What is common amongst these problems

- Fitting $k$ clusters to a set of $N$ points
- Fitting $L$ lines to a set of points in 2-dim plane
- Tossing two coins and getting the probabilities of heads from each from the observations
- A tourist asking for direction from a person in a country where the inhabitants only lie or speak the truth
- Getting the arc transition probabilities in a probabilistic FSM
- WSD from comparable corpora of two languages in unsupervised setting
- Fitting Gaussian distributions to a set of points
Maximum Likelihood considerations
EM: What is it?

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

EM ("how to know when you do not completely know")

29th August, 2012
Parameter estimation: an exercise in maximization

Problem:- Given $N_h$ no of heads obtained out of $N$ trials, what is probability of obtaining head?

- In case of one coin
  - Let probability of obtaining head $= P_H$
    - This implies probability of obtaining exactly $N_h$ successes (heads) out of $N$ trials (tosses)
      \[
      f(p_h) = \binom{N}{N_h} \times p_h^{N_h} \times (1 - p_h)^{N - N_h}
      \]
Most “likely” value of $P_H$

To obtain the most likely value of $P_H$, we take $\ln$ of the above equation and differentiate wrt $P_H$:

$$g(P_h) = \ln f(P_h) = \frac{N}{N_h} \ln C + N_h \ln P_h + (N - N_h) \ln (1 - P_h)$$

$$\frac{d}{dP_h} g(P_h) = \frac{N_h}{P_h} - \frac{N - N_h}{1 - P_h} = 0$$

$$\Rightarrow P_h = \frac{N_h}{N}$$
Value of $P_H$ in absence of any information

Suppose we know nothing about the properties of a coin then what can we say about probability of head? We have to use the entropy $E$:

- Let $P_H$ be the probability of head
- Let $P_T$ be the probability of head

$$P_H + P_T = 1 \tag{1}$$

$$E = -P_H \log_2 P_H - P_T \log_2 P_T$$
Entropy

- Entropy is defined as sum of the multiplication of probability and log of probability with − sign. It is the instrument to deal with uncertainty.

- So best we can do is to maximize the entropy. Maximize $E$ subject to the eq (1) and get the value of $P_H$. 
Finding $P_H$ and $P_T$

$$f(P_H, P_T) = -P_H \log_2 P_H - P_T \log_2 P_T - \lambda(P_H + P_T - 1)$$

$$\frac{\delta F}{\delta \lambda} = P_H + P_T - 1 = 0 \quad \text{(1)}$$

$$\frac{\delta F}{S \ P_H} = -k \ln P_H - k - \lambda = 0 \quad \text{(2)}$$

$$\frac{\delta F}{S \ P_T} = -k \ln P_T - k - \lambda = 0 \quad \text{(3)}$$

From 2 and 3

$$-k \ln P_H - k - \lambda = -k \ln P_T - k - \lambda$$

$$\therefore P_H = P_T \quad \text{(4)}$$

From 4 and 1

$$P_H = P_T = \frac{1}{2}$$
A deeper look at EM

- Problem: two coins are tossed, randomly picking a coin at a time. The number of trials is $N$, number of heads is $N_H$ and number of tails is $N_T$.

- How can one estimate the following probabilities:
  - $p$: prob. Of choosing coin $1$
  - $p_1$: prob. Of head from coin $1$
  - $p_2$: prob. Of head from coin $2$
Expectation Maximization (1 Coin Toss)

- Toss 1 coin
  - K = Number of heads
  - N = Number of trials
- X = observation of tosses
  = \(<x_1>, <x_2>, <x_3>...<x_n>\) - each can take values 0 or 1
- p = probability of Head
  = \(\frac{1}{N} \sum_{i=1}^{N} x_i\)
  (as per MLE – maximizes probability of observed data)
Expectation Maximization (1 Coin Toss)

- $Y = \langle x_1, z_1 \rangle, \langle x_2, z_2 \rangle, \langle x_3, z_3 \rangle \ldots \langle x_i, z_i \rangle \ldots \langle x_n, z_n \rangle$
  - $x_i = 1$ for Head
  - $x_i = 0$ for Tail
  - $z_i$ = indicator function
    - $= 1$ if the observation comes from the coin
  - In this case, $z_i = 1 \ \forall i$

- $P = \frac{1}{N} \sum_{i=1}^{N} x_i z_i$
Expectation Maximization (2 coin toss)

- $X = \langle x_1 \rangle, \langle x_2 \rangle, \langle x_3 \rangle, \ldots \langle x_i \rangle, \ldots \langle x_n \rangle$
- $Y = \langle x_1, z_{11}, z_{12} \rangle, \langle x_2, z_{21}, z_{22} \rangle, \langle x_3, z_{31}, z_{32} \rangle, \ldots \langle x_i, z_{i1}, z_{i2} \rangle, \ldots \langle x_n, z_{n1}, z_{n2} \rangle$
  - $x_i = 1$ for Head
  - $= 0$ for Tail
  - $z_{i1} = 1$ if the observation comes from coin 1 else 0
  - $z_{i2} = 1$ if the observation comes from coin 2 else 0
  - only 1 of $z_{i1}$ and $z_{i2}$ can be 1
  - $x_i$ is observed while $z_{i1}$ and $z_{i2}$ is unobserved
Expectation Maximization (2 coin toss)

- Parameters of the setting
  - $p_1 =$ probability of Head for coin 1
  - $p_2 =$ probability of Head for coin 2
  - $p =$ probability of choosing for coin 1 for the toss

- Express $p$, $p_1$ and $p_2$ in terms of observed and unobserved data

\[
p_1 = \frac{\sum_{i=1}^{N} x_i z_{i1}}{\sum_{i=1}^{N} z_{i1}}
\]
\[
p_2 = \frac{\sum_{i=1}^{N} x_i z_{i2}}{\sum_{i=1}^{N} z_{i2}}
\]
\[
p = \frac{\sum_{i=1}^{N} z_{i1}}{\sum_{i=1}^{N} (z_{i1} + z_{i2})} = \frac{\sum_{i=1}^{N} z_{i1}}{N}
\]
Expectation Maximization trick

- Replace $z_{i1}$ and $z_{i1}$ in $p, p_1, p_2$ with $E(z_{i1})$ and $E(z_{i2})$
  - $z_{i1}$: event of $x = x_i$ given that observation is from coin 1
  - $E(z_{i1})$ = expectation of $z_{i1}$

\[
E(z_{i1}) = P(\text{coin} = \text{coin1} | x = x_i)
\]
\[
= \frac{P(\text{coin} = \text{coin1})P(x = x_i | \text{coin} = \text{coin1})}{P(x = x_i)}
\]
\[
= \frac{P(\text{coin} = \text{coin1})P(x = x_i | \text{coin} = \text{coin1})}{P(\text{coin} = \text{coin1})P(x = x_i | \text{coin} = \text{coin1}) + P(\text{coin} = \text{coin2})P(x = x_i | \text{coin} = \text{coin2})}
\]
\[
= \frac{p.p_1}{p.p_1 + (1 - p).p_2}
\]
Summary

- \( X = \langle x_1 \rangle, \langle x_2 \rangle, \langle x_3 \rangle \ldots \langle x_i \rangle \ldots \langle x_n \rangle \)
- \( Y = \langle x_1, z_{11}, z_{12} \rangle, \langle x_2, z_{21}, z_{22} \rangle, \langle x_3, z_{31}, z_{32} \rangle \ldots \langle x_i, z_{i1}, z_{i2} \rangle \ldots \langle x_n, z_{n1}, z_{n2} \rangle \)

\[
p_1 = \frac{\sum_{i=1}^{N} x_i z_{i1}}{\sum_{i=1}^{N} z_{i1}} \quad p_2 = \frac{\sum_{i=1}^{N} x_i z_{i2}}{\sum_{i=1}^{N} z_{i2}} \quad p = \frac{\sum_{i=1}^{N} z_{i1}}{\sum_{i=1}^{N} (z_{i1} + z_{i2})} = \frac{\sum_{i=1}^{N} z_{i1}}{N}
\]

\[
E(z_{i1}) = \frac{p.p_1}{p.p_1 + (1-p).p_2} \quad E(z_{i2}) = \frac{(1-p).p_2}{p.p_1 + (1-p).p_2}
\]
Observations

- Any EM problem has observed and unobserved data
- Nature of distribution
  - two coins follow two different binomial distributions
- Oscillation between E and M
  - convergence to local maxima or minima guaranteed
  - greedy algorithm
EM: Baum-Welch algorithm: counts

String = abb aaa bbb aaa

Sequence of states with respect to input symbols

State seq

Output seq
Calculating probabilities from table

\[ P(q \xrightarrow{a} r) = \frac{5}{8} \]

\[ P(q \xrightarrow{b} r) = \frac{3}{8} \]

\[ P(s^i \xrightarrow{w_k} s^j) = \frac{c(s^i \xrightarrow{w_k} s^j)}{\sum_{i=1}^{T} \sum_{m=1}^{A} c(s^i \xrightarrow{w_m} s^l)} \]

\( T = \# \text{states} \)
\( A = \# \text{alphabet symbols} \)

Now if we have a non-deterministic transitions then multiple state seq possible for the given o/p seq (ref. to previous slide’s feature). Our aim is to find expected count through this.

<table>
<thead>
<tr>
<th>Src</th>
<th>Dest</th>
<th>O/P</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>q</td>
<td>r</td>
<td>a</td>
<td>5</td>
</tr>
<tr>
<td>q</td>
<td>q</td>
<td>b</td>
<td>3</td>
</tr>
<tr>
<td>r</td>
<td>q</td>
<td>a</td>
<td>3</td>
</tr>
<tr>
<td>r</td>
<td>q</td>
<td>b</td>
<td>2</td>
</tr>
</tbody>
</table>
Interplay Between Two Equations

\[ P(s^i \xrightarrow{w_k} s^j) = \frac{C(s^i \xrightarrow{w_k} s^j)}{\sum_{l=0}^{T} \sum_{m=0}^{A} C(s^i \xrightarrow{w_m} s^l)} \]

\[ C(s^i \xrightarrow{w_k} s^j) = \sum_{s_{0,n+1}} P(S_{0,n+1} | W_{0,n}) \times n(s^i \xrightarrow{w_k} s^j, S_{0,n+1}, W_{0,n}) \]

No. of times the transitions \( s^i \rightarrow s^j \) occurs in the string
Illustration

Actual (Desired) HMM

Initial guess
One run of Baum-Welch algorithm: string `ababb`

<table>
<thead>
<tr>
<th>State sequences</th>
<th>P(path)</th>
<th>$q \xrightarrow{a} r$</th>
<th>$r \xrightarrow{b} q$</th>
<th>$q \xrightarrow{a} q$</th>
<th>$q \xrightarrow{b} q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon \rightarrow a$</td>
<td>$a \rightarrow b$</td>
<td>$b \rightarrow a$</td>
<td>$a \rightarrow b$</td>
<td>$b \rightarrow \epsilon$</td>
<td>$q \rightarrow r$</td>
</tr>
<tr>
<td>$q \rightarrow r$</td>
<td>$q \rightarrow q$</td>
<td>$q \rightarrow q$</td>
<td>$q \rightarrow q$</td>
<td>$0.00442$</td>
<td>$0.00442$</td>
</tr>
<tr>
<td>$q \rightarrow q$</td>
<td>$q \rightarrow q$</td>
<td>$r \rightarrow q$</td>
<td>$q \rightarrow q$</td>
<td>$0.00442$</td>
<td>$0.00442$</td>
</tr>
<tr>
<td>$q \rightarrow q$</td>
<td>$q \rightarrow q$</td>
<td>$q \rightarrow q$</td>
<td>$q \rightarrow q$</td>
<td>$0.02548$</td>
<td>$0.0$</td>
</tr>
</tbody>
</table>

Rounded Total $\rightarrow$

| $0.035$ | $0.01$ | $0.01$ | $0.06$ | $0.095$ |

New Probabilities (P) $\rightarrow$

| $0.06$ | $1.0$ | $0.36$ | $0.581$ |

$\epsilon$ is considered as starting and ending symbol of the input sequence string. Through multiple iterations the probability values will converge.
EM based unsupervised Approach
ESTIMATING SENSE DISTRIBUTIONS

If sense tagged Marathi corpus were available, we could have estimated

\[
P(S_{1}^{mar}|maan) = \frac{\#(S_{1}^{mar}, maan)}{\#(S_{1}^{mar}, maan) + \#(S_{2}^{mar}, maan)}
\]

But such a corpus is not available
Framework: Figure 1 and Figure 2
E-M steps

**E-step**

\[
P(S_{1mar} | maan) \\approx \frac{P(S_1^{hin} | gardan) \cdot \#(gardan) + P(S_1^{hin} | galaa) \cdot \#(galaa)}{Z}
\]

where, \(Z = P(S_1^{hin} | gardan) \cdot \#(gardan) + P(S_1^{hin} | galaa) \cdot \#(galaa) + P(S_3^{hin} | aadar) \cdot \#(aadar) + P(S_3^{hin} | izzat) \cdot \#(izzat)\)

**M-step**

\[
P(S_1^{hin} | galaa) \approx \frac{P(S_1^{mar} | maan) \cdot \#(maan) + P(S_1^{mar} | greeva) \cdot \#(greeva)}{Z}
\]

\[
Z = P(S_1^{mar} | maan) \cdot \#(maan) + P(S_1^{mar} | greeva) \cdot \#(greeva) + P(S_3^{mar} | aawaaaj) \cdot \#(aawaaaj) + P(S_3^{mar} | swar) \cdot \#(swar)
\]

where, \(S_1^{mar} = \pi_{hin}(S_1^{hin})\) (see Figure 1)

\(S_3^{mar} = \pi_{mar}(S_3^{hin})\) (see Figure 1)

\((maan, greeva) \in \text{translations}_{mar}(galaa, S_1^{hin})\) (see Figure 2)

\((aawaaaj, swar) \in \text{translations}_{mar}(galaa, S_2^{hin})\) (see Figure 2)
Points to note...

- Symmetric formulation
- $E$ and $M$ steps are identical except for the change in language
- Either can be treated as the E-step, making the other as the M-step
- A back-and-forth traversal over translation correspondences in the two languages
- Does not require parallel corpus – only in-domain corpus is needed
In General..

E-Step:

\[ P(S_{k}^{L1}|u) \approx \frac{\sum_{v} P(\pi_{L2}(S_{k}^{L1})|v) \cdot \#(v)}{\sum_{S_{i}^{L1}} \sum_{y} P(\pi_{L2}(S_{i}^{L1})|y) \cdot \#(y)} \]

where, \( S_{k}^{L1}, S_{i}^{L1} \in \text{synsets}_{L1}(u) \)
\[ v \in \text{translations}_{L2}(u, S_{k}^{L1}) \]
\[ y \in \text{translations}_{L2}(u, S_{i}^{L1}) \]

M-Step:

\[ P(S_{j}^{L2}|v) \approx \frac{\sum_{a} P(\pi_{L1}(S_{j}^{L2})|a) \cdot \#(a)}{\sum_{S_{j}^{L2}} \sum_{b} P(\pi_{L1}(S_{j}^{L2})|b) \cdot \#(b)} \]

where, \( S_{j}^{L2}, S_{i}^{L2} \in \text{synsets}_{L2}(v) \)
\[ a \in \text{translations}_{L1}(v, S_{j}^{L2}) \]
\[ b \in \text{translations}_{L1}(v, S_{i}^{L2}) \]
Experimental Setup

- Languages: Hindi, Marathi
- Domains: Tourism and Health (largest domain-specific sense tagged corpus)

<table>
<thead>
<tr>
<th>Category</th>
<th>Polysemous words Tourism</th>
<th>Polysemous words Health</th>
<th>Monosemous words Tourism</th>
<th>Monosemous words Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>62336</td>
<td>24089</td>
<td>35811</td>
<td>18923</td>
</tr>
<tr>
<td>Verb</td>
<td>6386</td>
<td>1401</td>
<td>3667</td>
<td>5109</td>
</tr>
<tr>
<td>Adjective</td>
<td>18949</td>
<td>8773</td>
<td>28998</td>
<td>12138</td>
</tr>
<tr>
<td>Adverb</td>
<td>4860</td>
<td>2527</td>
<td>13699</td>
<td>7152</td>
</tr>
<tr>
<td>All</td>
<td>92531</td>
<td>36790</td>
<td>82175</td>
<td>43322</td>
</tr>
</tbody>
</table>

Table 2: Polysemous and Monosemous words per category in each domain for Hindi

<table>
<thead>
<tr>
<th>Category</th>
<th>Avg. degree of WordNet polysemy for polysemous words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>3.02</td>
</tr>
<tr>
<td>Verb</td>
<td>5.05</td>
</tr>
<tr>
<td>Adjective</td>
<td>2.66</td>
</tr>
<tr>
<td>Adverb</td>
<td>2.52</td>
</tr>
<tr>
<td>All</td>
<td>3.09</td>
</tr>
</tbody>
</table>

Table 4: Average degree of WordNet polysemy per category in the 2 domains for Hindi

<table>
<thead>
<tr>
<th>Category</th>
<th>Avg. degree of WordNet polysemy for polysemous words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>3.02</td>
</tr>
<tr>
<td>Verb</td>
<td>4.96</td>
</tr>
<tr>
<td>Adjective</td>
<td>2.60</td>
</tr>
<tr>
<td>Adverb</td>
<td>2.44</td>
</tr>
<tr>
<td>All</td>
<td>3.14</td>
</tr>
</tbody>
</table>

Table 5: Average degree of WordNet polysemy per category in the 2 domains for Marathi
Algorithms Being Compared

- EM (our approach)
- Personalized PageRank \(^{(\text{Agirre and Soroa, 2009})}\)
- State-of-the-art bilingual approach (using Mutual Information) \(^{(\text{Kaji and Morimoto, 2002})}\)
- Random Baseline
- Wordnet First sense baseline (supervised baseline)
Results

- Performs better than other state-of-the-art knowledge based and unsupervised approaches
- Does not beat the Wordnet First Sense Baseline which is a supervised baseline

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Average N</th>
<th>Average R</th>
<th>Average A</th>
<th>Average V</th>
<th>Average O</th>
</tr>
</thead>
<tbody>
<tr>
<td>WFS</td>
<td>60.00</td>
<td>68.64</td>
<td>52.39</td>
<td>39.65</td>
<td>57.29</td>
</tr>
<tr>
<td>EM</td>
<td>53.35</td>
<td>56.95</td>
<td>51.39</td>
<td>29.98</td>
<td>51.26</td>
</tr>
<tr>
<td>PPR</td>
<td>56.17</td>
<td>0.00</td>
<td>38.94</td>
<td>29.74</td>
<td>48.88</td>
</tr>
<tr>
<td>RB</td>
<td>34.74</td>
<td>44.32</td>
<td>39.38</td>
<td>17.21</td>
<td>34.79</td>
</tr>
<tr>
<td>MI</td>
<td>10.97</td>
<td>3.89</td>
<td>10.07</td>
<td>5.63</td>
<td>9.97</td>
</tr>
</tbody>
</table>

Average 4-fold cross validation results averaged over all Language-Domain pairs for all words