CMPT-413: Computational Linguistics

HMM6: Supervised learning of Hidden Markov Models

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

February 28, 2013

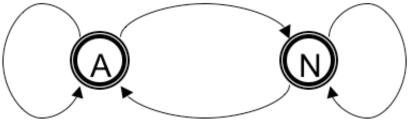
Hidden Markov Model Algorithms

- HMM as parser: compute the best sequence of states for a given observation sequence.
- HMM as language model: compute probability of given observation sequence.
- HMM as learner: given a corpus of observation sequences, learn its distribution, i.e. learn the parameters of the HMM from the corpus.
  - Learning from a set of observations with the sequence of states provided (states are not hidden) [Supervised Learning]
  - Learning from a set of observations without any state information. [Unsupervised Learning]

Hidden Markov Model

$$\text{Model } \theta = \begin{cases} \pi_i & \text{probability of starting at state } i \\ a_{i,j} & \text{probability of transition from state } i \text{ to state } j \\ b_i(o) & \text{probability of output } o \text{ at state } i \end{cases}$$

$$\text{Constraints}: \sum_i \pi_i = 1, \sum_j a_{i,j} = 1, \sum_o b_i(o) = 1$$



killer

crazy

clown

problem

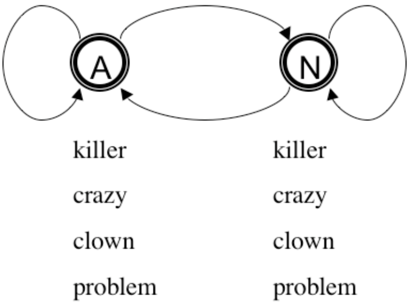
killer

crazy

clown

problem

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 L:

```
x1,y1: killer/N clown/N      (x1 = killer,clown; y1 = N,N)
x2,y2: killer/N problem/N    (x2 = killer,problem; y2 = N,N)
x3,y3: crazy/A problem/N     ...
x4,y4: crazy/A clown/N
x5,y5: problem/N crazy/A clown/N
x6,y6: clown/N crazy/A killer/N
```

Learning from Fully Observed Data

- Let's say we have m labeled examples:
  $L = (x_1, y_1), \ldots, (x_m, y_m)$
- Each $(x_\ell, y_\ell) = \{o_1, \ldots, o_T, s_1, \ldots, s_T\}$
- For each (x_ℓ, y_ℓ) we can compute the probability using the HMM:
  - $(x_1 = killer, clown; y_1 = N, N)$:
    $P(x_1, y_1) = \pi_N \cdot b_N(killer) \cdot a_{N,N} \cdot b_N(clown)$
  - $(x_2 = killer, problem; y_2 = N, N)$:
    $P(x_2, y_2) = \pi_N \cdot b_N(killer) \cdot a_{N,N} \cdot b_N(problem)$
  - $(x_3 = crazy, problem; y_3 = A, N)$:
    $P(x_3, y_3) = \pi_A \cdot b_A(crazy) \cdot a_{A,N} \cdot b_N(problem)$
  - $(x_4 = crazy, clown; y_4 = A, N)$:
    $P(x_4, y_4) = \pi_A \cdot b_A(crazy) \cdot a_{A,N} \cdot b_N(clown)$
  - $(x_5 = problem, crazy, clown; y_5 = N, A, N)$:
    $P(x_5, y_5) = \pi_N \cdot b_N(problem) \cdot a_{N,A} \cdot b_A(crazy) \cdot a_{A,N} \cdot b_N(clown)$
  - $(x_6 = clown, crazy, killer; y_6 = A, A, N)$:
    $P(x_6, y_6) = \pi_N \cdot b_N(clown) \cdot a_{N,A} \cdot b_A(crazy) \cdot a_{A,N} \cdot b_N(killer)$
- $\prod_\ell P(x_\ell, y_\ell) = {\pi_N}^4 \cdot {\pi_A}^2 \cdot {a_{N,N}}^2 \cdot {a_{N,A}}^2 \cdot {a_{A,N}}^4 \cdot {a_{A,A}}^0 \cdot b_N(killer)^3 \cdot b_N(clown)^4 \cdot b_N(problem)^3 \cdot b_A(crazy)^4$

Learning from Fully Observed Data

- 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:

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

Learning from Fully Observed Data

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

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

- For the labeled data $L = (x_1, y_1), \ldots, (x_\ell, y_\ell), \ldots, (x_m, y_m)$

$$\begin{aligned} P(L) &= \prod_{\ell=1}^{m} P(x_\ell, y_\ell) \\ &= \prod_{\ell=1}^{m} \left(\prod_i \pi_i^{f(i,x_\ell,y_\ell)} \cdot \prod_{i,j} a_{i,j}^{f(i,j,x_\ell,y_\ell)} \cdot \prod_{i,o} b_i(o)^{f(i,o,x_\ell,y_\ell)} \right) \end{aligned}$$

Learning from Fully Observed Data

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

$$P(L) = \prod_{\ell=1}^{m} \left(\prod_{i} \pi_i^{f(i,x_\ell,y_\ell)} \cdot \prod_{i,j} a_{i,j}^{f(i,j,x_\ell,y_\ell)} \cdot \prod_{i,o} b_i(o)^{f(i,o,x_\ell,y_\ell)} \right)$$

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

$$\begin{aligned} L(\theta) &= \sum_{\ell=1}^{m} \log P(x_\ell, y_\ell) \\ &= \sum_{\ell=1}^{m} \sum_{i} f(i, x_\ell, y_\ell) \log \pi_i + \\ &\quad \sum_{i,j} f(i, j, x_\ell, y_\ell) \log a_{i,j} + \\ &\quad \sum_{i,o} f(i, o, x_\ell, y_\ell) \log b_i(o) \end{aligned}$$

Learning from Fully Observed Data

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

$$\sum_{i} f(i, x_\ell, y_\ell) \log \pi_i + \sum_{i,j} f(i, j, x_\ell, y_\ell) \log a_{i,j} + \sum_{i,o} f(i, o, x_\ell, y_\ell) \log b_i(o)$$

- $\theta = (\pi, a, b)$
- $L(\theta)$ 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 $L(\theta)$
- Find the θ such that $\frac{dL(\theta)}{d\theta} = 0$

Learning from Fully Observed Data

$$L(\theta) = \sum_{\ell=1}^{m}$$
$$\sum_i f(i, x_\ell, y_\ell) \log \pi_i + \sum_{i,j} f(i, j, x_\ell, y_\ell) \log a_{i,j} + \sum_{i,o} f(i, o, x_\ell, y_\ell) \log b_i(o)$$

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

$$\begin{aligned}
\pi_i &= \frac{\sum_\ell f(i, x_\ell, y_\ell)}{\sum_\ell \sum_k f(k, x_\ell, y_\ell)} \\
a_{i,j} &= \frac{\sum_\ell f(i, j, x_\ell, y_\ell)}{\sum_\ell \sum_k f(i, k, x_\ell, y_\ell)} \\
b_i(o) &= \frac{\sum_\ell f(i, o, x_\ell, y_\ell)}{\sum_\ell \sum_{o' \in V} f(i, o', x_\ell, y_\ell)}
\end{aligned}$$

Learning from Fully Observed Data

- Labeled Data:

```
x1,y1: killer/N clown/N
x2,y2: killer/N problem/N
x3,y3: crazy/A problem/N
x4,y4: crazy/A clown/N
x5,y5: problem/N crazy/A clown/N
x6,y6: clown/N crazy/A killer/N
```

Learning from Fully Observed Data

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

$$\pi_i = \frac{\sum_\ell f(i, x_\ell, y_\ell)}{\sum_\ell \sum_k f(k, x_\ell, y_\ell)}$$

- $\pi_N = \frac{2}{3}$ and $\pi_A = \frac{1}{3}$ because:

$$\sum_\ell f(N, x_\ell, y_\ell) = 4$$
$$\sum_\ell f(A, x_\ell, y_\ell) = 2$$

Learning from Fully Observed Data

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

$$a_{i,j} \quad = \quad \frac{\sum_{\ell} f(i,j,x_\ell,y_\ell)}{\sum_{\ell}\sum_{k} f(i,k,x_\ell,y_\ell)}$$

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

$$\begin{aligned}
\sum_{\ell} f(N,N,x_\ell,y_\ell) &= 2 \qquad & \sum_{\ell} f(A,N,x_\ell,y_\ell) &= 4 \\
\sum_{\ell} f(N,A,x_\ell,y_\ell) &= 2 \qquad & \sum_{\ell} f(A,A,x_\ell,y_\ell) &= 0
\end{aligned}$$

Learning from Fully Observed Data

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

$$b_i(o) \quad = \quad \frac{\sum_\ell f(i, o, x_\ell, y_\ell)}{\sum_\ell \sum_{o' \in V} f(i, o', x_\ell, y_\ell)}$$

- $b_N(killer) = \frac{3}{10}$; $b_N(clown) = \frac{4}{10}$; $b_N(problem) = \frac{3}{10}$ and $b_A(crazy) = 1$ because:

$$\begin{aligned}
\sum_\ell f(N, killer, x_\ell, y_\ell) &= 3 & \sum_\ell f(A, killer, x_\ell, y_\ell) &= 0 \\
\sum_\ell f(N, clown, x_\ell, y_\ell) &= 4 & \sum_\ell f(A, clown, x_\ell, y_\ell) &= 0 \\
\sum_\ell f(N, crazy, x_\ell, y_\ell) &= 0 & \sum_\ell f(A, crazy, x_\ell, y_\ell) &= 4 \\
\sum_\ell f(N, problem, x_\ell, y_\ell) &= 3 & \sum_\ell f(A, problem, x_\ell, y_\ell) &= 0
\end{aligned}$$

Learning from Fully Observed Data

```
x1,y1: killer/N clown/N
x2,y2: killer/N problem/N
x3,y3: crazy/A problem/N
x4,y4: crazy/A clown/N
x5,y5: problem/N crazy/A clown/N
x6,y6: clown/N crazy/A killer/N
```

$\pi =$

A	0.25
N	0.75

$a =$

$a_{i,j}$	A	N
A	0.0	1.0
N	0.5	0.5

$b =$

$b_i(o)$	*clown*	*killer*	*problem*	*crazy*
A	0	0	0	1
N	0.4	0.3	0.3	0