CS 485: Mathematical Foundations for the Information Age

Lecture 38 ▪ April 22, 2009

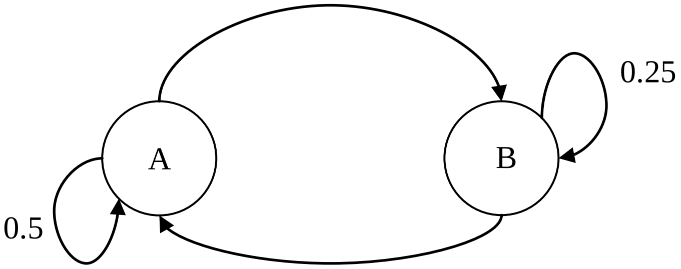
Jeff Davidson (jpd236) and Jeff Pankewicz (jhp36)

Approximate Counting Algorithm:
Problem – in a data stream, count the number of occurrences of a given symbol.
If there are n occurrences, this takes log n bits. We can approximate a solution by storing k where $n = 2^k$. This only takes log log n bits.

The trick is to probabilistically increment k by flipping a coin with probability of heads equal to $1/2^k$. Then at any given point, it will take an expected 2^k coin flips to actually increase k (and double the estimation), so the estimate is kept roughly close to the actual value.

Hidden Markov Models:

1. A finite state graph with transition probabilities
2. Initial state probability distribution π
3. Transition probabilities a_{ij}
4. Output symbol probability distribution for each state



A move consists of two steps:
When in a given state, flip a coin to go to another state. Then, flip another coin to select an output according to the probability distribution of that output (this distribution is based only on the state and does not depend on the previous state from which the transition occurred)

Outputs O:
$O_0O_1O_2\ldots O_T$, where T is the number of transitions (note that there is one more output symbol than the number of transitions)

Three questions we might ask:
1. What is the probability of a given output sequence?
2. What is the most likely sequence of states to produce a given output (Viterbi algorithm)?
3. Given an upper bound on the number of states and watching the outputs for a long time, what is the model?

Probability of an output sequence given a Hidden Markov Model:
$O_0O_1O_2\ldots O_T$
Probability of seeing first symbol $O_0 = \sum_i \pi(i) P(O_0|i)$

What is the probability of being in state j given O_0O_1?

Let $b_j(O_i)$ = probability of output O_i given arrived in state j

Record the maximum value and where you came from to get that value, then trace the path back to find the sequence of states that is most likely (one would do this by, as moving forward to build the table, keeping track of the backward arrows)

$$\left(\sum_i a_{ij}\pi(i)b_i(O_0)\right)b_j(O_1)$$

We want to discover:

- a_{ij}, the transition probability
- $b_j(O_{t+1})$, the probability of outputting O_{t+1} given that the Hidden Markov Model is in state j at time t+1

We also have the parameters:

- $\alpha_t(i)$, the probability of seeing $O_0O_1O_2\ldots O_T$ ending in state i at time t
- $\beta_{t+1}(j)$, the probability of seeing remainder of sequence given state j at time t+1

This is clearly an NP-complete problem. We can approximate it by getting initial guesses for a_{ij} and $b_j(O_{t+1})$, assigning probabilities, and calculating probabilities of being in various stages. The values should converge to a local minimum, if not the actual minimum.

$$a_{ij} = \frac{\text{expected \# of times through edge (i,j)}}{\text{expected \# of times at i}}$$

To be continued on Friday.