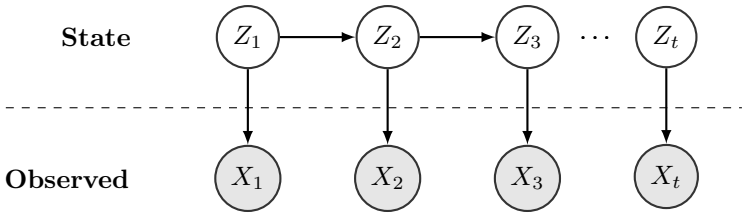
1 HMM Overview

1.1 Structure



Above we have a Hidden Markov Model (HMM). It can be thought of in a similar context to how we discussed Bayes Nets. We have an acyclic graph, every node indicates a random variable. Like before, a node is independent of its non-descendants give its parents.

1.2 State (Hidden or Latent Nodes)

State can be thought us in a similar way to state space from search. A state is a representation of the problem at that instance in time. Based on the Markov Assumption, we know that every state is conditionally independent of all other previous states when given the state immediately before it. For example:

$$P(Z_t|Z_{t-1}, Z_{t-2}...Z_1) = P(Z_t|Z_{t-1})$$

In the general case, the problem is broken into discrete time steps, and each step is represented by a new state/observation pair.

1.3 Observations

Observed are, as the name implies, your observation from that discrete time step. Because the state is hidden, you must derive your contextual understanding at each timestep from the observed variables. An algorithm for doing so will be discussed in more detail later on.

Note: While state is represented as a random variable that takes on discrete values, observations can be any value such that a probability mass function can assigned to it (mandates you can derive a probability for every observation). This means observations can be discrete or continuous. *We will generally be looking at discrete cases in this class, this is just an interesting extension.*

1.4 Wrap Up

Hidden Markov Models are quite expressive despite their simplicity. Of course, they cannot be used for everything. There are, however, many problems which can be modeled using this framework in a natural way.

2 HMM Symbolic Representation

Just like how Bayes Nets required a CPT for every node, there is a necessity for predefined distributions with every HMM. Lets try and be a little more formal about what is required to represent a HMM.

2.1 Necessary Components

2.1.1 Transition Probability (Stationary Assumption)

I will denote $T(i, j)$ to be the probability that the next state will be j given the current state is i. Written symbolically:

$$T(i, j) = P(Z_{k+1} = j | Z_k = i)$$

These values are fixed, and can be thought of as an m x m matrix when the state can take on values $1, 2, 3, ...m$.

2.1.2 Emission Probabilities

I will denote $\varepsilon_i(x)$ to be the probability that we observe x when the state is i. Written symbolically:

$$\varepsilon_i(x) = P(X_k = x | Z_k = i)$$

This distribution is the same no matter what k is.

2.1.3 Initial Distribution

I will denote $\pi(i)$ to be the probability that the first state is i. Written symbolically:

$$\pi(i) = P(Z_1 = i)$$

This is again a distribution over possible states.

2.2 Tying It All Together

Once these things are defined, we can calculate $P(x_1, x_2, ...x_t, z_1, z_2, ..., z_t)$. For example, we know from conditional independencies in the HMM that:

$$P(x_1, ...x_t, z_1, ..., z_t) = P(z_1)P(x_1|z_1)\prod_{i=2}^{t} P(z_i|z_{i-1})P(x_i|z_i)$$

Using what we defined previously, we can rewrite this more succinctly by stating:

$$P(x_1, ...x_t, z_1, ..., z_t) = \pi(z_1)\varepsilon_{z_1}(x_1)\prod_{i=2}^{t} T(z_{i-1}, z_i)\varepsilon_{z_i}(x_i)$$

3 Inference

Inference in a HMM can be done using the forward-backward algorithm. The forward backward algorithm provides a framework for calculating:

$$P(Z_k|x_1, x_2, x_3, ...x_n)$$

Where k is the distribution of a state at some discrete point, and n are all the observations your received from the first n time steps. Note, $1 \leq k \leq n$.

The algorithm breaks such a computation into two calculations. The forward part:

$$P(Z_k, x_1, x_2, ..., x_k)$$

And the backward part:

$$P(x_{k+1}, x_{k+2}, ...x_n|Z_k)$$

In this helper hours, we will focus primarily on the forward part, and applications where $n == k$. In other words, you want to know the distribution over current state using all your observations up through that state.

3.1 Forward Part

3.1.1 Mathematical Intuition

The first thing we have to do to calculate Z_K is marginalize over predecessors, Z_{k-1}

$$P(Z_k, x_1, x_2, ..., x_k) = \sum_{z_{k-1}} P(Z_k, Z_{k-1} = z_{k-1}, x_{1:k})$$

If we write this out using conditional independence properties we know:

$$P(Z_k, x_1, x_2, ..., x_k) = \sum_{z_{k-1}} P(x_k|Z_k)P(Z_k|Z_{k-1} = z_{k-1})P(Z_{k-1} = z_{k-1}, x_{1:k-1})$$

If you are not convinced, try writing out the D-Separation relationships on the HMM from the first section to see why these conditional independencies exist.

We have now created a recurrence. Z_k can be expressed in terms of Z_{k-1}, which can be expressed in terms of Z_{k-2} and so on until we get to $P(Z_1|x_1)$ which we can derive from the initial distribution and emission probabilities.

3.1.2 Computational Significance

Rather than having to compute an entire distribution with m^k states (k steps, and every state can take on m values) and marginalize over the whole thing, we can instead work from the bottom up, calculating one level at a time. Instead of the runtime being $O(m^k)$, it is simplified to $O(km^2)$ (m^2 work at every level for k levels).

4 Example

4.1 Problem Statement

You just finished CS 270, returned home for the summer, and want to help a local physician improve his flu screening for next year. So you decide to using an HMM as your modeling tool. The doctor observes whether or not a patient has a fever every time they come in for a checkup. He wants to be able to calculate, from all previous and current observations, the likelihood that the patient currently has a flu.

4.2 Distributions

4.2.1 State

The state will be either s or $\neg s$ to indicate sick or not sick respectively.

4.2.2 Observations

The observations will be f or $\neg f$ to indicate fever or no fever respectively.

4.2.3 Transition

The probability they are sick now given they were sick before is 0.75, and the probability they are sick now given they weren't sick before is 0.15: $P(s_k|s_{k-1}) = 0.75$ and $P(s_k|\neg s_{k-1}) = 0.15$

4.2.4 Emission

The probability they have a fever if they are sick is 0.8, and the probability they have a fever if they are not sick is 0.05: $P(f|s) = 0.8$ and $P(f|\neg s) = 0.05$

4.2.5 Initial Distribution

The patient is sick at first checkup, on average, one in ten times: $P(s_1) = 0.1$.

4.3 Computation

The forward inference algorithm gives us a means of solving a problem like this, so let's apply it to see what we get.

4.3.1 Example Data

For ease of computation, let's assume the person has only come in for two checkups, meaning we have observed two events, o_1 and o_2. We now want to know the probability that they are sick $P(S_2, o_1, o2)$. In this case lets say they did not have a fever the first visit, but now they do: $o_1 = \neg f$, $o_2 = f$.

4.3.2 Solving

Lets start by calculating the base case to our recurrence:

$$P(S_1|o_1)$$

This is proportional to $P(o_1|S_1)P(S_1)$. In this case S_1 can take on values s or $\neg s$. From initial distribution and emission, we know that:

$$P(s_1)P(\neg f_1|s_1) = 0.1 * 0.2$$
$$P(\neg s_1)P(\neg f_1|\neg s_1) = 0.9 * 0.95$$

Remember, we must renormalize: After doing so, we arrive at the following distribution:

$$P(s_1|o_1) = 0.023$$

Now, knowing $P(S_1|o_1)$ we can calculate $P(S_2|o_1, o_2)$ using the rule we derived before. We know $P(S_2|o_1, o_2)$ is proportional to:

$$P(o_2|S_2)P(S_2|S_1)P(S_1|o_1)$$

S_2 and S_1 can take on values $s, \neg s$. Lets calculate each of the four possibilities:

$$P(f|S_2 = s)P(S_2 = s|S_1 = s)P(S_1 = s|o_1) = 0.8 * 0.75 * 0.023$$
$$P(f|S_2 = \neg s)P(S_2 = \neg s|S_1 = s)P(S_1 = s|o_1) = 0.05 * 0.25 * 0.023$$
$$P(f|S_2 = s)P(S_2 = s|S_1 = \neg s)P(S_1 = \neg s|o_1) = 0.8 * 0.15 * 0.977$$
$$P(f|S_2 = \neg s)P(S_2 = \neg s|S_1 = \neg s)P(S_1 = \neg s|o_1) = 0.05 * 0.85 * 0.977$$

After renormalizing, we get:

$$P(s_2|o_1, o_2) = 0.758$$

Which makes sense considering the patient has a fever and this is a strong indicator of sickness.

4.4 Expanding

4.5 Next Observation

How would you calculate S_3 given a third observations (so now you have access to o_1, o_2, and o_3)? What would/wouldn't you have to calculate?

$S_2|o_1, o_2$ doesn't change based on Markov Assumption, so we get $P(S_3|o_1, o_2, o_3)$ is proportional to:

$$P(o_3|S_3)P(S_3|S_2)P(S_2|o_1, o_2)$$

And we have already calculated $P(S_2|o_1, o_2)$

4.6 Increasing Efficiency

What if the doctor needed this computation for every patient, on every visit? Should we start from scratch every time?

Probably not. We could instead write out S_i to a file for every visit, then read it in to calculate S_{i+1}, and rewrite it out to a file. This would only require a constant amount of computation every visit independent of the number of times the patient had been seen.

4.7 Increasing Observations

What if we observed two factors rather than one. Assuming both are binary, we now have 4 possibilities for an observation rather than 2. How might this make our HMM more expressive?

Open ended.