CPS260/BGT204.1 Algorithms in Computational Biology **October 16, 2003**

Lecture 14: Hidden Markov Models

Lecturer: Ron Parr *Scribe: Wenbin Pan*

In the last lecture we studied probability theories, and using probabilities as predictions of some events, like the probability that Bush will win the second run for the U.S. president. However, the predictions we have looked so far are mostly atemporal. That is, when predicting the probability of event A, we do not consider the events that happen before A and might have an influence on A. Thus in real world, for example when tracking an airplane, the probability distribution of a plane's position at time t has a close connection with the plane's position at time $t - \epsilon$. In order to describe our world that changes over time, we need a much stronger model.

14.1 Definition of HMM

14.1.1 States

First we introduce the notion of *atomic event*. An atomic event is an assignment to every random variable in the domain. For example, "it is raining today" and "it is not raining today" are two atomic events. We can use a binary variable $raining$ to describe these two events. If it is raining, we assign $raining$ to 1. Apparently for n random variables, there are 2^n possible atomic events.

States are atomic events that can transfer from one to another. Suppose a model has n states $\{S1, S2, ..., Sn\}$, we can describe how a system behaves with a state-transition diagram.

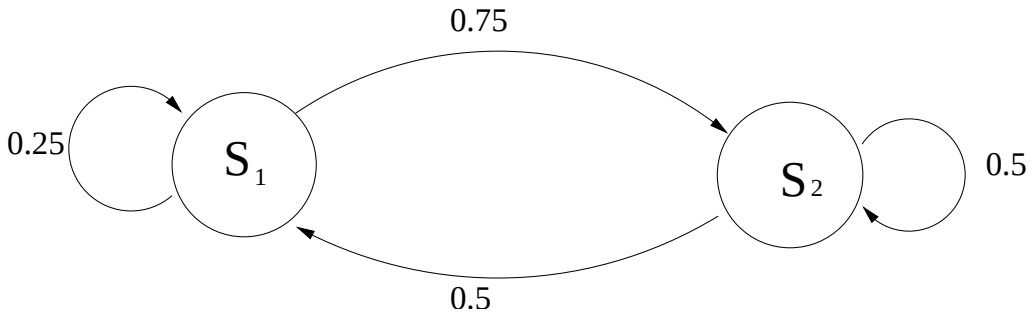

Figure 14.1: state-transition diagram

In this diagram, $P(Si|Sj), (1 \leq i, j \leq n)$ are called *transition probabilities*. Transitions among the states are

governed by these transition probabilities. If we consider that time moves in uniform, discrete increments, $P(Si|Sj)$ represent the probability that in time $t+1$, the system is in state Si, given that in time t, the system is in state Sj. For example in the diagram, in a time interval t if the system is in state $S1$, then in time $t+1$, there is a $\frac{3}{4}$ probability that the system is in state $S2$, and a $\frac{1}{4}$ probability that the system is still in state $S1$.

Notice that transition probabilities should also satisfy the normal stochastic constraints.

$$0 \leq P(Si|Sj) \leq 1,\ (1 \leq i, j \leq n)$$

and

$$\sum_{i=1}^{n} P(Si|Sj) = 1,\ (1 \leq j \leq n)$$

The second formula shows that for any state Sj, the sum of the probabilities that Sj will transfer to a state is 1. However, formula $\sum_{j=1}^{n} P(Si|Sj) = 1,\ (1 \leq i \leq n)$ does not necessarily hold for all i's. That is to say, the sum of the probabilities that a state will transfer to state Si does not have to be 1. If it is larger than 1, the system has a little higher probability to be in state Si.

14.1.2 Markov Model

In the state-transition diagram, we actually make the following assumptions:

- Transition probabilities are stationary. They do not change over times. In our diagram, for any time t, if the system is in state $S1$, then in time $t+1$, there is always a $\frac{3}{4}$ probability that the system is in state $S2$, no matter if $t = 0$ or $t = 1000$. This is often called the stationary assumption.
- The event space does not change over time. We will not get a new state as time goes on.
- Probability distribution over next states depends only on the current state.

The third assumption is the famous *Markov Assumption*. As given in the definition, transition probabilities are

$$P(Si|Sj), (1 \leq i, j \leq n)$$

In other words it is assumed that the next state is dependent only upon the current state. Mathematically, let S_t be a random variable for the state at time *t*(notice that S_t is different from St, which is the *t*-th state), then

$$P(S_t|S_{t-1}, ..., S_0) = P(S_t|S_{t-1})$$

Actually, Markov assumption is a special kind of conditional independence. It shows that given the current state, future state is independent of all past states. It seems that this assumption is very limited, but actually most cases of the real world can satisfy this assumption given our states are well defined. Target tracking, patient monitoring and speech recognition are all this kind of applications.

A system with states that obey the Markov assumption is called a *Markov Model*. A sequence of states resulting from such a model is called a *Markov Chain*. Markov model has a very nice property that its description can be maintained within quadratic space(as to the number of states in the model). Potentially

we can get an infinite time sequence.

We can use a transition matrix P to describe a Markov model. The (i, j)-th entry of this matrix is $P(Sj|Si)$. Then the properties of the system can be analyzed in terms of properties of the transition matrix. Suppose

$$x = (x_1, x_2, \ldots, x_n)$$

is the distribution over states at time t, where x_i is the probability of state Si(of course here $\sum_{i=1}^{n} x_i = 1$ should hold.) Then we can compute the distribution at time $t + 1$

$$(\sum_{j=1}^{n} x_j \cdot P(S1|Sj), \sum_{j=1}^{n} x_j \cdot P(S2|Sj), \ldots, \sum_{j=1}^{n} x_j \cdot P(Sn|Sj)) = xP$$

Similarly, the distribution over states at time $t + k$ can be calculated as xP^k.

14.1.3 Hidden Markov Models

In the Markov Model we introduce E_t as the outcome or *observation* at time t. Observations are generated according to the associated probability distribution. Given the current state S, the probability we have the observation E is defined as emission probability $P(E|S)$. Here we also make the stationary assumption, that emission probabilities do not change over time. Besides, very similar to the Markov assumption, we assume that the current observation is only depended upon the current state. Or in an other word, observations are conditionally independent of other variables given the current state. Mathematically this assumption is represented as

$$P(E_t|S_t, S_{t-1}, ..., S_0) = P(E_t|S_t)$$

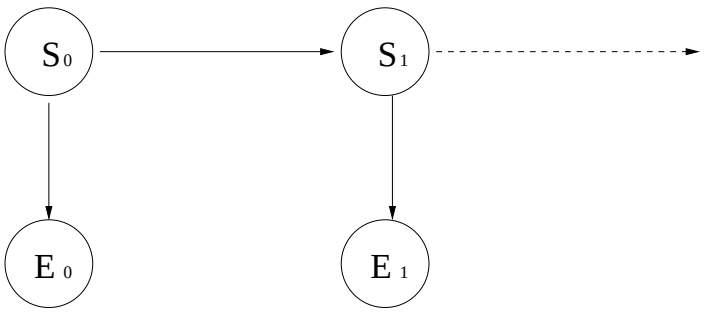

Figure 14.2: graphical model of HMMs

Notice that to an external observer, only the observations E are visible, the states S are hidden to the outside. This is where the name Hidden Markov Models comes from.

14.2 Use of HMMs

14.2.1 Basic Problems

Given a hidden Markov model and an observation sequence $E = e_1, e_2, \ldots, e_t, (1 \leq t \leq T)$ generated by this model, we can get the following information of the corresponding Markov chain

- We can compute the current hidden states S_T. This is often called *monitoring* or *filtering*. This is most useful in the problem like patient monitoring. Here the symptoms of the patient are our observations. The object is to know the current health status of the patient based on the observations to decide the treatment.
- We can predict the future observations e_t for $t > T$. Radar tracking problems is mostly like this prediction problem. The past positions of the plane are our observations. We want to predict the future position of this plane so that our radar will not lose it.
- We can update our view of past state $S_t, t < T$ based on the observation sequence. This problem is a *smoothing* or *hindsight* problem. Suppose we observed a car crash at time $t = 20$, we want to know the status of the car at time $t = 5$ to see the probability that the driver made a mistake at that time. This is an example of hindsight problem.
- A similar problem to filtering/hindsight problem is one of finding the most likely path through the state space. That is, what is the most likely sequence of events (from start to finish) to explain what we have seen. Like in the car crash problem, we now want to study all the status of the car from time $t = 1$ to 20 to study what caused the crash.

14.2.2 Extended Bayes' Rule

Before we study the problems, we first extend the Bayes' Rule

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

by introducing extra evidence C

$$P(A|BC) = \frac{P(B|AC) \cdot P(A|C)}{P(B|C)}$$

This can be interpreted as we limit the whole probability space to a corner where C always holds, and the every other things behave the same.

14.2.3 Monitoring

In the monitoring problem, actually we want

$$P(S_T|e_T \ldots e_1)$$

Using extended Bayes' rule, we can write the notation as

$$P(S_T|e_T \dots e_1) = \frac{P(e_T|S_T e_{T-1} \dots e_1) \cdot P(S_T|e_{T-1} \dots e_1)}{P(e_T|e_{T-1} \dots e_1)}$$

Since e_T is only dependent upon S_T, the $P(e_T|S_T e_{T-1} \dots e_1)$ factor in the numerator can be written as $P(e_T|S_T)$. And since the denominator $P(e_T|e_{T-1} \dots e_1)$ is independent to S_T, we can also write $P(S_T|e_T \dots e_1)$ as proportional to the numerator. Thus the equation changes to

$$\begin{aligned} P(S_T|e_T \dots e_1) &= \alpha P(e_T|S_T) \cdot P(S_T|e_{T-1} \dots e_1) \\ &= \alpha P(e_T|S_T) \cdot \sum_{S_{T-1}} (P(S_T|S_{T-1}) \cdot P(S_{T-1}|e_{T-1} \dots e_1)) \end{aligned}$$

Now we have a recursive relation so that we can calculate $P(S_T|e_T \dots e_1)$.

Now let's look at an example. Suppose a professor want to know if his student is working by if the student has produce some results. This problem is just a monitoring problem. Hidden state W stands for the student is working. Observation R stands for the student has produce some results. Then we can construct a hidden Markov model. Suppose the transition probabilities and emission probabilities are

$$\begin{aligned} P(W_{t+1}|W_t) &= 0.8 \\ P(W_{t+1}|\overline{W}_t) &= 0.3 \\ P(R_t|W_t) &= 0.6 \\ P(R_t|\overline{W}_t) &= 0.2 \end{aligned}$$

Suppose the professor knows that the student starts in a working state W_0. And after that, the professor has observed two consecutive months without results $\overline{R_1}$ and $\overline{R_2}$. The professor want to know if the student was working in the second month.

Using the recursive relation, first we calculate $P(W_1)$ and $P(\overline{W}_1)$

$$\begin{aligned} P(W_1|\overline{R}_1) &= \alpha_1 \cdot P(\overline{R}_1|W_1) \sum_{W_0} (P(W_1|W_0)P(W_0)) \\ &= \alpha_1 \cdot 0.4(0.8 \times 1.0 + 0.3 \times 0.0) \\ &= 0.32\alpha_1 \end{aligned}$$

$$\begin{aligned} P(\overline{W}_1|\overline{R}_1) &= \alpha_1 \cdot P(\overline{R}_1|W_1) \sum_{W_0} (P(\overline{W}_1|W_0)P(W_0)) \\ &= \alpha_1 \cdot 0.8(0.2 \times 1.0 + 0.7 \times 0.0) \\ &= 0.16\alpha_1 \end{aligned}$$

Use $P(W_1|\overline{R}_1) + P(\overline{W}_1|\overline{R}_1) = 1$ to get

$$\begin{aligned} P(W_1|\overline{R}_1) &= 0.67 \\ P(\overline{W}_1|\overline{R}_1) &= 0.32 \end{aligned}$$

Similarly, $P(W_1)$ and $P(\overline{W}_1)$ are

$$\begin{aligned} P(W_2|\overline{R}_2\overline{R}_1) &= \alpha_2 \cdot P(\overline{R}_2|W_2)\sum_{W_1}(P(W_2|W_1)P(W_1)) \\ &= \alpha_2 \cdot 0.4(0.8 \times 0.67 + 0.3 \times 0.33) \\ &= 0.25\alpha_2 \end{aligned}$$

$$\begin{aligned} P(\overline{W}_2|\overline{R}_2\overline{R}_1) &= \alpha_2 \cdot P(\overline{R}_2|\overline{W}_2)\sum_{W_1}(P(\overline{W}_2|W_1)P(W_1)) \\ &= \alpha_2 \cdot 0.8(0.2 \times 0.67 + 0.7 \times 0.33) \\ &= 0.292\alpha_2 \end{aligned}$$

$$\begin{aligned} P(W_2|\overline{R}_2\overline{R}_1) &= 0.46 \\ P(\overline{W}_2|\overline{R}_2\overline{R}_1) &= 0.54 \end{aligned}$$

So the probability that the student is working in the second month is only 0.46.

14.2.4 Hindsight

Now let's look at the hindsight problem. As explained before, hindsight problem is to calculate

$$P(S_t|e_T \dots e_1),\ 1 \le t < T$$

Similarly to the monitoring problem, we have

$$\begin{aligned} P(S_t|e_T \dots e_1) &= \frac{P(e_T \dots e_{t+1}|S_t e_t \dots e_1) \cdot P(S_t|e_t \dots e_1)}{P(e_T \dots e_{t+1}|e_t \dots e_1)} \\ &= \alpha P(e_T \dots e_{t+1}|S_t e_t \dots e_1) \cdot P(S_t|e_t \dots e_1) \\ &= \alpha P(e_T \dots e_{t+1}|S_t) \cdot P(S_t|e_t \dots e_1) \end{aligned}$$

For $P(e_T \dots e_{t+1}|S_t)$, we can write a recursive relation to solve it

$$\begin{aligned} P(e_T \dots e_{t+1}|S_t) &= \sum_{S_{t+1}} P(e_T \dots e_{t+1}|S_t S_{t+1})P(S_{t+1}|S_t) \\ &= \sum_{S_{t+1}} P(e_T \dots e_{t+1}|S_{t+1})P(S_{t+1}|S_t) \\ &= \sum_{S_{t+1}} P(e_{t+1}|S_{t+1})P(e_T \dots e_{t+2}|S_{t+1})P(S_{t+1}|S_t) \end{aligned}$$

So backward smoothing requires two recursive passes. One for $P(S_t|e_t \dots e_1)$(from 1 to t), and one for $P(e_T \dots e_{t+1}|S_t)$(from $T-1$ to t).

Now back to our student problem. The professor observed that the student has produced no result for two consecutive months. Then what's the probability that the student was working in the first month.

We have calculated $P(W_1|\overline{R}_1)$ and $P(\overline{W}_1|\overline{R}_1)$, and

$$\begin{aligned} P(\overline{R}_2|W_1) &= \sum_{W_2}(P(\overline{R}_2|W_2)P(W_2|W_1)) \\ &= (0.4 \times 0.8 + 0.8 \times 0.2) \\ &= 0.48 \end{aligned}$$

$$\begin{aligned} P(\overline{R}_2|\overline{W}_1) &= \sum_{W_2}(P(\overline{R}_2|W_2)P(W_2|\overline{W}_1)) \\ &= (0.4 \times 0.3 + 0.8 \times 0.7) \\ &= 0.68 \end{aligned}$$

$$P(\overline{W}_1|\overline{R}_1\overline{R}_2) = \alpha 0.33 \times 0.68 = 0.4556\alpha = 0.5864$$
$$P(W_1|\overline{R}_1\overline{R}_2) = \alpha 0.67 \times 0.48 = 0.3216\alpha = 0.4138$$

In monitoring and hindsight, we both calculated the probability that the student is working in the first month. However, this probability drop from 0.67 to 0.4138. This is because in monitoring, when we calculate the probability, we only use the observation $\overline{R}_1$ that the student has no result in the first month. When hindsight, we have additional observation $\overline{R}_2$ that the student still has no result in the second month. In this case, the probability calculated in hindsight is more likely to be the fact.

14.2.5 Most Likely Path

We want to find the most likely state sequence for a given sequence of observations. The solution to this problem depends upon the way 'most likely state sequence' is defined. One approach is to find the most likely state S_t at $1 \leq t \leq T$ and to concatenate all such S_t's. However, sometimes this method does not give a physically meaningful state sequence. Therefore we would go for another method which has no such problems.

In this method, commonly known as *Viterbi algorithm*, the whole state sequence with the maximum likelihood is found. The main observation is that the most likely path from $S_0, \ldots, S_k, \ldots, S_t$ can be decomposed as the most likely path from $S_0, \ldots, S_k$ combined with the most like path from $S_k, \ldots, S_t$. So we can build the most likely path recursively.

Most likely path is very useful for reconstructing events depending on the observation. In the car crash problem, we can get the most likely status of the car in each time interval and try to find out the reason of the accident. Another example is speech recognition. Most likely path will get a word sequence that is most probable.

14.2.6 Learning of HMMs

Generally, the learning problem is how to adjust the HMM parameters, so that the given set of observations (called the training set) is represented by the model in the best way for the intended application. Typically, we

know the time sequences of states and observations. The goal is to find the transition probability and emission probability for the hidden Markov model. Thus it would be clear that the 'quantity' we wish to optimize during the learning process can be different from application to application. In other words there may be several optimization criteria for learning, out of which a suitable one is selected depending on the application.

Such conditional probabilities is always computed by maximum likelihood. An EM-algorithm is used in such condition. The basic idea is to start with an initial guess about probabilities and then modify our guess. When we have a guess about the probabilities, we can use forward/backward method to compute the probability distribution over every states. Then we can update our transition probabilities based on state occupancy probabilities.

EM-algorithm will always converge. However, it may not get the optimal result, but stuck in a local optima instead. The initial guess may be crucial in this case.