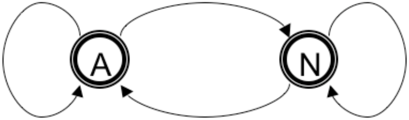
CMPT-413
Computational Linguistics

Anoop Sarkar
http://www.cs.sfu.ca/∼anoop

March 5, 2008

Hidden Markov Model

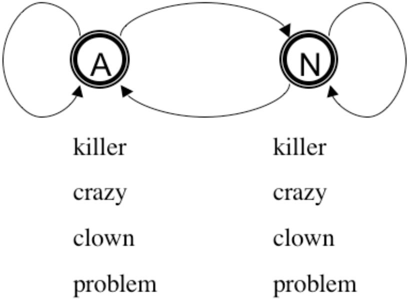
- Model $\theta = \{\pi_i, a_{i,j}, b_i(o)\}$
  - π_i: probability of starting at state i
  - $a_{i,j}$: probability of transition from state i to state j
  - $b_i(o)$: probability of output o at state i



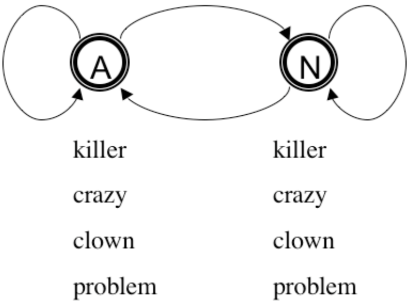
killer	killer
crazy	crazy
clown	clown
problem	problem

HMM Learning from Labeled Data

- Model $\theta = \{\pi_i, a_{i,j}, b_i(o)\}$
  - π_i: probability of starting at state i
  - $a_{i,j}$: probability of transition from state i to state j
  - $b_i(o)$: probability of output o at state i



HMM Learning from Labeled Data



- The task: to find the values for the parameters of the HMM:
  - π_A, π_N
  - $a_{A,A}, a_{A,N}, a_{N,N}, a_{N,A}$
  - $b_A(killer), b_A(crazy), b_A(clown), b_A(problem)$
  - $b_N(killer), b_N(crazy), b_N(clown), b_N(problem)$

Learning from Fully Observed Data

- Labeled Data:

```
x1,y1: killer/N clown/N        x3,y3: crazy/A problem/N
x2,y2: killer/N problem/N      x4,y4: crazy/A clown/N
```

- Let's say we have m labeled examples: $(x_1, y_1), \ldots, (x_m, y_m)$
- Each $(x_l, y_l) = \{o_1, \ldots, o_T, s_1, \ldots, s_T\}$
- For each (x_l, y_l) we can compute the probability using the HMM:
  - $(x_1, y_1) : \pi_N \cdot b_N(killer) \cdot a_{N,N} \cdot b_N(clown)$
  - $(x_2, y_2) : \pi_N \cdot b_N(killer) \cdot a_{N,N} \cdot b_N(problem)$
  - $(x_3, y_3) : \pi_A \cdot b_A(crazy) \cdot a_{A,N} \cdot b_N(problem)$
  - $(x_4, y_4) : \pi_A \cdot b_A(crazy) \cdot a_{A,N} \cdot b_N(clown)$

Learning from Fully Observed Data

- Labeled Data:

```
x1,y1: killer/N clown/N        x3,y3: crazy/A problem/N
x2,y2: killer/N problem/N      x4,y4: crazy/A clown/N
```

- We can easily collect frequency of observing a word with a state (tag)
  - $f(i, x, y)$ = number of times i is the initial state in (x, y)
  - $f(i, j, x, y)$ = number of times j follows i in (x, y)
  - $f(i, o, x, y)$ = number of times i is paired with observation o
- Then according to our HMM the probability of x, y is:

$$P(x, y) = \prod_{i:f(i,x,y)=1} \pi_i^{f(i,x,y)} \cdot \prod_{i,j} a_{i,j}^{f(i,j,x,y)} \cdot \prod_{i,o} b_i(o)^{f(i,o,x,y)}$$

Learning from Fully Observed Data

- According to our HMM the probability of x, y is:

$$P(x,y) = \prod_{i:f(i,x,y)=1} \pi_i^{f(i,x,y)} \cdot \prod_{i,j} a_{i,j}^{f(i,j,x,y)} \cdot \prod_{i,o} b_i(o)^{f(i,o,x,y)}$$

- The probability of the labeled data $(x_1, y_1), \ldots, (x_m, y_m)$ according to HMM with parameters θ is:

$$\begin{aligned} L(\theta) &= \sum_{l=1}^{m} \log P(x_l, y_l) \\ &= \sum_{l=1}^{m} \sum_{i:f(i,x,y)=1} f(i, x_l, y_l) \log \pi_i + \\ & \quad \sum_{i,j} f(i, j, x_l, y_l) \log a_{i,j} + \\ & \quad \sum_{i,o} f(i, o, x_l, y_l) \log b_i(o) \end{aligned}$$

Learning from Fully Observed Data

$$L(\theta) = \sum_{l=1}^{m} \sum_{i} f(i, x_l, y_l) \log \pi_i + \sum_{i,j} f(i, j, x_l, y_l) \log a_{i,j} + \sum_{i,o} f(i, o, x_l, y_l) \log b_i(o)$$

- $L(\theta)$ is the probability of the labeled data $(x_1, y_1), \ldots, (x_m, y_m)$
- We want to find a θ that will give us the maximum value of $L(\theta)$
- We find the θ such that $\frac{dL(\theta)}{d\theta} = 0$

Learning from Fully Observed Data

$$L(\theta) = \sum_{l=1}^{m}$$

$$\sum_{i} f(i, x_l, y_l) \log \pi_i + \sum_{i,j} f(i, j, x_l, y_l) \log a_{i,j} + \sum_{i,o} f(i, o, x_l, y_l) \log b_i(o)$$

- The values of $\pi_i, a_{i,j}, b_i(o)$ that maximize $L(\theta)$ are:

$$\pi_i = \frac{\sum_l f(i, x_l, y_l)}{\sum_l \sum_k f(k, x_l, y_l)}$$

$$a_{i,j} = \frac{\sum_l f(i, j, x_l, y_l)}{\sum_l \sum_k f(i, k, x_l, y_l)}$$

$$b_i(o) = \frac{\sum_l f(i, o, x_l, y_l)}{\sum_l \sum_{o' \in V} f(i, o', x_l, y_l)}$$

Learning from Fully Observed Data

- Labeled Data:

```
x1,y1: killer/N clown/N        x3,y3: crazy/A problem/N
x2,y2: killer/N problem/N      x4,y4: crazy/A clown/N
```

- The values of π_i that maximize $L(\theta)$ are:

$$\pi_i = \frac{\sum_l f(i, x_l, y_l)}{\sum_l \sum_k f(k, x_l, y_l)}$$

- $\pi_N = \frac{2}{4}$ and $\pi_A = \frac{2}{4}$ because:

$$\sum_l f(N, x_l, y_l) = 2$$

$$\sum_l f(A, x_l, y_l) = 2$$

Learning from Fully Observed Data

- Labeled Data:

```
x1,y1: killer/N clown/N        x3,y3: crazy/A problem/N
x2,y2: killer/N problem/N      x4,y4: crazy/A clown/N
```

- The values of $a_{i,j}$ that maximize $L(\theta)$ are:

$$a_{i,j} = \frac{\sum_l f(i,j,x_l,y_l)}{\sum_l \sum_k f(i,k,x_l,y_l)}$$

- $a_{N,N} = \frac{2}{4} = \frac{1}{2}$; $a_{N,A} = 0$; $a_{A,N} = \frac{1}{2}$ and $a_{A,A} = 0$ because:

$$\sum_l f(N,N,x_l,y_l) = 2 \qquad \sum_l f(A,N,x_l,y_l) = 2$$
$$\sum_l f(N,A,x_l,y_l) = 0 \qquad \sum_l f(A,A,x_l,y_l) = 0$$

Learning from Fully Observed Data

- Labeled Data:

```
x1,y1: killer/N clown/N        x3,y3: crazy/A problem/N
x2,y2: killer/N problem/N      x4,y4: crazy/A clown/N
```

- The values of $b_i(o)$ that maximize $L(\theta)$ are:

$$b_i(o) = \frac{\sum_l f(i, o, x_l, y_l)}{\sum_l \sum_{o' \in V} f(i, o', x_l, y_l)}$$

- $b_N(killer) = \frac{2}{6} = \frac{1}{3}$; $b_N(clown) = \frac{1}{3}$; $b_N(problem) = \frac{1}{3}$ and $b_A(crazy) = 1$ because:

$$\begin{aligned}
\sum_l f(N, killer, x_l, y_l) &= 2 & \sum_l f(A, killer, x_l, y_l) &= 0 \\
\sum_l f(N, clown, x_l, y_l) &= 2 & \sum_l f(A, clown, x_l, y_l) &= 0 \\
\sum_l f(N, crazy, x_l, y_l) &= 0 & \sum_l f(A, crazy, x_l, y_l) &= 2 \\
\sum_l f(N, problem, x_l, y_l) &= 2 & \sum_l f(A, problem, x_l, y_l) &= 0
\end{aligned}$$