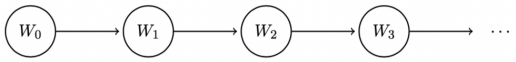
# Hidden Markov Models

## Markov Models

In previous notes, we talked about Bayes' nets and how they are a wonderful structure used for compactly representing relationships between random variables. We'll now cover a very intrinsically related structure called a **Markov model**, which for the purposes of this course can be thought of as analogous to a chainlike, infinite-length Bayes' net. The running example we'll be working with in this section is the day-to-day fluctuations in weather patterns. Our weather model will be time-dependent (as are Markov models in general), meaning we'll have a separate random variable for the weather on each day. If we define $W_i$ as the random variable representing the weather on day i, the Markov model for our weather example would look like this




What information should we store about the random variables involved in our Markov model? To track how our quantity under consideration (in this case, the weather) changes over time, we need to know both it's **initial distribution** at time t = 0 and some sort of **transition model** that characterizes the probability of moving from one state to another between timesteps. The initial distribution of a Markov model is enumerated by the probability table given by P(W0) and the transition model of transitioning from state i to i+1 is given by $P(W_{i+1}|W_i$. Note that this transition model implies that the value of $W_{i+1}$ is conditionally dependent only on the value of $W_i$. In other words, the weather at time t = i + 1 satisfies the Markov property or memoryless property, and is independent of the weather at all other timesteps besides t = i. Using our Markov model for weather, if we wanted to reconstruct the joint between $W_0$, $W_1$, and $W_2$ using the chain rule, we would want:

$$P(W_0, W_1, W_2) = P(W_0)P(W_1|W_0)P(W_2|W_1, W_0)$$

However, with our assumption that the Markov property holds true and $W_0$ is independent of $W_2|W_1$, the joint simplifies to:

$$P(W_0, W_1, W_2) = P(W_0)P(W_1|W_0)P(W_2|W_1)$$

And we have everything we need to calculate this from the Markov model. More generally, Markov models make the following independence assumption at each timestep: $W_{i+1}$ is independent of $W_0, ..., W_{i-1}|W_i$. This allows us to reconstruct the joint distribution for the first n+1 variables via the chain rule as follows:

$$P(W_0, W_1, ..., W_n) = P(W_0)P(W_1|W_0)P(W_2|W_1)...P(W_n|W_{n-1}) = P(W_0)\prod_{i=1}^{n-1} P(W_{i+1}|W_i)$$

A final assumption that's typically made in Markov models is that the transition model is **stationary**. In other words, for all values of i (all timesteps), $P(W_{i+1}|W_i)$ is identical. This allows us to represent a Markov model with only two tables: one for $P(W_0)$ and one for $P(W_{i+1}|W_i)$.

## The Mini Forward Algorithm

We now know how to compute the joint distribution across timesteps of a Markov model. However, this doesn't explicitly help us answer the question of the distribution of the weather on some given day t. Naturally, we can compute the joint then marginalize (sum out) over all other variables, but this is typically extremely inefficient, since if we have j variables each of which can take on d values, the size of the joint distribution is $O(d^j)$. Instead, we'll present a more efficient technique called the mini-forward algorithm. Here's how it works. By properties of marginalization, we know that

$$P(W_i + 1) = \sum_{W_i} P(W_i, W_{i+1})$$

By the chain rule we can re-express this as follows:

$$P(W_i + 1) = \sum_{w_i} P(W_{i+1}|(W_i)P(W_i)$$

This equation should make some intuitive sense — to compute the distribution of the weather at timestep i+1, we look at the probability distribution at timestep i given by $P(W_i)$ and "advance" this model a timestep with our transition model $P(W_{i+1}|(W_i)$. With this equation, we can iteratively compute the distribution of the weather at any timestep of our choice by starting with our initial distribution $P(W_0)$ and using it to compute $P(W_1)$ , then in turn using $P(W_1)$ to compute $P(W_2)$ and so on. Let's walk through an example, using the following initial distribution and transition model:

| $W_0$ | $P(W_0)$ |
|---|---|
| *sun* | 0.8 |
| *rain* | 0.2 |

| $W_{i+1}$ | $W_i$ | $P(W_{i+1}\|W_i)$ |
|---|---|---|
| *sun* | *sun* | 0.6 |
| *rain* | *sun* | 0.4 |
| *sun* | *rain* | 0.1 |
| *rain* | *rain* | 0.9 |

Using the mini-forward algorithm we can compute $P(W_1)$ as follows:

$$P(W_1 = sun) = \sum_{w_0} P(W_1 = sun|w_0)P(w_0)$$

$$P(W_1 = sun) = P(W_1 = sun|W_0 = sun)P(W_0 = sun) + P(W_1 = sun|W_0 = rain)P(W_0 = rain)$$

$$= 0.60.8 + 0.10.2 = 0.5$$

$$P(W_1 = rain) = P(W_1 = rain|w_0)P(w_0)$$

$$P(W_1 = rain) = P(W_1 = rain|W_0 = sun)P(W_0 = sun) + P(W_1 = rain|W_0 = rain)P(W_0 = rain)$$

$$= 0.4 0.8 + 0.9 0.2 = 0.5$$

Notably, the probability that it will be sunny has decreased from 80% at time t = 0 to only 50% at time t = 1. This is a direct result of our transition model, which favors transitioning to rainy days over sunny days. This gives rise to a natural follow-up question: does the probability of being in a state at a given timestep ever converge? We'll address the answer to this problem in the following section.

### Stationary Distribution

To solve the problem stated above, we must compute the stationary distribution of the weather. As the name suggests, the stationary distribution is one that remains the same after the passage of time, i.e.

$$P(W_{t+1}) = P(W_t)$$

We can compute these converged probabilities of being in a given state by combining the above equivalence with the same equation used by the mini-forward algorithm:

$$P(W_{t+1}) = P(W_t) = \sum_{w_t} P(W_{t+1}|w_t)P(w_t)$$

For our weather example, this gives us the following two equations:

$$P(W_t = sun) = P(W_{t+1} = sun|W_t = sun)P(W_t = sun)+P(W_{t+1} = sun|W_t = rain)P(W_t = rain)$$

$$= 0.6P(W_t = sun) + 0.1P(W_t = rain)$$

$$P(W_t = rain) = P(W_{t+1} = rain|W_t = sun)P(W_t = sun)+P(W_{t+1} = rain|W_t = rain)P(W_t = rain)$$

$$= 0.4P(W_t = sun) + 0.9P(W_t = rain)$$

We now have exactly what we need to solve for the stationary distribution, a system of 2 equations in 2 unknowns! We can get a third equation by using the fact that $P(W_t)$ is a probability distribution and so must sum to 1:

$$P(W_t = sun) = 0.6P(W_t = sun) + 0.1P(W_t = rain)$$

$$P(W_t = rain) = 0.4P(W_t = sun) + 0.9P(W_t = rain)$$

$$1 = P(W_t = sun) + P(W_t = rain)$$

Solving this system of equations yields $P(W_t = sun) = 0.2$ and $P(W_t = rain) = 0.8$. Hence the table for our stationary distribution, which we'll henceforth denote as $P(W_\infty)$, is the following:

$$P(W_{\infty+1} = sun) = P(W_{\infty+1} = sun|W_\infty = sun)P(W_\infty = sun)+P(W_{\infty+1} = sun|W_\infty = rain)P(W_\infty = rain)$$

$$= 0.6 \times 0.2 + 0.1 \times 0.8 = 0.2$$

$$P(W_{\infty+1} = rain) = P(W_\infty+1 = rain|W_\infty = sun)P(W_\infty = sun)+P(W_{\infty+1} = rain|W_\infty = rain)P(W_\infty = rain)$$

$$= 0.4 \times 0.2 + 0.9 \times 0.8 = 0.8$$

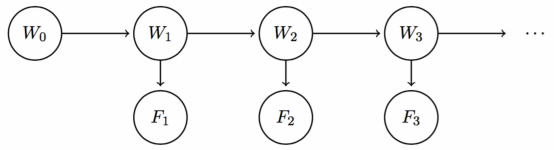
As expected, $P(W_\infty + 1) = P(W_\infty)$. In general, if $W_t$ had a domain of size k, the equivalence

$$P(W_t) = \sum_{w_t} P(W_{t+1}|w_t)P(w_t)$$

yields a system of k equations, which we can use to solve for the stationary distribution.

### Hidden Markov Models

With Markov models, we saw how we could incorporate change over time through a chain of random variables. For example, if we want to know the weather on day 10 with our standard Markov model from above, we can begin with the initial distribution $P(W_0)$ and use the mini-forward algorithm with our transition model to compute $P(W_10)$. However, between time t = 0 and time t = 10, we may collect new meteorological evidence that might affect our belief of the probability distribution over the weather at any given timestep. In simpler terms, if the weather forecasts an 80% chance of rain on day 10, but there are clear skies on the night of day 9, that 80% probability might drop drastically. This is exactly what the Hidden Markov Model helps us with - it allows us to observe some evidence at each timestep, which can potentially affect the belief distribution at each of the states. The Hidden Markov Model for our weather model can be described using a Bayes' net structure that looks like the following:




Unlike vanilla Markov models, we now have two different types of nodes. To make this distinction, we'll call each $W_i$ a state variable and each weather forecast $F_i$ an evidence variable. Since $W_i$ encodes our belief of the probability distribution for the weather on day i, it should be a natural result that the weather forecast for day i is conditionally dependent upon this belief. The model implies similar conditional independence relationships as standard Markov models, with an additional set of relationships for the evidence variables:

$$\begin{aligned}
F_1 &\perp\!\!\!\perp W_0|W_1 \\
\forall i &= 2,\ldots,n; \qquad W_i \perp\!\!\!\perp \{W_0,\ldots,W_{i-2},F_1,\ldots,F_{i-1}\}|W_{i-1} \\
\forall i &= 2,\ldots,n; \qquad F_i \perp\!\!\!\perp \{W_0,\ldots,W_{i-1},F_1,\ldots,F_{i-1}\}|W_i
\end{aligned}$$

Just like Markov models, Hidden Markov Models make the assumption that the transition model $P(W_i + 1|W_i)$ is stationary. Hidden Markov Models make the additional simplifying assumption that the sensor model $P(F_i|W_i)$ is stationary as well. Hence any Hidden Markov Model

can be represented compactly with just three probability tables: the initial distribution, the transition model, and the sensor model.
As a final point on notation, we'll define the belief distribution at time i with all evidence F1,...,Fi observed up to date:

$$B(W_i) = P(W_i|f_1, ..., f_i)$$

Similarly, we'll define $B(W_i)$ as the belief distribution at time i with evidence $f_1, ..., f_i1$ observed:

$$B(W_i) = P(W_i|f_1, ..., f_i1)$$

Defining $e_i$ as evidence observed at timestep i, you might sometimes see the aggregated evidence from timesteps $1 \leq i \leq t$ reexpressed in the following form:

$$e_{1:t} = e_1, ..., e_t$$

Under this notation, $P(W_i|f_1, ..., f_{i1})$ can be written as $P(W_i|f_1 : (i1))$. This notation will become relevant in the upcoming sections, where we'll discuss time elapse updates that iteratively incorporate new evidence into our weather model.