Outline

1 Markov model
   - Definition
   - Parameter estimation

2 Hidden Markov Model
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.
Markov model

Definition

**Special case and our focus for the rest of the course**

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

- Transition probability becomes a table/matrix $A$ whose elements are

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

- Initial probability becomes a vector $\pi$ 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.
High-order Markov

Previously, we have assumed

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

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

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

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

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

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

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**

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

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_T = x_T) = P(X_1 = x_1) \prod_{t=2}^{T} P(X_t = x_t | X_{t-1} = x_{t-1})$$

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

$$
\begin{align*}
P(X_1, X_2, X_3) &= P(X_3 | X_1, X_2) P(X_1, X_2) \\
&= P(X_3 | X_2) P(X_1, X_2) \\
&= P(X_3 | X_2) P(X_2 | X_1) P(X_1)
\end{align*}
$$
Maximum likelihood estimation

\[
\sum_m \log P(x^m) = \sum_m \log P(x^m_1) + \sum_m \sum_t \log P(x^m_t | x^m_{t-1})
\]

\[
= \sum_m \log \pi_{x^m_1} + \sum_m \sum_t \log a_{x^m_{t-1} x^m_t}
\]

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:

\[
\begin{align*}
&1 \ 0 \ 0 \ 1 \\
&0 \ 1 \ 1 \ 1 \\
&1 \ 1 \ 1 \ 1
\end{align*}
\]

Thus

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

and

\[
\begin{align*}
a_{00} &= \frac{1}{3}, \quad a_{01} = \frac{2}{3} \\
a_{10} &= \frac{2}{6}, \quad a_{11} = \frac{4}{6}
\end{align*}
\]
Outline

1. Markov model

2. Hidden Markov Model
   - Definition
   - Key inference problems in HMMs
   - Forward-backward algorithms
   - Viterbi algorithm
   - Parameter estimation
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
  \[ Z_1, Z_2, \ldots, Z_t, \ldots, Z_T \]
  \[ Z_t \in \{s_1, s_2, s_3, \ldots, s_S\} \], a discrete set of \( S \) values

- Observed variable, i.e, a set of random variables
  \[ X_1, X_2, \ldots, X_t, \ldots, X_T \]
  \[ X_t \in \{o_1, o_2, o_3, \ldots, o_N\} \], a discrete set of \( N \) values
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, $\lambda = (A, B, \pi)$
The ice cream example

- States: $s_1 = \text{cold}$ and $s_2 = \text{hot}$ 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

\[
P(X_1, X_2, \cdots, X_T, Z_1, Z_2, \cdots, Z_T) \\
= P(Z_1, Z_2, \cdots, Z_T)P(X_1, X_2, \cdots, X_T|Z_1, Z_2, \cdots, Z_T)
\]

- 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 \(Z_t\).
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
  \[ (z_1^*, z_2^*, \cdots, z_T^*) = \operatorname{arg\,max} P(z_1, z_2, \cdots, z_T | x_1, x_2, \cdots, x_T) \]

- 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

$$\alpha_t(j) = P(Z_t = s_j | x_1:t)$$

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.
Hidden Markov Model
Forward-backward algorithms

**Forward message can be computed recursively**

\[
\alpha_t(j) = \frac{P(Z_t = s_j, x_{1:t-1}, x_t)}{P(x_{1:t})} = \frac{P(x_t|Z_t = s_j, x_{1:t-1})P(Z_t = s_j, x_{1:t-1})}{P(x_{1:t})}
\]

due to independence

\[
= \frac{P(x_t|Z_t = s_j)P(Z_t = s_j, x_{1:t-1})}{P(x_{1:t})}
\]

\[
= \frac{P(x_t|Z_t = s_j) \sum_i P(Z_t = s_j, Z_{t-1} = s_i, x_{1:t-1})}{P(x_{1:t})}
\]

\[
= \frac{P(x_t|Z_t = s_j) \sum_i P(Z_t = s_j|Z_{t-1} = s_i, x_{1:t-1})P(Z_{t-1} = s_i, x_{1:t-1})}{P(x_{1:t})}
\]

\[
= \frac{P(x_t|Z_t = s_j) \sum_i P(Z_t = s_j|Z_{t-1} = s_i)P(Z_{t-1} = s_i, x_{1:t-1})}{P(x_{1:t})}
\]

\[
= \frac{P(x_t|Z_t = s_j) \sum_i P(Z_t = s_j|Z_{t-1} = s_i)P(Z_{t-1} = s_i|x_{1:t-1})}{P(x_{1:t})/P(x_{1:t-1})}
\]
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

something$_1 \times$ something$_2 \times$ something$_3 \cdots \times$ something$_t \times \cdots \times$ 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 was described in Prof. Sha’s lecture

- 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 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 on blackboard!*
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?

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

**Yet another recursion!**

Define the most likely path ending with $j$ at time $t$

\[
\delta_t(j) = \max_{z_1, z_2, \ldots, z_{t-1}} P(Z_1 = z_1, Z_2 = z_2, \ldots Z_{t-1} = z_{t-1}, Z_t = s_j | x_{1:T})
\]

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 on board!*
Central Techniques: EM Algorithm

Therefore maximizing $F(q, \theta)$ is equivalent to maximizing the expected complete log-likelihood $Q(\theta|\theta_n) = \sum_z q(z|x, \theta_n) \log p(x, z|\theta)$. We can reexpress the EM algorithm as follows:

**E-step:** compute $Q(\theta|\theta_n) = E_q(z|x, \theta_n) [\log p(x, z|\theta)]$

**M-step:** $\theta_{n+1} = \arg \max_{\theta} Q(\theta|\theta_n)$. 
Can you guess what the parameter estimation formulae looks like?

*Hint*. Check the parameter estimation formulae for Markov model. Imagine the “#” there should be replaced with soft-counts (when we derive EM for Gaussian mixture models), i.e., the probabilities of states as well as state pairs.
Computational complexity

- **Space Complexity:** $O(TS)$ since it uses an $T$ by $S$ matrix
- **Time Complexity:** $O(TS^2)$. There are $TS$ cells, and each cell is computed by looking at $S$ previous cells.