Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
Estimate $\theta$ from partly observed data — recap

- What if FAHN observed, but not S?
- Can’t calculate MLE
  \[ \theta \leftarrow \arg\max_\theta \log \prod_k P(f_k, a_k, s_k, h_k, n_k | \theta) \]
- Let X be all observed variable values (over all examples)
- Let Z be all unobserved variable values
- Can’t calculate MLE
  \[ \theta \leftarrow \arg\max_\theta \log P(X, Z | \theta) \]
- Estimate:
  \[ \theta \leftarrow \arg\max_\theta \mathbb{E}_{Z|X,\theta}[\log P(X, Z | \theta)] \]
**EM algorithm — informally**

- 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

\[
E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]
\]
E Step: Use $X, \theta$ to Calculate $P(Z|X,\theta)$

- observed $X=$\{F,A,H,N\}
- unobserved $Z=$\{S\}

- How? Bayes net inference problem

\[ P(S_k = 1|f_k, a_k, h_k, n_k, \theta) = \]
E Step: Use $X$, $\theta$ to Calculate $P(Z|X,\theta)$

- observed $X=${F,A,H,N}
- unobserved $Z=${S}

How? Bayes net inference problem

$$P(S_k = 1 | f_k a_k h_k n_k, \theta) = \frac{P(S_k = 1, f_k a_k h_k n_k | \theta)}{P(S_k = 1, f_k a_k h_k n_k | \theta) + P(S_k = 0, f_k a_k h_k n_k | \theta)}$$
EM and estimating $\theta_{s|i}$

- observed $X=\{F,A,H,N\}$; unobserved $Z=\{S\}$

- **E Step:** Calculate $P(Z_k|X_k; \theta)$ for each training example, $k$

  \[
P(S_k = 1|f_k, a_k, h_k, n_k, \theta) = E[s_k] = \frac{P(S_k = 1, f_k, a_k, h_k, n_k|\theta)}{P(S_k = 1, f_k, a_k, h_k, n_k|\theta) + P(S_k = 0, f_k, a_k, h_k, n_k|\theta)}
\]

- **M Step:** update all relevant parameters.

What was MLE? $\theta_{s|i} =$
EM and estimating $\theta_{s|ij}$

- observed $X=$\{F,A,H,N\}; unobserved $Z=$\{S\}

- **E Step:** Calculate $P(Z_k|X_k; \theta)$ for each training example, $k$

  $$P(S_k = 1 | f_k a_k h_k n_k, \theta) = E[s_k] = \frac{P(S_k = 1, f_k a_k h_k n_k | \theta)}{P(S_k = 1, f_k a_k h_k n_k | \theta) + P(S_k = 0, f_k a_k h_k n_k | \theta)}$$

- **M Step:** update all relevant parameters. For example:

  $$\theta_{s|ij} \leftarrow \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j) \ E[s_k]}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$

  Recall MLE was:

  $$\theta_{s|ij} = \frac{\sum_{k=1}^{K} \delta(f_k = i, a_k = j, s_k = 1)}{\sum_{k=1}^{K} \delta(f_k = i, a_k = j)}$$
Generalizing: EM and estimating $\theta$

- More generally, given observed set $X$, unobserved set $Z$ of boolean values

- **E Step**: Calculate for each training example, $k$ the expected value of each unobserved variable

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

\[
\delta(Y = 1) \rightarrow E_{Z|X,\theta}[Y] \quad \delta(Y = 0) \rightarrow (1 - E_{Z|X,\theta}[Y])
\]
Using (partially) 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>

semi-supervised learning
EM and estimating $\theta$

- E step: Calculate for each training example $k$, the expected value of each unobserved variable $Y$
EM and estimating $\theta$

- Observed set $X$
- Partially unobserved set $Y$ of boolean values

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

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

**M step:** Calculate estimates similar to MLE, but replacing each count by its expected count
EM and estimating $\theta$

- Observed set $X$
- Partially unobserved set $Y$ of boolean values

**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

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

MLE would be:

$$\hat{P}(X_i = j|Y = m) = \frac{\sum_k \delta((y(k) = m) \land (x_i(k) = j))}{\sum_k \delta(y(k) = m)}$$
EM algorithm — summary

- EM is a general procedure for learning from partly observed data
- Given observed variables $X$, unobserved $Z$ ($X=$\{F,A,H,N\}, $Z=$\{S\})
- Define \( Q(\theta'|\theta) = \mathbb{E}_{P(Z|X,\theta)}[\log P(X, Z|\theta')]\)

Begin with arbitrary choice for parameters $\theta$

Iterate until convergence:

- E Step: Use $X$ and current $\theta$ to calculate $P(Z|X,\theta)$
- M Step: Replace current $\theta$ by $\theta \leftarrow \text{arg max}_{\theta'} Q(\theta'|\theta)$

- Guaranteed to find local maximum. Each iteration increases

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

---

un
semi-supervised learning

---

<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</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>?</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>?</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>?</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>
Unsupervised clustering

Just extreme case of EM with zero labeled examples…