parameters \( \theta \) to estimate unobserved \( Z \) values
  - **M step:** use estimated values of \( Z \) to retrain Bayes net params \( \theta \)

\[
\theta_{K=1|S=i,A=j} = P(K = 1|S = i, A = j) = \frac{\sum_{m=1}^{M} P_{\theta}(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} P_{\theta}(s_m = i, a_m = j)}
\]

wait – how do we compute these probabilities??

m\(^{th}\) training example
Only One Unobserved Variable:

How do we calculate \( P(K=1 \mid S=s, A=a, E=e, H=h) \)?

\[
P(K = 1 \mid S = s, A = a, E = e, H = h) = \frac{P(S = s, A = a, K = 1, E = e, H = h)}{P(S = s, A = a, E = e, H = h)} \\
= \frac{P(S = s, A = a, K = 1, E = e, H = h)}{P(S = s, A = a, K = 1, E = e, H = h) + P(S = s, A = a, K = 0, E = e, H = h)}
\]

where:

\[
P(S = s, A = a, K = k, E = e, H = h) = P(S = s)P(A = a)P(K = k \mid S = s, A = a)P(E = e \mid K = k)P(H = 1 \mid K = k)
\]

Efficient: \( O(2^n) \) for \( n \) Boolean variables.
EM algorithm

- Let X be all *observed* variable values (over all examples)
- Let Z be all *unobserved* variable values

**EM algorithm:**

- Iterate until convergence:
  - **E step:** use current Bayes net parameters $\theta$ to estimate unobserved Z values
  - **M step:** use estimated values of Z to retrain Bayes net params $\theta$

$$
\theta_{K=1|S=i,A=j} = P(K = 1|S = i, A = j) = \frac{\sum_{m=1}^{M} P_{\theta}(s_m = i, a_m = j, k_m = 1)}{\sum_{m=1}^{M} P_{\theta}(s_m = i, a_m = j)}
$$

$m^{th}$ training example
EM Algorithm - Precisely

EM is a general procedure for learning from partly observed data.

Given observed training feature values $X$, unobserved $Z$, from all examples.

Iterate until convergence:

- **E Step**: Use $X$ and current $\theta$ to calculate $P(Z|X, \theta)$
- **M Step**: Replace current $\theta$ by

$$
\theta \leftarrow \operatorname{arg \ max}_{\theta'} E_{P(Z|X, \theta)} \left[ \log P(X, Z|\theta') \right]
$$

Guaranteed to find $\theta$ that is local maximum of $E_{P(Z|X, \theta)} \left[ \log P(X, Z|\theta') \right]$
Using Unlabeled Data to Help Train Naïve Bayes Classifier

Learn $P(Y|X)$

<table>
<thead>
<tr>
<th></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>
EM for semi-supervised Naïve Bayes

Given observed set X, unobserved set Y of values (only missing values are labels Y for some examples)

E step: Calculate for each training example, k
the expected value of each unobserved value of variable Y

\[ E_{P(Y|X_1,...,X_N)}[y(k)] = P(y(k) = 1|x_1(k),...,x_N(k); \theta) = \frac{P(y(k) = 1) \prod_i P(x_i(k)|y(k) = 1)}{\sum_{j=0}^{1} P(y(k) = j) \prod_i P(x_i(k)|y(k) = j)} \]

M step: Calculate estimates similar to MLE, but replacing each count by its expected count
EM for semi-supervised Naïve Bayes

Given observed set X, unobserved set Y of values (only missing values are labels Y for some examples)

E step: Calculate for each training example, k
the expected value of each unobserved value of variable Y

\[
E_{P(Y|X_1...X_N)}[y(k)] = P(y(k) = 1|x_1(k), \ldots x_N(k); \theta) = \frac{P(y(k) = 1) \prod_i P(x_i(k)|y(k) = 1)}{\sum_{j=0}^1 P(y(k) = j) \prod_i P(x_i(k)|y(k) = j)}
\]

Why is expected value of Boolean-valued Y just P(Y=1)?

Answer: the definition of expected value:

\[
E_{P(Y)}[Y] = \sum_{y\in\{0,1\}} P(Y = y) \cdot y
\]

\[
= [P(Y = 1) \cdot 1] + [P(Y = 0) \cdot 0]
\]

\[
= P(Y = 1)
\]
EM for semi-supervised Naïve Bayes

Given observed set X, unobserved set Y of values (only missing values are labels Y for some examples)

E step: Calculate for each training example, k

the expected value of each unobserved value of variable Y

\[ E_{P(Y|X_1...X_N)}[y(k)] = P(y(k) = 1|x_1(k),...x_N(k); \theta) = \frac{P(y(k) = 1) \prod_i P(x_i(k)|y(k) = 1)}{\sum_{j=0}^1 P(y(k) = j) \prod_i P(x_i(k)|y(k) = j)} \]

M step: Calculate estimates similar to MLE, but replacing each count by its expected count
Given observed set $X$, unobserved set $Y$ of values (only missing values are labels $Y$ for some examples)

**E step:** Calculate for each training example, $k$

the expected value of each unobserved variable $Y$

$$E_{P(Y|X_1...X_N)}[y(k)] = P(y(k) = 1|x_1(k),...x_N(k); \theta) = \frac{P(y(k) = 1) \prod_i P(x_i(k)|y(k) = 1)}{\sum_{j=0}^1 P(y(k) = j) \prod_i P(x_i(k)|y(k) = j)}$$

**M step:** Calculate estimates similar to MLE, but replacing each count by its expected count

$$\theta_{ij|m} = \hat{P}(X_i = j|Y = m) = \frac{\sum_k P(y(k) = m|x_1(k)\ldots x_N(k)) \delta(x_i(k) = j)}{\sum_k P(y(k) = m|x_1(k)\ldots x_N(k))}$$

MLE would be:

$$\hat{P}(X_i = j|Y = m) = \frac{\sum_k \delta((y(k) = m) \wedge (x_i(k) = j))}{\sum_k \delta(y(k) = m)}$$
20 Newsgroups
Question: What if our data provides no Y labels, but we believe $P(Y,X_1,X_2,X_3,X_4)$ is defined by this Naïve Bayes net structure?

Can we still use EM to learn $P(Y,X_1,X_2,X_3,X_4)$?
Question: What if our data provides no Y labels, but we believe \( P(Y,X1,X2,X3,X4) \) is defined by this Naïve Bayes net structure?

→ Unsupervised clustering
→ Y is the unobserved indicator of which cluster each X belongs to. 
  \( P(Y=1|X) \), \( P(Y=0|X) \) indicate the prob. that X belongs to each cluster
→ Or, if we want to consider more clusters, we define Y to have more values (i.e., Y in \{0,1,2,…,N\} )

Unobserved cluster label to be learned
Question: What if our data provides no Y labels, but we believe \( P(Y, X_1, X_2, X_3, X_4) \) is defined by this Naïve Bayes net structure?

→ Unsupervised clustering
→ \( Y \) is the unobserved indicator of which cluster each \( X \) belongs to. \( P(Y=1|X), P(Y=0|X) \) indicate the prob. that \( X \) belongs to each cluster

Suppose we assume \( P(X_1, X_2, X_3, X_4) \) is a mixture of two distributions (two clusters). Then:

\[
P(X_1, X_2, X_3, X_4) = P(Y=1) P(X_1, X_2, X_3, X_4 | Y=1) + P(Y=0) P(X_1, X_2, X_3, X_4 | Y=0)
\]
Question: What if our data provides no Y labels, but we believe \( P(Y,X_1,X_2,X_3,X_4) \) is defined by this Naïve Bayes net structure?

\[ \rightarrow \text{Unsupervised clustering} \]
\[ \rightarrow Y \text{ is the unobserved indicator of which cluster each } X \text{ belongs to.} \]
\[ P(Y=1|X), P(Y=0|X) \text{ indicate the prob. that } X \text{ belongs to each cluster} \]

Suppose we assume \( P(X_1,X_2,X_3,X_4) \) is a mixture of two distributions (two clusters). Then:
\[
P(X_1,X_2,X_3,X_4) = P(Y=1)P(X_1,X_2,X_3,X_4 | Y=1) + P(Y=0)P(X_1,X_2,X_3,X_4 | Y=0)
\]

This form is called a “mixture distribution”
Question: What if our data provides no Y labels, but we believe $P(Y,X_1,X_2,X_3,X_4)$ is defined by this Naïve Bayes net structure?

→ Unsupervised clustering : EM

Learned probabilistic cluster label

<table>
<thead>
<tr>
<th>Pr</th>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Question: What if our data provides no Y labels, but we believe $P(Y,X_1,X_2,X_3,X_4)$ is defined by this Naïve Bayes net structure?

→ Unsupervised clustering : EM

What if real-valued $X_i$’s?
Question: What if our data provides no Y labels, but we believe \( P(Y, X_1, X_2, X_3, X_4) \) is defined by this Naïve Bayes net structure?

\[ Y \rightarrow X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \]

→ Unsupervised clustering : EM

What if real-valued \( X_i \)'s?
Need different form of \( P(X_i|Y) \)
e.g., Gaussian

\[
P(X_i = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma_{iy}^2}} \exp\left(-\frac{1}{2\sigma_{iy}^2} (x - \mu_{iy})^2\right)
\]

<table>
<thead>
<tr>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>0.1</td>
<td>7.2</td>
<td>3.1</td>
<td>1.4</td>
</tr>
<tr>
<td>?</td>
<td>9.9</td>
<td>2.1</td>
<td>5.0</td>
<td>0.2</td>
</tr>
<tr>
<td>?</td>
<td>8.0</td>
<td>0.7</td>
<td>5.1</td>
<td>0.9</td>
</tr>
<tr>
<td>?</td>
<td>1.1</td>
<td>6.2</td>
<td>2.9</td>
<td>2.1</td>
</tr>
<tr>
<td>?</td>
<td>1.4</td>
<td>8.3</td>
<td>2.7</td>
<td>1.8</td>
</tr>
</tbody>
</table>

\[ P(X) \rightarrow \]

\[ X \rightarrow \]
Question: What if our data provides no Y labels, but we believe \( P(Y, X_1, X_2, X_3, X_4) \) is defined by this Naïve Bayes net structure?

→ Unsupervised clustering : EM

\[
\begin{array}{c|cccc}
Y & X_1 & X_2 & X_3 & X_4 \\
\hline
? & 0.1 & 7.2 & 3.1 & 1.4 \\
? & 9.9 & 2.1 & 5.0 & 0.2 \\
? & 8.0 & 0.7 & 5.1 & 0.9 \\
? & 1.1 & 6.2 & 2.9 & 2.1 \\
? & 1.4 & 8.3 & 2.7 & 1.8 \\
\end{array}
\]

\[
\begin{array}{c|ccccc}
Pr & Y & X_1 & X_2 & X_3 & X_4 \\
\hline
0.8 & 1 & 0.1 & 7.2 & 3.1 & 1.4 \\
0.2 & 0 & 0.1 & 7.2 & 3.1 & 1.4 \\
0.3 & 1 & 9.9 & 2.1 & 5.0 & 0.2 \\
0.7 & 0 & 9.9 & 2.1 & 5.0 & 0.2 \\
0.4 & 1 & 8.0 & 0.7 & 5.1 & 0.9 \\
0.6 & 0 & 8.0 & 0.7 & 5.1 & 0.9 \\
0.7 & 1 & 1.1 & 6.2 & 2.9 & 2.1 \\
0.3 & 0 & 1.1 & 6.2 & 2.9 & 2.1 \\
0.6 & 1 & 1.4 & 8.3 & 2.7 & 1.8 \\
0.4 & 0 & 1.4 & 8.3 & 2.7 & 1.8 \\
\end{array}
\]
EM for Mixture of Gaussians Clustering

Let's simplify to make this easier:

1. Assume $X = \langle X_1 \ldots X_n \rangle$, and the $X_i$ are conditionally independent given $Z$. (the Naïve Bayes assumption).
   $P(X|Z = j) = \prod_i N(X_i|\mu_{ji}, \sigma_{ji})$

2. Assume only 2 clusters ($Z$ in $\{0, 1\}$), and $\forall i, j, \sigma_{ji} = \sigma$
   $P(X) = \sum_{j=1}^{2} P(Z = j|\pi) \prod_i N(x_i|\mu_{ji}, \sigma)$

3. Assume $\sigma$ known, $\pi_1 \ldots \pi_K, \mu_{li} \ldots \mu_{Ki}$ unknown

Observed: $X = \langle X_1 \ldots X_n \rangle$
Unobserved: $Z$
EM for Gaussian mixture model clustering

Given observed real-valued variables $X_i$, unobserved $Z$

where $\theta = \langle \pi, \mu_{ji} \rangle$

$\pi \equiv P(Z = 1)$

$\mu_{ji} \equiv \text{mean of Gaussian for } P(X_i | Z = j)$

Iterate until convergence:

- **E Step**: For each observed example $X(n)$, calculate $P(Z(n) \mid X(n), \theta)$

$$P(z(n) = k \mid x(n), \theta) = \frac{\prod_i N(x_i(n) \mid \mu_{k,i}, \sigma)}{\sum_j \prod_i N(x_i(n) \mid \mu_{j,i}, \sigma)} \left( \frac{\pi^k (1 - \pi)^{(1-k)}}{\pi^j (1 - \pi)^{(1-j)}} \right)$$

- **M Step**: Update

$$\pi \leftarrow \frac{1}{N} \sum_{n=1}^{N} E[z(n)]$$

$$\mu_{ji} \leftarrow \frac{\sum_{n=1}^{N} P(z(n) = j \mid x(n), \theta) \cdot x_i(n)}{\sum_{n=1}^{N} P(z(n) = j \mid x(n), \theta)}$$
Goal: Learn mixture distribution, interpreting $Z$ as cluster label

Learn $P(X_1, X_2 | \theta) =$

$P(Z=1 | \theta) P(X_1, X_2 | Z=1, \theta) \]

$+ P(Z=0 | \theta) P(X_1, X_2 | Z=0, \theta)$

<table>
<thead>
<tr>
<th>$Z$</th>
<th>$X_1$</th>
<th>$X_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>?</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>?</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Goal: Learn mixture distribution, interpreting Z as cluster label

Learn \( P(X_1, X_2 \mid \theta) = \)
\( P(Z=1 \mid \theta) P(X_1, X_2 \mid Z=1, \theta) \)
\( + P(Z=0 \mid \theta) P(X_1, X_2 \mid Z=0, \theta) \)

---

EM Algorithm

1. Choose any initial \( \theta \)

2. Iterate until convergence:
   - E Step: Use \( X \) and current \( \theta \) to calculate \( P(Z \mid X, \theta) \)
   - M Step: Replace current \( \theta \) by
     \( \theta \leftarrow \arg \max_{\theta'} {E_{P(Z \mid X, \theta)}[\log P(X, Z \mid \theta')]} \)

---

<table>
<thead>
<tr>
<th>Z</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>?</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>?</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Observed data \( X_1, X_2 \), unknown cluster assignment \( Z \)
Observed data $X_1, X_2$, unknown cluster assignment $Z$

Goal: Learn mixture distribution, interpreting $Z$ as cluster label

Learn $P(X_1, X_2 | \theta) =$

\[ P(Z=1 | \theta) P(X_1, X_2 | Z=1, \theta) + P(Z=0 | \theta) P(X_1, X_2 | Z=0, \theta) \]

EM Algorithm
1. Choose any initial $\theta$
2. Iterate until convergence:
   - **E Step:** Use $X$ and current $\theta$ to calculate $P(Z|X, \theta)$
   - **M Step:** Replace current $\theta$ by
     \[ \theta \leftarrow \arg \max_{\theta'} E_{P(Z|X, \theta)} \left[ \log P(X, Z|\theta') \right] \]

<table>
<thead>
<tr>
<th>Z</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>?</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>?</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

E-Step

<table>
<thead>
<tr>
<th>Probability</th>
<th>Z</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

... ... ...
Observed data $X_1, X_2$, unknown cluster assignment $Z$

Goal: Learn mixture distribution, interpreting $Z$ as cluster label

Learn $P(X_1, X_2 \mid \theta) =$
\[ P(Z=1 \mid \theta) \ P(X_1, X_2 \mid Z=1, \theta) + P(Z=0 \mid \theta) \ P(X_1, X_2 \mid Z=0, \theta) \]

<table>
<thead>
<tr>
<th>Probability</th>
<th>Z</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

EM Algorithm
1. Choose any initial $\theta$
2. Iterate until convergence:
   • E Step: Use $X$ and current $\theta$ to calculate $P(Z \mid X, \theta)$
   • M Step: Replace current $\theta$ by
     \[ \theta \leftarrow \text{arg max} \ E_{P(Z \mid X, \theta)}[\log P(X, Z \mid \theta')] \]

\[ \theta_{Z=1} \equiv P(Z = 1) \leftarrow \frac{1}{N} \sum_{n=1}^{N} P_{Z \mid X, \theta}(Z_n = 1) \]
\[ = \frac{0.8 + 0.3 + 0.6 + \ldots}{3 + \ldots} \]
Goal: Learn mixture distribution, interpreting $Z$ as cluster label

Learn $P(X_1, X_2 | \theta) =$

$P(Z=1 | \theta) P(X_1, X_2 | Z=1, \theta)$
$+ P(Z=0 | \theta) P(X_1, X_2 | Z=0, \theta)$

**EM Algorithm**

1. Choose any initial $\theta$
2. Iterate until convergence:

   - **E Step**: Use $X$ and current $\theta$ to calculate $P(Z|X, \theta)$
   - **M Step**: Replace current $\theta$ by $\theta \leftarrow \arg \max_{\theta'} E_{P(Z|X, \theta)}[\log P(X, Z|\theta')]$

<table>
<thead>
<tr>
<th>Probability</th>
<th>Z</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

... ... ... ...

M-Step $\theta_{Z=1}$

$$\theta_{Z=1} \equiv P(Z = 1) \leftarrow \frac{1}{N} \sum_{n=1}^{N} P_{Z|X, \theta}(Z_n = 1)$$

$$= \frac{0.8 + 0.3 + 0.6 + \ldots}{3 + \ldots}$$

Note if $Z$ observed, we would have

$$\theta_{Z=1} \equiv P(Z = 1) \leftarrow \frac{1}{N} \sum_{n=1}^{N} Z$$
Goal: Learn mixture distribution, interpreting Z as cluster label

Learn \( P(X_1, X_2 \mid \theta) = \)
\( P(Z=1 \mid \theta) P(X_1, X_2 \mid Z=1, \theta) \)
\( + P(Z=0 \mid \theta) P(X_1, X_2 \mid Z=0, \theta) \)

<table>
<thead>
<tr>
<th>Probability</th>
<th>Z</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

\( \mu_{X_i \mid Z=j} \)

\[
\mu_{X_i \mid Z=j} \leftarrow \frac{\sum_{n=1}^{N} P(Z_n = j) X_{i,n}}{\sum_{n=1}^{N} P(Z_n = j)}
\]

e.g.,

\[
\mu_{x_2 \mid Z=1} = \frac{0.8(-1.3) + 0.3(1.2) + 0.6(-0.6) + \ldots}{0.8 + 0.3 + 0.6 + \ldots}
\]
Goal: Learn mixture distribution, interpreting $Z$ as cluster label

Learn $P(X_1, X_2 | \theta) =$

\[ P(Z=1 | \theta) \ P(X_1, X_2 | Z=1, \theta) \]
\[ + \ P(Z=0 | \theta) \ P(X_1, X_2 | Z=0, \theta) \]

<table>
<thead>
<tr>
<th>Probability</th>
<th>Z</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0.9</td>
<td>-1.3</td>
</tr>
<tr>
<td>0.3</td>
<td>1</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>-1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>-0.4</td>
<td>-0.6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

M-Step: $\mu_{X_i|Z=j}$

\[ \mu_{X_i|Z=j} = \frac{\sum_{n=1}^{N} P(Z_n = j) X_{i,n}}{\sum_{n=1}^{N} P(Z_n = j)} \]

e.g.,

\[ \mu_{X_2|Z=1} = \frac{0.8(-1.3) + 0.3(1.2) + 0.6(-0.6) + \ldots}{0.8 + 0.3 + 0.6 + \ldots} \]
Final $P(Z) = [0.4893 \ 0.5107]$
Example: Mixture of Three (Spherical) Gaussians
EM assuming mixture of 3 Gaussian components: no conditional indep assumptions, so non-spherical Gaussians

10 iterations

20 iterations

60 iterations
EM assuming mixture of 3 Gaussian components: no conditional indep assumptions, so non-spherical Gaussians

10 iterations

20 iterations

60 iterations

2 components
EM assuming mixture of 3 Gaussian components: no conditional indep assumptions, so non-spherical Gaussians.
EM assuming mixture of 3 Gaussian components: no conditional indep assumptions, so non-spherical Gaussians

How should we choose the number of clusters?
How to choose number $k$ of clusters?

- We can try multiple values of $k$, evaluating each by the data likelihood $P(\text{Data} \mid k \text{ component mixture model})$.

- Note if we do this on the training data, the $k$ that maximizes $P(\text{trainData} \mid k \text{ component mixture model})$ will be $k = \text{number of training examples}$!

- Use held-out test data to chose $k$ $P(\text{testData} \mid k \text{ component mixture model})$. 
Applications of GMM in computer vision

1- Image segmentation:

\[ X = (R, G, B)^T \]
What you should know about EM mixture model clustering

- Another application of EM to learn from partially observed data
- Unobserved variable: cluster label
- Based on Bayes net that models mixture distribution
- Can use this for both discrete-valued, real-valued $X_i$
- Doesn’t answer the question of *how many* clusters to assume
  - But cross validation can reveal which choice is best on held-out data
Learning Bayes Net Structure
How can we learn Bayes Net graph structure?

In general case, open problem
• can require lots of data (else high risk of overfitting)
• can use Bayesian priors, or other kinds of prior assumptions about graph structure to constrain search

One key result:
• Chow-Liu algorithm: finds “best” tree-structured network
• What’s best?
  – suppose $P(X)$ is true distribution, $T(X)$ is distribution of our tree-structured network, where $X = <X_1, \ldots, X_n>$
  – Chow-Liu minimizes Kullback-Leibler divergence:

$$KL(P(X) \mid\mid T(X)) \equiv \sum_k P(X = k) \log \frac{P(X = k)}{T(X = k)}$$
Kullback-Leibler Divergence

- \( KL(P(X) \| T(X)) \) is a measure of the difference between probability distributions \( P(X) \) and \( T(X) \)

\[
KL(P(X) \| T(X)) \equiv \sum_k P(X = k) \log \frac{P(X = k)}{T(X = k)}
\]

- It is asymmetric, always greater or equal to 0
- It is 0 iff \( P(X) = T(X) \)
Chow-Liu Algorithm

**Key result:** To minimize \( KL(P \| T) \) over possible tree networks \( T \) approximating true \( P \), it suffices to find the tree network \( T \) that maximizes the sum of mutual informations over its edges.

Mutual information for an edge between variable \( A \) and \( B \):

\[
I(A, B) = \sum_a \sum_b P(a, b) \log \frac{P(a, b)}{P(a)P(b)}
\]

This works because for tree networks with nodes \( \mathbf{X} \equiv \langle X_1 \ldots X_n \rangle \)

\[
KL(P(\mathbf{X}) \| T(\mathbf{X})) \equiv \sum_k P(\mathbf{X} = k) \log \frac{P(\mathbf{X} = k)}{T(\mathbf{X} = k)}
\]

\[
= -\sum_i I(X_i, Pa(X_i)) + \sum_i H(X_i) - H(X_1 \ldots X_n)
\]
Chow-Liu Algorithm

1. for each pair of variables A, B, use training data to estimate \( P(A,B), \ P(A), \text{ and } P(B) \)

2. for each pair A, B calculate mutual information

\[
I(A, B) = \sum_a \sum_b P(a, b) \log \frac{P(a, b)}{P(a)P(b)}
\]

3. calculate the maximum spanning tree over the set of variables, using edge weights \( I(A, B) \)
   (given N vars, this costs only \( O(N^2) \) time)

4. add arrows to edges to form a directed-acyclic graph

5. learn the CPD’ss for this graph
Chow-Liu algorithm example
Greedy Algorithm to find Max-Spanning Tree

[courtesy A. Singh, C. Guestrin]
Bayes Nets – What You Should Know

• Representation
  – Bayes nets represent joint distribution as a DAG + Conditional Distributions
  – D-separation lets us decode conditional independence assumptions

• Inference
  – NP-hard in general
  – For some graphs, closed form inference is feasible
  – Approximate methods too, e.g., Monte Carlo methods, …

• Learning
  – Easy for known graph, fully observed data (MLE’s, MAP est.)
  – EM for partly observed data, known graph
  – Learning graph structure: Chow-Liu for tree-structured networks
  – Hardest when graph unknown, data incompletely observed