CSCI567 Machine Learning (Spring 2018)

Michael Shindler

Lecture 19: March 28

Outline

1. Administration
2. Review of last lecture
3. Markov model
4. Hidden Markov Model

Outline

Quiz 2 coming up

- It is now end of week 11
- Friday April 6 is coming up. Quiz 2.
- Quiz 2:
  - Bring a pencil
  - Bring your USC ID.
  - Be sure to fill out the ID section on Scantron
  - Be sure to STOP when time called
    - Stop writing immediately.
    - Look up, not at your exam or desk.
  - Know your name, ID#, lecture room, enrolled discussion time

Outline

1. Administration
2. Review of last lecture
   - Generative vs. Discriminative Approaches
   - Kernel Density Estimation
3. Markov model
4. Hidden Markov Model

Generative

- Aim to model the joint distribution $p(x, y)$. For naive Bayes and Gaussian discriminant analysis, this is done by assuming the form

$$p(x, y) = p(y)p(x|y)$$

  and then model $p(x|y)$ and $p(y)$ separately.

- Parameters of the distribution are estimated by maximizing the likelihood

$$\max_{\theta} \sum_{n} \log p(x_n, y_n; \theta)$$

- To classify, compute the posterior probability and identify the latest one

$$\arg\max_y p(y|x) = \arg\max_y p(x, y)$$

Discriminative (such as logistic regression)

- Aim to model the conditional distribution

$$p(y|x)$$

- Parameters of the distribution are estimated by maximizing the conditional likelihood

$$\max_{\theta} \sum_{n} \log p(y_n|x_n; \theta)$$

- To classify, compute the model out

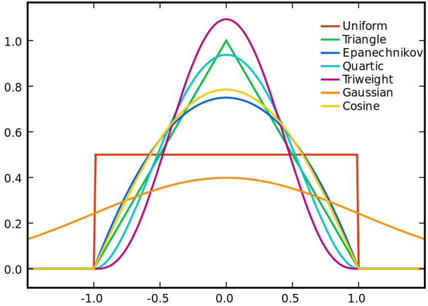
$$\arg\max_y p(y|x) = \arg\max_y p(y|x)$$

Parzen window method

Given a set of data points $x_1, x_2, \cdots, x_n$, the corresponding density estimator is

$$\hat{p}(x) = \sum_{n=1}^{N} \frac{1}{h} K\left(\frac{x - x_n}{h}\right) \frac{1}{N} = \sum_{n=1}^{N} K_h(x - x_n) \frac{1}{N}$$

where $K(\cdot)$ is a kernel function.



Outline

Markov chain

Definition

Given a sequentially ordered random variables $X_1, X_2, \cdots, X_t, \cdots, X_T$, called *states*,

- **Transition probability** for describing how the state at time $t-1$ changes to the state at time t,

$$P(X_t = \text{value}' | X_{t-1} = \text{value})$$

- **Initial probability** for describing the initial state at time $t = 1$.

$$P(X_1 = \text{value})$$

value represents possible values $\{X_t\}$ can take. Note that we will assume that all the random variables (at different times) can take value from the same set and assume that the transition probability does not change with respect to time t, i.e., a stationary Markov chain.

Special case and our focus for the rest of the course

When X_t are discrete, taking values from $\{1, 2, 3, \cdots, N\}$

- Transition probability becomes a table/matrix $\boldsymbol{A}$ whose elements are

$$a_{ij} = P(X_t = j | X_{t-1} = i)$$

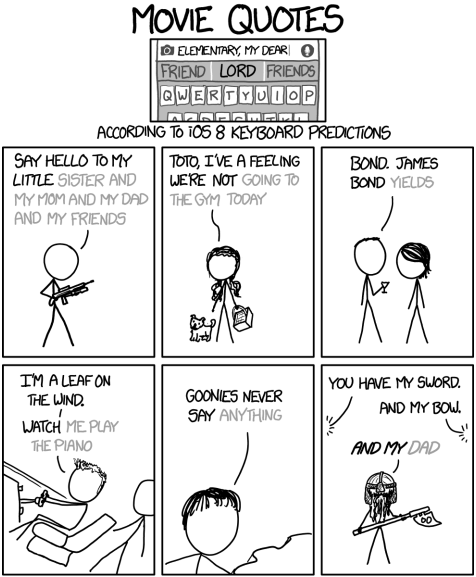
- Initial probability becomes a vector $\boldsymbol{\pi}$ whose elements are

$$\pi_i = P(X_1 = i)$$

where i or j index over from 1 to N. We have the following constraints

$$\sum_j a_{ij} = 1 \quad \sum_i \pi_i = 1$$

Additionally, all those numbers should be non-negative.


MOVIE QUOTES
ELEMENTARY, MY DEAR
FRIEND
LORD
FRIENDS
Q W E R T Y U I O P
ACCORDING TO iOS 8 KEYBOARD PREDICTIONS
SAY HELLO TO MY LITTLE SISTER AND MY MOM AND MY DAD AND MY FRIENDS
TOTO, I'VE A FEELING WE'RE NOT GOING TO THE GYM TODAY
BOND. JAMES BOND YIELDS
I'M A LEAF ON THE WIND.
WATCH ME PLAY THE PIANO
GOONIES NEVER SAY ANYTHING
YOU HAVE MY SWORD.
AND MY BOW.
AND MY DAD

Examples

- Example 1 (Language model)

$$P(\text{next\_word} = \text{cream}|\text{current\_word} = \text{ice})$$

is a gigantic matrix of $N \times N$ where N is the number of words in the dictionary. It can be used to inform us what likely the next word(s) is/are.

- Example 2 (Temperature)

$$P(\text{temperature at month } j|\text{temperature at month}(j-1))$$

is a matrix of $N \times N$ where N is the number of possible temperature bucket: extremely cold, very cold, cold, cool, warm, hot, very hot, extremely hot.

More language examples

Collected from https://www.reddit.com/user/nfl_ss, March 25 2018.

- Everything else is saying the Rams stopped doing that.
- Please point to and what is it is new years eve. Browns have a original team logo with a 9-7 team drafting outside the hashes are no other team and they probably wouldn't need to be happy.

More language examples

Collected from https://www.reddit.com/user/nfl_ss, March 25 2018.

- Everything else is saying the Rams stopped doing that.
- Please point to and what is it is new years eve. Browns have a original team logo with a 9-7 team drafting outside the hashes are no other team and they probably wouldn't need to be happy.
- His moms house... definitely looks like something the Chargers owned the Colts lol. To be fair, he wasn't playing and started acting out to ~17.9... slightly worse than 28-3.
- How many teams have a #2 corner.

More language examples

Collected from https://www.reddit.com/user/nfl_ss, March 25 2018.

- Everything else is saying the Rams stopped doing that.
- Please point to and what is it is new years eve. Browns have a original team logo with a 9-7 team drafting outside the hashes are no other team and they probably wouldn't need to be happy.
- His moms house... definitely looks like something the Chargers owned the Colts lol. To be fair, he wasn't playing and started acting out to ~17.9... slightly worse than 28-3.
- How many teams have a #2 corner.
- Didn't know Mitch Hedberg was a 70 yard Hail Mary, then he was trying to protect the field last season, but when Michael was with Dallas he came in with them, if I should have clarified that I don't think it's Hunt.

High-order Markov

We have assumed the following Markov property

$$P(X_t|X_1, X_2, \cdots, X_{t-1}) = P(X_t|X_{t-1})$$

that is why we are only concerning with ourselves the *immediate* history.

We can extend to use more histories, thus high-order Markov

$$P(X_t|X_1, X_2, \cdots, X_{t-1}) = P(X_t|X_{t-1}, X_{t-2}, \cdots, X_{t-H})$$

For instance, the language model previously is an order-one HMM. Obviously, languages have long-range dependency so the past history (not just a single word) matters.

Parameter estimation for Markov models

Given a training dataset $\mathcal{D}$, how do we estimate the parameters A and π?

$$\mathcal{D} = \{\boldsymbol{x}^1 = (x_1^1, x_2^1, \cdots, x_T^1), \boldsymbol{x}^2 = (x_1^2, x_2^2, \cdots, x_T^2), \cdots \\ \boldsymbol{x}^M = (x_1^M, x_2^M, \cdots, x_T^M)\}$$

Note that we have assumed all M *observed* sequences have equal length T — extending to unequal lengths is left as an exercise

Maximum likelihood estimation

$$\boldsymbol{A}^*, \boldsymbol{\pi}^* = \arg\max \log P(\mathcal{D}) = \arg\max \sum_m \log P(\boldsymbol{x}^m)$$

How to compute the probability of a sequence?

We need to compute

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_T = x_T)$$

We use the Markov property to factorize

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_T = x_T) = \tag{1}$$

$$P(X_1 = x_1) \prod_{t=2}^{T} P(X_t = x_t | X_{t-1} = x_{t-1}) \tag{2}$$

How to derive this? Details as an exercise but you should leverage the property in the following way:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_3|X_1, X_2)P(X_1, X_2) \\ &= P(X_3|X_2)P(X_1, X_2) = P(X_3|X_2)P(X_2|X_1)P(X_1) \end{aligned}$$

Maximum likelihood estimation

$$
\begin{aligned}
\sum_m \log P(\boldsymbol{x}^m) &= \sum_m \log P(x_1^m) + \sum_m \sum_t \log P(x_t^m | x_{t-1}^m) \\
&= \sum_m \log \pi_{x_1^m} + \sum_m \sum_t \log a_{x_{t-1}^m \, x_t^m}
\end{aligned}
$$

Maximizing this, we will get (derivation is left as an exercise)

$$
\pi_i = \frac{\#\text{of sequences starting with } i}{\#\text{of sequences}}
$$

and

$$
a_{ij} = \frac{\#\text{of transitions starting with } i \text{ but ending with } j}{\#\text{of transitions starting with } i}
$$

Example

Suppose we have two possible states $X_t \in \{0, 1\}$, and we have observed the following 3 sequences

$$1\ 0\ 0\ 1$$
$$0\ 1\ 1\ 1$$
$$1\ 1\ 1\ 1$$

Thus

$$\pi_0 = \frac{1}{3}, \quad \pi_1 = \frac{2}{3}$$

and

$$a_{00} = \frac{1}{3}, \quad a_{01} = \frac{2}{3}$$
$$a_{10} = \frac{2}{6}, \quad a_{11} = \frac{4}{6}$$

Example

- If it's rainy one day, 5050 it rains next day too
- If it's sunny, 80% chance sun next day.
- Let rainy be state zero; transition matrix:

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{bmatrix}$$

- If today is rainy, what is probability that after 10 days, we will have a rainy day?

Example continued

We can represent this:

Example continued

What about "in ten days will it be rainy?"

Outline

Motivation example

Underlying process is Markov chain

Say, the temperature fluctuation in each month:
cold, cold, hot, hot, cold, hot, ...

But we observe only indirectly, through a related quantity

Say, we can measure how many scoops of ice creams that have been consumed
1, 3, 3, 2, 1, 1, ...

Question

How do we infer the trace of the temperatures from how much we have eaten the ice creams?

Hidden Markov Models

Brief History:

- The foundations that we know today were laid down in three papers by LE Baum and colleagues in 1966, 1970 and 1972.
- A bit earlier than this (1960-61), the foundations of what we now call (linear) state-space models were being developed by R Kalman, RS Bucy and others.
- In the mid-1970s, it was realized that HMMs are discrete non-linear state-space models; equivalently, linear state space models are linear continuous HMMs. Soon afterwards, hybrid forms appeared.
- Applications: first in finance, later in text modeling and speech recognition, in 80s genetics and molecular biology, and now everywhere.

Formal definition of Hidden Markov Models (HMMs)

What are the variables?

- Underlying Markov chain, i.e., a set of random variables
    1. $Z_1, Z_2, \cdots, Z_t, \cdots Z_T$
    2. $Z_t \in \{s_1, s_2, s_3, \cdots, s_S\}$, a discrete set of S values
- Observed variable, i.e, a set of random variables
    1. $X_1, X_2, \cdots, X_t, \cdots X_T$
    2. $X_t \in \{o_1, o_2, o_3, \cdots, o_N\}$, a discrete set of N values

Key difference: Zs are never observed. However, their values can be inferred from the observed values Xs.

Explanation of notations

To avoid confusion, we will use

- Capitalized letter such as Z_t represents a random variable
- Lower-case letter such as z_t represents the value Z_t has taken
- Lower-case letter such as s_1, s_2 represent the value Z_t could take, i.e., its domain.

In other words, we can have

- $P(Z_t)$ mean probability of the random variable (of taking some value)
- $P(Z_t = z_t)$ or $P(z_t)$ mean probability taking value z_t
- $P(Z_t = s_1)$ mean probability take value s_1

Parameters for specifying the probabilistic structures

- For the Markov chain
  1. Transition probability: $a_{ij} = P(Z_t = s_j | Z_{t-1} = s_i)$ for $1 \leq i, j \leq S$
  2. Initial probability: $\pi_i = P(Z_1 = s_i)$ for $1 \leq i \leq S$.
- For observation model

$$b_i(k) = P(X_t = o_k | Z_t = s_i)$$

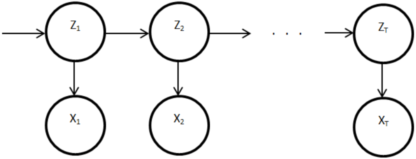
for $1 \leq k \leq N$ and $1 \leq i \leq S$.

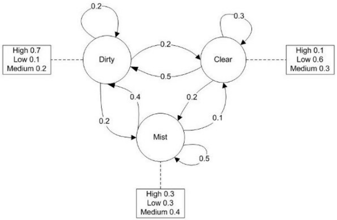
Collectively, $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$

The ice cream example

- States: $s_1 =$ cold and $s_2 =$ hot with $S = 2$
- Observed variables: $o_1 = 1$, $o_2 = 2$, and $o_3 = 3$ with $N = 3$

Graphical Model Representation of HMM





HMM defines a joint probability

$$\begin{aligned}
&P(X_1, X_2, \cdots, X_T, Z_1, Z_2, \cdots, Z_T) \\
&= P(Z_1, Z_2, \cdots, Z_T) P(X_1, X_2, \cdots, X_T | Z_1, Z_2, \cdots, Z_T)
\end{aligned}$$

- Markov assumption simplifies the first term

$$P(Z_1, Z_2, \cdots, Z_T) = P(Z_1) \prod_{t=2}^{T} P(Z_t | Z_{t-1})$$

- The *independence* assumption simplifies the second term

$$P(X_1, X_2, \cdots, X_T | Z_1, Z_2, \cdots, Z_T) = \prod_{t=1}^{T} P(X_t | Z_t)$$

Namely, each X_t is conditionally independent of anything else, if conditioned on Z_t.

Intuition of these assumptions

Ice cream example

- How hot every month is governed by a physical process whose current state depends on the previous state.
  For example, how far we are from the Sun
- How much ice cream you like to eat depends on the temperature of that day.
  - If I tell you what the temprature today is, then you know how much ice cream you will eat
  - If I tell you only what the temperature yesterday is, then how much ice cream you eat is correlated with that
  - If I tell you what you have eaten yesterday, then how much ice cream you eat yesterday correlates with how much you eat today
  - If I tell you what you have eaten yesterday, or yesterday's temperature and today's temperature, then the information is a bit redundant since you have known today's temperature

In HMMs, we are often interested in the following problems

- Total probability of observing a whole sequence

$$P(x_1, x_2, \cdots, x_T)$$

- The most likely path of the Markov chain's states

$$(z_1^*, z_2^*, \cdots, z_T^*) = \arg\max P(z_1, z_2, \cdots, z_T | x_1, x_2, \cdots, x_T)$$

- The likelihood of a state at a given time

$$P(z_t | x_1, x_2, \cdots, x_T)$$

- The likelihood of two consecutive states at a given time

$$P(z_{t-1}, z_t | x_1, x_2, \cdots, x_T)$$

They are all related to how HMMs is to be used, as well as how to estimate parameters of HMMs from data.

How to compute $P(x_1, x_2, \cdots, x_T)$?

We need to marginalize all the hidden variables

$$P(x_1, x_2, \cdots, x_T) = \sum_{Z_1} \sum_{Z_2} \cdots \sum_{Z_T} P(x_1, x_2, \cdots, x_T, Z_1, Z_2, \cdots, Z_T)$$

and there are exponential number of sums. But the structure of HMMs will enable an efficient way of computing it.

We will start with an auxiliary quantity[1]

$$\alpha_t(j) = P(Z_t = s_j | x_{1:t})$$

where $x_{1:t}$ represents $x_1, x_2, \cdots, x_t$. This quantity is called "forward message". The intuition is, if we observe up to time t, what is the likelihood of the Markov chain in state s_j?

Note that, this quantity can be defined differently, resulting slightly different algorithms.

Forward message can be computed recursively

$$
\begin{aligned}
\alpha_t(j) &= \frac{P(Z_t = s_j, x_{1:t-1}, x_t)}{P(x_{1:t})} \\
&= \frac{P(x_t|Z_t = s_j, x_{1:t-1})P(Z_t = s_j, x_{1:t-1})}{P(x_{1:t})} \\
&\overset{\text{due to independence}}{=} \frac{P(x_t|Z_t = s_j)P(Z_t = s_j, x_{1:t-1})}{P(x_{1:t})} \\
&= \frac{P(x_t|Z_t = s_j)\sum_i P(Z_t = s_j, Z_{t-1} = s_i, x_{1:t-1})}{P(x_{1:t})} \\
&= \frac{P(x_t|Z_t = s_j)\sum_i P(Z_t = s_j|Z_{t-1} = s_i, x_{1:t-1})P(Z_{t-1} = s_i, x_{1:t-1})}{P(x_{1:t})} \\
&= \frac{P(x_t|Z_t = s_j)\sum_i P(Z_t = s_j|Z_{t-1} = s_i)P(Z_{t-1} = s_i, x_{1:t-1})}{P(x_{1:t})} \\
&= \frac{P(x_t|Z_t = s_j)\sum_i P(Z_t = s_j|Z_{t-1} = s_i)P(Z_{t-1} = s_i|x_{1:t-1})}{P(x_{1:t})/P(x_{1:t-1})}
\end{aligned}
$$

Recursion

$$\alpha_t(j) = \frac{P(x_t|Z_t = s_j)\sum_i a_{ij}\alpha_{t-1}(i)}{\text{something}_t}$$

Do we need to compute something_t? There is an easy way:

$$\text{something}_t = \sum_j P(x_t|Z_t = s_j)\sum_i a_{ij}\alpha_{t-1}(i)$$

because we need to make sure $\sum_j \alpha_t(j) = 1$ (because $\alpha_t(j)$ is a probability).

Base case

When $t = 1$

$$\alpha_1(j) = P(Z_1 = s_1|x_1) = \frac{P(x_1|Z_1 = s_j)P(Z_1 = s_j)}{P(x_1)} = \frac{\pi_j P(x_1|Z_1 = s_j)}{P(x_1)}$$

where $P(x_1)$ is

$$\sum_j \pi_j P(x_1|Z_1 = s_j)$$

So what is $P(x_{1:T})$?

$$P(x_{1:T}) = P(x_1)\frac{P(x_{1:2})}{P(x_1)}\frac{P(x_{1:3})}{P(x_{1:2})}\cdots\frac{P(x_{1:t})}{P(x_{1:t-1})}\cdots\frac{P(x_{1:T})}{P(x_{1:T-1})}$$

which is

$\text{something}_1 \times \text{something}_2 \times \text{something}_3 \cdots \times \text{something}_t \times \cdots \times \text{something}_T$

Note that this is the formula given in the textbook by Kevin Murphy.

Forward procedure

- Compute $\alpha_1(j)$ for all $1 \leq j \leq S$.
- Compute something$_1 = P(x_1)$.
- Use the recursion to compute $\alpha_t(j)$ and make sure you keep something$_t$ for $2 \leq t \leq T$
- Compute $P(x_{1:T})$ using all the accumulated something$_t$

Alternative method

This is somewhat simpler and more stable numerically

- Define $\alpha_t(j) = P(Z_t = s_j, x_{1:t})$ — note that this is not a conditional probability
- Base case: $\alpha_1(j) = P(x_1|Z_1 = s_j)P(Z_1 = s_j) = \pi_j P(x_1|Z_1 = s_j)$
- Use recursion

$$\alpha_t(j) = P(x_t|Z_t = s_j) \sum_i a_{ij}\alpha_{t-1}(i)$$

- Compute

$$P(x_{1:T}) = \sum_j \alpha_T(j)$$

Showing the correctness of this procedure is left as an exercise. This procedure has one advantage: we do not have to keep something$_t$ along the way.

Backward algorithm

We can define backward messages

$$\beta_t(j) = P(x_{t+1:T} | Z_t = s_j)$$

The interpretation is: if we are told that the Markov chain at time t is in the state s_j, then what are the likelihood of observing *future* observations from $t+1$ to T?

Recursion

$$\beta_{t-1}(i) = \sum_j \beta_t(j) a_{ij} p(x_t | z_t = s_j)$$

with the base case of $\beta_T(j) = 1$ for any j.

Derivation

$$
\begin{aligned}
\beta_t(j) &= P(x_{t+1:T}|Z_t = s_j) \\
\beta_{t-1}(i) &= P(x_{t:T}|Z_{t-1} = s_i) \\
&= P(x_t, x_{t+1:T}|Z_{t-1} = S_i) \\
&= \sum_j P(x_t, x_{t+1:T}, Z_t = s_j|Z_{t-1} = s_i) \\
&= \sum_j P(Z_t = s_j|Z_{t-1} = s_i) \cdot P(x_t|Z_t = s_j \wedge Z_{t-1} = s_i) \\
&\qquad\qquad\qquad\qquad\qquad \cdot P(X_{t+1:T}|x_t, Z_t = s_j, Z_{t-1} = s_i)
\end{aligned}
$$

But what is $P(Z_t = s_j|Z_{t-1} = s_i)$?

Derivation

$$
\begin{aligned}
\beta_t(j) &= P(x_{t+1:T} | Z_t = s_j) \\
\beta_{t-1}(i) &= P(x_{t:T} | Z_{t-1} = s_i) \\
&= P(x_t, x_{t+1:T} | Z_{t-1} = S_i) \\
&= \sum_j P(x_t, x_{t+1:T}, Z_t = s_j | Z_{t-1} = s_i) \\
&= \sum_j P(Z_t = s_j | Z_{t-1} = s_i) \cdot P(x_t | Z_t = s_j \wedge Z_{t-1} = s_i) \\
&\qquad \cdot P(X_{t+1:T} | x_t, Z_t = s_j, Z_{t-1} = s_i)
\end{aligned}
$$

But what is $P(Z_t = s_j | Z_{t-1} = s_i)$? α_{ij}
What about $P(x_t | Z_t = s_j \wedge Z_{t-1} = s_i)$?

Derivation

$$
\begin{aligned}
\beta_t(j) &= P(x_{t+1:T}|Z_t = s_j) \\
\beta_{t-1}(i) &= P(x_{t:T}|Z_{t-1} = s_i) \\
&= P(x_t, x_{t+1:T}|Z_{t-1} = S_i) \\
&= \sum_j P(x_t, x_{t+1:T}, Z_t = s_j|Z_{t-1} = s_i) \\
&= \sum_j P(Z_t = s_j|Z_{t-1} = s_i) \cdot P(x_t|Z_t = s_j \wedge Z_{t-1} = s_i) \\
&\qquad \cdot P(X_{t+1:T}|x_t, Z_t = s_j, Z_{t-1} = s_i)
\end{aligned}
$$

But what is $P(Z_t = s_j|Z_{t-1} = s_i)$? α_{ij}
What about $P(x_t|Z_t = s_j \wedge Z_{t-1} = s_i)$? $P(x_t|Z_t = s_j)$

Derivation

$$
\begin{aligned}
\beta_t(j) &= P(x_{t+1:T}|Z_t = s_j) \\
\beta_{t-1}(i) &= P(x_{t:T}|Z_{t-1} = s_i) \\
&= P(x_t, x_{t+1:T}|Z_{t-1} = S_i) \\
&= \sum_j P(x_t, x_{t+1:T}, Z_t = s_j|Z_{t-1} = s_i) \\
&= \sum_j P(Z_t = s_j|Z_{t-1} = s_i) \cdot P(x_t|Z_t = s_j \wedge Z_{t-1} = s_i) \\
&\qquad\qquad \cdot P(X_{t+1:T}|x_t, Z_t = s_j, Z_{t-1} = s_i)
\end{aligned}
$$

But what is $P(Z_t = s_j|Z_{t-1} = s_i)$? α_{ij}
What about $P(x_t|Z_t = s_j \wedge Z_{t-1} = s_i)$? $P(x_t|Z_t = s_j)$
$P(X_{t+1:T}|x_t, Z_t = s_j, Z_{t-1} = s_i)$?

Derivation

$$
\begin{aligned}
\beta_t(j) &= P(x_{t+1:T}|Z_t = s_j) \\
\beta_{t-1}(i) &= P(x_{t:T}|Z_{t-1} = s_i) \\
&= P(x_t, x_{t+1:T}|Z_{t-1} = S_i) \\
&= \sum_j P(x_t, x_{t+1:T}, Z_t = s_j|Z_{t-1} = s_i) \\
&= \sum_j P(Z_t = s_j|Z_{t-1} = s_i) \cdot P(x_t|Z_t = s_j \wedge Z_{t-1} = s_i) \\
&\qquad \cdot P(X_{t+1:T}|x_t, Z_t = s_j, Z_{t-1} = s_i)
\end{aligned}
$$

But what is $P(Z_t = s_j|Z_{t-1} = s_i)$? α_{ij}
What about $P(x_t|Z_t = s_j \wedge Z_{t-1} = s_i)$? $P(x_t|Z_t = s_j)$
$P(X_{t+1:T}|x_t, Z_t = s_j, Z_{t-1} = s_i)$? $P(X_{t+1:T}|Z_t = s_j)$

Derivation

$$
\begin{aligned}
\beta_{t-1}(i) &= \sum_j \alpha_{ij} \times P(x_t|Z_t = s_j) \times P(X_{t+1:T}|Z_t = s_j) \\
&= \sum_j \alpha_{ij} \times P(x_t|Z_t = s_j) \times \beta_t(j)
\end{aligned}
$$

And this is exactly the recursive solution we were interested in!

Why we need both forward and backward?

How to compute $P(x_{1:T})$ from backward messages?

$$P(x_{1:T}) = \sum_i \beta_1(i)\pi_i P(x_1|Z_1 = s_i)$$

This is a good trick to check whether your forward/backward code is implemented correctly!

How to compute the likelihood of a state at a given time?

$$\gamma_t(j) = P(Z_t = s_j|x_{1:T}) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j'} \alpha_t(j')\beta_t(j')}$$

How to compute the likelihood of two consecutive states at a given time?

$$\xi_{t,t+1}(i,j) = p(Z_t = s_i, Z_{t+1} = s_j|x_{1:T}) = \frac{\alpha_t(i)p(x_{t+1}|z_{t+1} = s_j)\beta_{t+1}(j)a_{ij}}{\text{something}}$$

Making the most likely path

Yet another recursion!

Define the most likely path ending with j at time t

$$\delta_t(j) = \max_{z_1, z_2, \cdots z_{t-1}} P(Z_1 = z_1, Z_2 = z_2, \cdots Z_{t-1} = z_{t-1}, Z_t = s_j, x_{1:t})$$

It relates to

$$\delta_t(j) = \max_i \delta_{t-1}(i) a_{ij} P(x_t | Z_t = s_j)$$

The probability of the most likely path is then

$$\arg\max_j \delta_T(j)$$

Derivation of most likely path

$$P(Z_t = s_j | x_{1:T}) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j'} \alpha_t(j')\beta_t(j')}$$

$$\xi_{t,t+1}(i,j) = p(Z_t = s_i, Z_{t+1} = s_j | x_{1:T}) = \frac{\alpha_t(i) p(x_{t+1}|z_{t+1} = s_j)\beta_{t+1}(j) a_{ij}}{\text{something}}$$

$$\begin{aligned}
\delta_t(j) &= \max_{z_1, z_2, \cdots z_{t-1}} P(Z_1 = z_1, Z_2 = z_2, \cdots Z_{t-1} = z_{t-1}, Z_t = s_j, x_{1:t}) \\
&= \max_i \max_{z_1, z_2, \cdots z_{t-2}} P(Z_1 = z_1, Z_2 = z_2, \cdots Z_{t-1} = s_i, x_{1:t}) \\
&\quad \times P(Z_t = s_j | Z_1 = z_1, Z_2 = z_2, \cdots Z_{t-1} = s_i, x_{1:t})
\end{aligned}$$

But let's use :

$$\max_{z_1, z_2, \cdots z_{t-2}} P(Z_1 = z_1, Z_2 = z_2, \cdots Z_{t-1} = s_i, x_{1:t}) = \delta_{t-1}(i)$$

and what about

$$P(Z_t = s_j | Z_1 = z_1, Z_2 = z_2, \cdots Z_{t-1} = s_i, x_{1:t})$$

Derivation continued

What about

$$
\begin{aligned}
P(Z_t = s_j | Z_1 = z_1, Z_2 = z_2, \cdots Z_{t-1} = s_i, x_{1:t}) &= P(Z_t = s_j | Z_{t-1} = s_i, x_{1:T}) \\
&= \frac{P(Z_t = s_j, Z_{t-1} = s_i | x_{1:T})}{P(Z_{t-1} = s_i | x_{1:T})} \\
&= \frac{\alpha_{t-1}(i)\beta_{t-1}(i)\alpha_{ij} P(x_t | Z_t = s_j)}{\alpha_{t-1}(i)\beta_{t-1}(i)} \\
&= \alpha_{ij} P(x_t | Z_t = s_j)
\end{aligned}
$$

So

$$\delta_t(j) = \max_i \delta_{t-1}(i)\alpha_{ij} P(x_t | Z_t = s_j)$$