CSCI567 Machine Learning (Spring 2021)

Sirisha Rambhatla

University of Southern California

April 2, 2021

Outline

1. Logistics
2. (Hidden) Markov models I

Outline

1 Logistics

2 (Hidden) Markov models I

Logistics

- **April 7, 2021** is a Wellness day, there will be no class.

Outline

1 Logistics

2 (Hidden) Markov models I
- Markov chain
- Hidden Markov Model

Markov Models

Markov models are powerful probabilistic tools to analyze **sequential data**:

- text or speech data
- stock market data
- gene data
- $\cdots$

Definition

A **Markov chain** is a stochastic process with **Markov property**:

Definition

A **Markov chain** is a stochastic process with **Markov property**: a sequence of random variables $Z_1, Z_2, \cdots$ s.t.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t) \qquad \text{(Markov property)}$$

i.e. *the current state only depends on the most recent state* (notation $Z_{1:t}$ denotes the sequence $Z_1, \ldots, Z_t$).

Definition

A **Markov chain** is a stochastic process with **Markov property**: a sequence of random variables $Z_1, Z_2, \cdots$ s.t.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t) \qquad \text{(Markov property)}$$

i.e. *the current state only depends on the most recent state* (notation $Z_{1:t}$ denotes the sequence $Z_1, \ldots, Z_t$).

We only consider the following case:

- All Z_t's take value from the same discrete set $\{1, \ldots, S\}$

Definition

A **Markov chain** is a stochastic process with **Markov property**: a sequence of random variables $Z_1, Z_2, \cdots$ s.t.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t) \qquad \text{(Markov property)}$$

i.e. *the current state only depends on the most recent state* (notation $Z_{1:t}$ denotes the sequence $Z_1, \ldots, Z_t$).

We only consider the following case:

- All Z_t's take value from the same discrete set $\{1, \ldots, S\}$
- $P(Z_{t+1} = s' \mid Z_t = s) = a_{s,s'}$, known as transition probability

Definition

A **Markov chain** is a stochastic process with **Markov property**: a sequence of random variables $Z_1, Z_2, \cdots$ s.t.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t) \qquad \text{(Markov property)}$$

i.e. *the current state only depends on the most recent state* (notation $Z_{1:t}$ denotes the sequence $Z_1, \ldots, Z_t$).

We only consider the following case:

- All Z_t's take value from the same discrete set $\{1, \ldots, S\}$
- $P(Z_{t+1} = s' \mid Z_t = s) = a_{s,s'}$, known as transition probability
- $P(Z_1 = s) = \pi_s$

Definition

A **Markov chain** is a stochastic process with **Markov property**: a sequence of random variables $Z_1, Z_2, \cdots$ s.t.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t) \qquad \text{(Markov property)}$$

i.e. *the current state only depends on the most recent state* (notation $Z_{1:t}$ denotes the sequence $Z_1, \ldots, Z_t$).

We only consider the following case:

- All Z_t's take value from the same discrete set $\{1, \ldots, S\}$
- $P(Z_{t+1} = s' \mid Z_t = s) = a_{s,s'}$, known as transition probability
- $P(Z_1 = s) = \pi_s$
- $(\{\pi_s\}, \{a_{s,s'}\}) = (\boldsymbol{\pi}, \boldsymbol{A})$ are parameters of the model

Examples

- Example 1 (**Language model**)

  States $[S]$ represent a dictionary of words,

$$a_{\text{ice,cream}} = P(Z_{t+1} = \text{cream} \mid Z_t = \text{ice})$$

  is an example of the transition probability.

Examples

- Example 1 (**Language model**)

  States $[S]$ represent a dictionary of words,

  $$a_{\text{ice,cream}} = P(Z_{t+1} = \text{cream} \mid Z_t = \text{ice})$$

  is an example of the transition probability.

- Example 2 (**Weather**)

  States $[S]$ represent weather at each day

  $$a_{\text{sunny,rainy}} = P(Z_{t+1} = \text{rainy} \mid Z_t = \text{sunny})$$

High-order Markov chain

Is the Markov assumption reasonable?

High-order Markov chain

Is the Markov assumption reasonable? Not completely for the language model for example.

High-order Markov chain

Is the Markov assumption reasonable? Not completely for the language model for example.

Higher order Markov chains make it more reasonable, e.g.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t, Z_{t-1}) \qquad \textbf{(second-order Markov)}$$

i.e. the current word only depends on the last two words.

High-order Markov chain

Is the Markov assumption reasonable? Not completely for the language model for example.

Higher order Markov chains make it more reasonable, e.g.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t, Z_{t-1}) \qquad \textbf{(second-order Markov)}$$

i.e. the current word only depends on the last two words.

Learning higher order Markov chains is similar, but more expensive.

High-order Markov chain

Is the Markov assumption reasonable? Not completely for the language model for example.

Higher order Markov chains make it more reasonable, e.g.

$$P(Z_{t+1} \mid Z_{1:t}) = P(Z_{t+1} \mid Z_t, Z_{t-1}) \qquad \text{(}\textbf{second-order Markov}\text{)}$$

i.e. the current word only depends on the last two words.

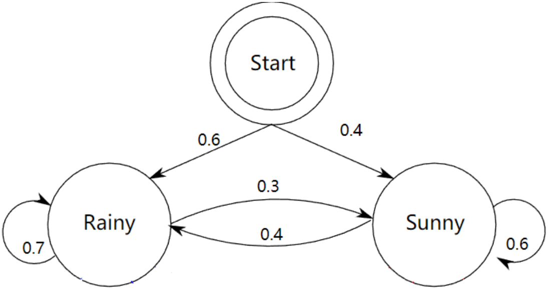
Learning higher order Markov chains is similar, but more expensive.

We only consider standard Markov chains.

Graph Representation

picture from Wikipedia

It is intuitive to represent a Markov model as a **graph**



Learning from examples

Now suppose we have observed N **sequences of examples**, say $\mathcal{D}$:

- $z_{1,1}, \ldots, z_{1,T}$
- $\cdots$
- $z_{n,1}, \ldots, z_{n,T}$
- $\cdots$
- $z_{N,1}, \ldots, z_{N,T}$

Learning from examples

Now suppose we have observed N **sequences of examples** , say $\mathcal{D}$:

- $z_{1,1}, \ldots, z_{1,T}$
- $\cdots$
- $z_{n,1}, \ldots, z_{n,T}$
- $\cdots$
- $z_{N,1}, \ldots, z_{N,T}$

where

- for simplicity we assume each sequence has the same length T

Learning from examples

Now suppose we have observed N **sequences of examples** , say $\mathcal{D}$:

- $z_{1,1}, \ldots, z_{1,T}$
- $\cdots$
- $z_{n,1}, \ldots, z_{n,T}$
- $\cdots$
- $z_{N,1}, \ldots, z_{N,T}$

where

- for simplicity we assume each sequence has the same length T
- lower case $z_{n,t}$ represents the value of the random variable $Z_{n,t}$

Learning from examples

Now suppose we have observed N **sequences of examples**, say $\mathcal{D}$:

- $z_{1,1}, \ldots, z_{1,T}$
- $\cdots$
- $z_{n,1}, \ldots, z_{n,T}$
- $\cdots$
- $z_{N,1}, \ldots, z_{N,T}$

where

- for simplicity we assume each sequence has the same length T
- lower case $z_{n,t}$ represents the value of the random variable $Z_{n,t}$

From these observations how do we *learn the model parameters* $\boldsymbol{\theta} := (\boldsymbol{\pi}, \boldsymbol{A})$?

Finding the MLE

Same story, find the **MLE**.

Finding the MLE

Same story, find the **MLE**. The log-likelihood of a sequence $z_1, \ldots, z_T$ is

$$\ln p(Z_{1:T} = z_{1:T}; \boldsymbol{\theta})$$

Finding the MLE

Same story, find the **MLE**. The log-likelihood of a sequence $z_1, \ldots, z_T$ is

$$\begin{aligned}
&\ln p(Z_{1:T} = z_{1:T}; \boldsymbol{\theta}) \\
&= \sum_{t=1}^{T} \ln p(Z_t = z_t \mid Z_{1:t-1} = z_{1:t-1}; \boldsymbol{\theta}) && \text{(always true)}
\end{aligned}$$

Finding the MLE

Same story, find the **MLE**. The log-likelihood of a sequence $z_1, \ldots, z_T$ is

$$\begin{aligned}
&\ln p(Z_{1:T} = z_{1:T}; \boldsymbol{\theta}) \\
&= \sum_{t=1}^{T} \ln p(Z_t = z_t \mid Z_{1:t-1} = z_{1:t-1}; \boldsymbol{\theta}) && \text{(always true)} \\
&= \sum_{t=1}^{T} \ln p(Z_t = z_t \mid Z_{t-1} = z_{t-1}; \boldsymbol{\theta}) && \text{(Markov property)}
\end{aligned}$$

Finding the MLE

Same story, find the **MLE**. The log-likelihood of a sequence $z_1, \ldots, z_T$ is

$$\begin{aligned}
&\ln p(Z_{1:T} = z_{1:T}; \boldsymbol{\theta}) \\
&= \sum_{t=1}^{T} \ln p(Z_t = z_t \mid Z_{1:t-1} = z_{1:t-1}; \boldsymbol{\theta}) && \text{(always true)} \\
&= \sum_{t=1}^{T} \ln p(Z_t = z_t \mid Z_{t-1} = z_{t-1}; \boldsymbol{\theta}) && \text{(Markov property)} \\
&= \ln \pi_{z_1} + \sum_{t=2}^{T} \ln a_{z_{t-1}, z_t}
\end{aligned}$$

Finding the MLE

Same story, find the **MLE**. The log-likelihood of a sequence $z_1, \ldots, z_T$ is

$$\ln p(Z_{1:T} = z_{1:T}; \boldsymbol{\theta})$$

$$= \sum_{t=1}^{T} \ln p(Z_t = z_t \mid Z_{1:t-1} = z_{1:t-1}; \boldsymbol{\theta}) \qquad \text{(always true)}$$

$$= \sum_{t=1}^{T} \ln p(Z_t = z_t \mid Z_{t-1} = z_{t-1}; \boldsymbol{\theta}) \qquad \text{(Markov property)}$$

$$= \ln \pi_{z_1} + \sum_{t=2}^{T} \ln a_{z_{t-1}, z_t}$$

$$= \sum_{s} \mathbb{I}[z_1 = s] \ln \pi_s + \sum_{s,s'} \left(\sum_{t=2}^{T} \mathbb{I}[z_{t-1} = s, z_t = s'] \right) \ln a_{s,s'}$$

Finding the MLE

So the MLE is

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \ln p(\mathcal{D};\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln p(Z_{n,1:T} = z_{n,1:T};\boldsymbol{\theta})$$

$$= \operatorname*{argmax}_{\boldsymbol{\pi},\boldsymbol{A}} \sum_{s} (\textbf{\#initial states with value } s) \ln \pi_s$$

$$+ \sum_{s,s'} (\textbf{\#transitions from } s \textbf{ to } s') \ln a_{s,s'}$$

subject to

$$\sum_{s'} a_{s,s'} = 1 \;\text{ and }\; a_{s,s'} \geq 0 \;\; \forall s \in [S]$$

Finding the MLE

So the MLE is

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \ln p(\mathcal{D};\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln p(Z_{n,1:T} = z_{n,1:T};\boldsymbol{\theta})$$

$$= \operatorname*{argmax}_{\boldsymbol{\pi},\boldsymbol{A}} \sum_{s} (\textbf{\#initial states with value } s) \ln \pi_s$$

$$+ \sum_{s,s'} (\textbf{\#transitions from } s \textbf{ to } s') \ln a_{s,s'}$$

subject to

$$\sum_{s'} a_{s,s'} = 1 \text{ and } a_{s,s'} \geq 0 \quad \forall s \in [S]$$

We have seen this many times. The solution is:

$$\pi_s \propto \#\text{initial states with value } s$$

$$a_{s,s'} \propto \#\text{transitions from } s \text{ to } s'$$

See also: [MLaPP 17.2.2] and `http://cs229.stanford.edu/section/cs229-hmm.pdf`

Example

Suppose we observed the following 2 sequences of length 5

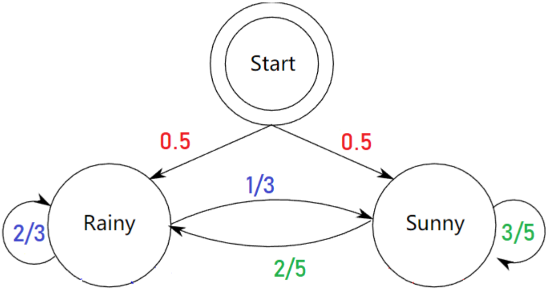
- sunny, sunny, rainy, rainy, rainy
- rainy, sunny, sunny, sunny, rainy

Example

Suppose we observed the following 2 sequences of length 5

- sunny, sunny, rainy, rainy, rainy
- rainy, sunny, sunny, sunny, rainy

MLE is the following model



Markov Model with outcomes

Now suppose each state Z_t also "emits" some **outcome** $X_t \in [O]$ based on the following model

$$P(X_t = o \mid Z_t = s) = b_{s,o} \qquad \text{(\textbf{emission probability})}$$

independent of anything else.

Markov Model with outcomes

Now suppose each state Z_t also "emits" some **outcome** $X_t \in [O]$ based on the following model

$$P(X_t = o \mid Z_t = s) = b_{s,o} \qquad \textbf{(emission probability)}$$

independent of anything else.

For example, in the language model, X_t is the speech signal for the underlying word Z_t (very useful for speech recognition).

Markov Model with outcomes

Now suppose each state Z_t also "emits" some **outcome** $X_t \in [O]$ based on the following model

$$P(X_t = o \mid Z_t = s) = b_{s,o} \qquad \textbf{(emission probability)}$$

independent of anything else.

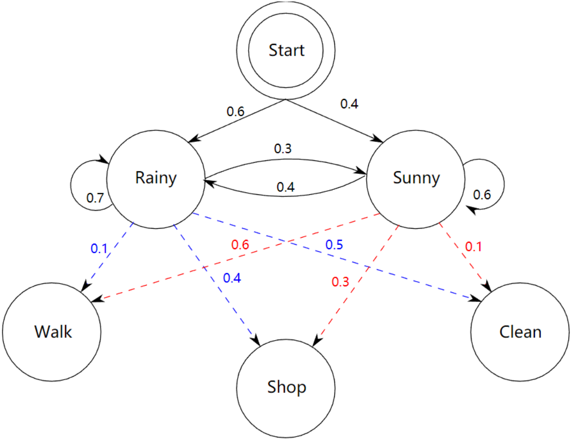
For example, in the language model, X_t is the speech signal for the underlying word Z_t (very useful for speech recognition).

Now the model parameters are $(\{\pi_s\}, \{a_{s,s'}\}, \{b_{s,o}\}) = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$.

Another example

picture from Wikipedia

On each day, we also observe **Bob's activity: walk, shop, or clean**, which only depends on the weather of that day.



Joint likelihood

The joint log-likelihood of a state-outcome sequence $z_1, x_1, \ldots, z_T, x_T$ is

$$\ln P(Z_{1:T} = z_{1:T}, X_{1:T} = x_{1:T})$$

Joint likelihood

The joint log-likelihood of a state-outcome sequence $z_1, x_1, \ldots, z_T, x_T$ is

$$
\begin{aligned}
&\ln P(Z_{1:T} = z_{1:T}, X_{1:T} = x_{1:T}) \\
&= \ln P(Z_{1:T} = z_{1:T}) + \ln P(X_{1:T} = x_{1:T} \mid Z_{1:T} = z_{1:T}) \quad \text{(always true)}
\end{aligned}
$$

Joint likelihood

The joint log-likelihood of a state-outcome sequence $z_1, x_1, \ldots, z_T, x_T$ is

$$
\begin{aligned}
&\ln P(Z_{1:T} = z_{1:T}, X_{1:T} = x_{1:T}) \\
&= \ln P(Z_{1:T} = z_{1:T}) + \ln P(X_{1:T} = x_{1:T} \mid Z_{1:T} = z_{1:T}) \quad \text{(always true)} \\
&= \sum_{t=1}^{T} \ln P(Z_t = z_t \mid Z_{t-1} = z_{t-1}) + \sum_{t=1}^{T} \ln P(X_t = x_t \mid Z_t = z_t)
\end{aligned}
$$

(due to all the independence)

Joint likelihood

The joint log-likelihood of a state-outcome sequence $z_1, x_1, \ldots, z_T, x_T$ is

$$\begin{aligned}
&\ln P(Z_{1:T} = z_{1:T}, X_{1:T} = x_{1:T}) \\
&= \ln P(Z_{1:T} = z_{1:T}) + \ln P(X_{1:T} = x_{1:T} \mid Z_{1:T} = z_{1:T}) \quad \text{(always true)} \\
&= \sum_{t=1}^{T} \ln P(Z_t = z_t \mid Z_{t-1} = z_{t-1}) + \sum_{t=1}^{T} \ln P(X_t = x_t \mid Z_t = z_t) \\
&\qquad \text{(due to all the independence)} \\
&= \ln \pi_{z_1} + \sum_{t=2}^{T} \ln a_{z_{t-1}, z_t} + \sum_{t=1}^{T} \ln b_{z_t, x_t}
\end{aligned}$$

Learning the model

If we observe N state-outcome sequences: $z_{n,1}, x_{n,1}, \ldots, z_{n,T}, x_{n,T}$ for $n = 1, \ldots, N$, the MLE can again be obtained in a similar way (verify yourself):

$$\begin{aligned} \pi_s &\propto \textbf{\#initial states with value } s \\ a_{s,s'} &\propto \textbf{\#transitions from } s \textbf{ to } s' \\ b_{s,o} &\propto \textbf{\#state-outcome pairs } (s, o) \end{aligned}$$

Learning the model

However, *most often we do not observe the states!*

Learning the model

However, *most often we do not observe the states!* Think about the speech recognition example.

Learning the model

However, *most often we do not observe the states!* Think about the speech recognition example.

This is called **Hidden Markov Model (HMM)**, widely used in practice

Learning the model

However, *most often we do not observe the states!* Think about the speech recognition example.

This is called **Hidden Markov Model (HMM)**, widely used in practice

How to learn HMMs?

Learning the model

However, *most often we do not observe the states!* Think about the speech recognition example.

This is called **Hidden Markov Model (HMM)**, widely used in practice

How to learn HMMs? **Roadmap**:

- first discuss how to **infer** when the model is known (key: dynamic programming)

Learning the model

However, *most often we do not observe the states!* Think about the speech recognition example.

This is called **Hidden Markov Model (HMM)**, widely used in practice

How to learn HMMs? **Roadmap**:

- first discuss how to **infer** when the model is known (key: dynamic programming)
- then discuss how to **learn** the model (key: EM)