CSCI-567: Machine Learning

Prof. Victor Adamchik

U of Southern California

July 29, 2020

Your model is only as good as your data.

Outline

1. Markov chain
2. Hidden Markov Models

Outline

1. Markov chain
   - Exercise
2. Hidden Markov Models

Markov Models

Markov models are powerful **probabilistic models** to analyze sequential data. A.A.Markov (1856-1922) introduced the Markov chains in 1906 when he produced the first theoretical results for stochastic processes. They are now commonly used in

- text or speech recognition
- stock market prediction
- bioinformatics
- $\cdots$

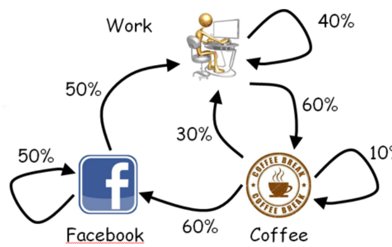
Markov chain

Directed strongly connected graph with self-loops.

Each edge labeled by a positive probability.

At each state, the probabilities on outgoing edges sum up to 1.

Transition (or stochastic) matrix: $A = a_{ij} = P(i \to j \text{ in 1 step})$.



$$A = \begin{array}{c} \\ W \\ C \\ F \end{array} \begin{array}{c} \begin{array}{ccc} W & C & F \end{array} \\ \begin{pmatrix} .4 & .6 & 0 \\ .3 & .1 & .6 \\ .5 & 0 & .5 \end{pmatrix} \end{array}$$

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 $t = 1$.

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

All X_t's take value from the same discrete set $\{1, \dots, N\}$.
We will assume that the transition probability does not change with respect to time t, i.e., a stationary Markov chain.

Markov chain

- Transition probabilities make 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.

Examples

- Example 1 (**Language model**)

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

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

  is an example of the transition probability.

- Example 2 (**Weather**)

  States $[N]$ represent weather at each day

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

Definition

A Markov chain is a stochastic process with the **Markov property**: a sequence of random variables $X_1, X_2, \dots$ s.t.

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

i.e. *the current state only depends on the most recent state*.

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

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

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

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

Chain Rule

In all derivations we will be using the chain rule:

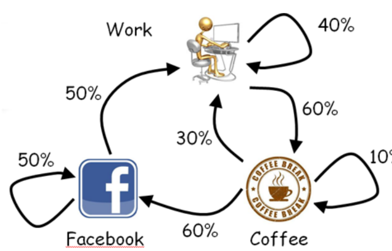
$$P(X, Y) = P(X \mid Y)\, P(Y) = P(Y \mid X)\, P(X)$$

$$P(X, Y, Z) = P(X, Y \mid Z)\, P(Z)$$

$$P(X, Y, Z) = P(X \mid Y, Z)\, P(Y \mid Z)\, P(Z)$$

Exercise 1

Consider the following Markov model. Given that now I am having Coffee, what's the probability that the next step is Facebook and the next is Work?



$P(X_3 = W, X_2 = F | X_1 = C) =$

$$= \frac{P(X_3 = W, X_2 = F, X_1 = C)}{P(X_1 = C)}$$

$$= \frac{P(X_3 = W | X_2 = F, X_1 = C) P(X_2 = F | X_1 = C) P(X_1 = C)}{P(X_1 = C)}$$

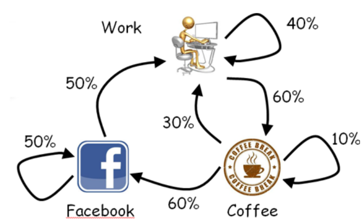
(chain rule)

$= P(X_3 = W | X_2 = F) P(X_2 = F | X_1 = C)$ (Markov rule)

$= 0.5 \times 0.6 = 0.3$

Exercise 2

Given that now I am having Coffee, what is the probability that in two steps I am at Work?



$P(X_3 = W | X_1 = C) =$

$= \sum_s P(X_3 = W, X_2 = s | X_1 = C) =$

$= P(X_3 = W | X_2 = W) P(X_2 = W | X_1 = C)$ (marginalization)

$+ P(X_3 = W | X_2 = C) P(X_2 = C | X_1 = C)$

$+ P(X_3 = W | X_2 = F) P(X_2 = F | X_1 = C)$

$= 0.3 \times 0.4 + 0.1 \times 0.3 + 0.6 \times 0.5 = 0.45$

Using a transition matrix:

$$P(X_3 = j | X_1 = i) = \sum_{k=1}^{N} a_{ik}\, a_{kj} = a^2_{ij}$$

Parameter estimation for Markov models

Now suppose we have observed M **sequences of examples**:

- $x_{1,1}, \ldots, x_{1,T}$
- $\cdots$
- $x_{M,1}, \ldots, x_{M,T}$

where

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

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

Finding the MLE

Same story, **Maximum Likelihood Estimation**:

$$\operatorname*{argmax}_{\boldsymbol{\pi}, \boldsymbol{A}} \ln P(X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T)$$

First, we need to compute this joint probability. Applying the chain rule for random variables, we get

$$\begin{aligned} & P(X_1, X_2, \ldots, X_T) \\ &= P(X_2, X_3, \ldots, X_T | X_1) P(X_1) \\ &= P(X_3, \ldots, X_T | X_1, X_2) P(X_2 | X_1) P(X_1) \\ &= \cdots = \\ &= P(X_1) \prod_{t=2}^{T} P(X_t | X_1, \ldots, X_{t-1}) \\ &= P(X_1) \prod_{t=2}^{T} P(X_t | X_{t-1}) \qquad \text{(Markov property)} \end{aligned}$$

Finding the MLE

The log-likelihood of a sequence $x_1, \ldots, x_T$ is

$$\begin{aligned} & \ln P(X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T) \\ &= P(X_1 = x_1) + \sum_{t=2}^{T} \ln P(X_t = x_t \mid X_{t-1} = x_{t-1}) \\ &= \ln \pi_{x_1} + \sum_{t=2}^{T} \ln a_{x_{t-1}, x_t} \\ &= \sum_n \mathbb{I}[x_1 = n] \ln \pi_n + \sum_{n,n'} \left(\sum_{t=2}^{T} \mathbb{I}[x_{t-1} = n, x_t = n'] \right) \ln a_{n,n'} \end{aligned}$$

Finding the MLE

So MLE is

$$\begin{aligned} \operatorname*{argmax}_{\boldsymbol{\pi}, \boldsymbol{A}} & \sum_n (\textbf{\#initial states with value } n) \ln \pi_n \\ & + \sum_{n,n'} (\textbf{\#transitions from } n \textbf{ to } n') \ln a_{n,n'} \end{aligned}$$

We have seen this many times. The solution is (derivation is left as an exercise):

$$\pi_n = \frac{\#\text{of sequences starting with } n}{\#\text{of sequences}}$$

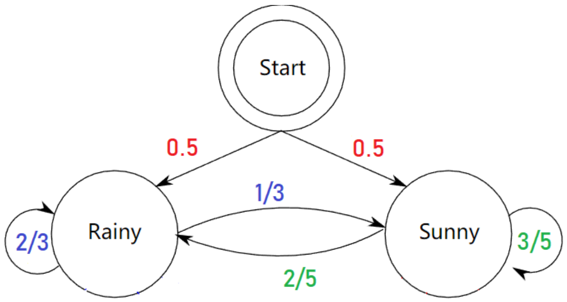
$$a_{n,n'} = \frac{\#\text{of transitions from } n \text{ to } n'}{\#\text{of transitions starting with } n}$$

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



Problem 1

Suppose that we didn't know the emission probabilities or transition probabilities for this HMM. Instead, we had to estimate them from data. Consider the following data set:

```
state: S S V V V S S S S S V S V V S V S S V V
  obs: G F G G F F F F G F G G G G F G F F G G
```

Based on this data, estimate the emission and the transition probabilities for this HMM.

Solution

Outline

1. Markov chain
2. Hidden Markov Models
   - Forward and backward messages
   - Viterbi Algorithm
   - Viterbi Algorithm: Example
   - Exercise
   - Learning HMMs

Markov Model with outcomes

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

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

independent of anything else.

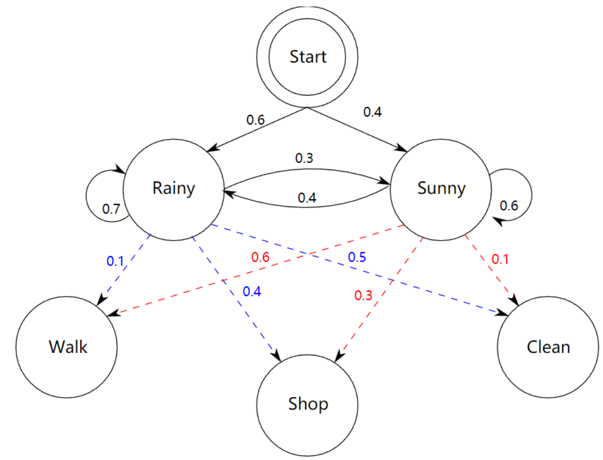
For example, in the language model, O_t is the speech signal for the underlying word X_t (very useful for speech recognition).

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.



HMM defines a joint probability

$$P(X_1, X_2, \cdots, X_T, O_1, O_2, \cdots, O_T)$$
$$= P(X_1, X_2, \cdots, X_T)\, P(O_1, O_2, \cdots, O_T \mid X_1, X_2, \cdots, X_T)$$

- Markov assumption simplifies the first term

$$P(X_1, X_2, \cdots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1})$$

- The *independence* assumption simplifies the second term

$$P(O_1, O_2, \cdots, O_T \mid X_1, X_2, \cdots, X_T) = \prod_{t=1}^{T} P(O_t \mid X_t)$$

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

Joint likelihood

The joint log-likelihood is

$$\ln P(X_1 = x_1, X_2 = x_2, \cdots, X_T = x_T, O_1 = o_1, O_2 = o_2, \cdots, O_T = o_T)$$
$$= \ln P(X_1 = x_1) \prod_{t=2}^{T} P(X_t = x_t \mid X_{t-1} = x_{t-1}) \prod_{t=1}^{T} P(O_t = o_t \mid X_t = x_t)$$
$$= \ln P(X_1 = x_1) + \sum_{t=2}^{T} \ln P(X_t = x_t \mid X_{t-1} = x_{t-1})$$
$$+ \sum_{t=1}^{T} \ln P(O_t = o_t \mid X_t = x_t)$$
$$= \ln \pi_{x_1} + \sum_{t=2}^{T} \ln a_{x_{t-1}, x_t} + \sum_{t=1}^{T} \ln b_{x_t, o_t}$$

Learning the model

If we observe M state-outcome sequences: $x_{m,1}, o_{m,1}, \ldots, x_{m,T}, o_{m,T}$ for $m = 1, \ldots, M$, the MLE is again very simple (verify yourself):

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

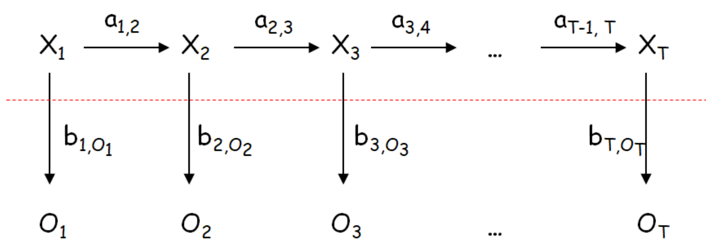
Learning the model

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

This is called **Hidden Markov Model (HMM)**.

Notice that "hidden" is referred to the states of the Markov chain, not to the parameters of the model.

A generic hidden Markov model is illustrated in this picture:



HMM problems

There are three fundamental problems that we solve:

- **Problem 1**: Scoring and evaluation

  Given an observation sequence $O_1, O_2, \ldots, O_T$ and a model $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, how to compute efficiently the probability of $P(O_1, O_2, \ldots, O_T)$?

HMM problems

There are three fundamental problems that we solve:

- **Problem 2**: Decoding (Viterbi algorithm)

  Given an observation sequence $O_1, O_2, \ldots, O_T$ and a model $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, how do we determine the optimal corresponding state sequence $X_1, X_2, \ldots, X_T$ that best explains how the observations were generated?

HMM problems

There are three fundamental problems that we solve:

- **Problem 3**: Training

  Given an observation sequence $O_1, O_2, \ldots, O_T$, how to adjust the parameters $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ to maximize the probability of $P(O_1, O_2, \ldots, O_T)$? In the other words, find a model to best fit the observed data. we will solve this by the Baum–Welch algorithm.

Chain Rule

In all derivations we will be using the chain rule to calculate any member of the joint distribution using only conditional probabilities.

$$P(X, Y) = P(X \mid Y)\ P(Y) = P(Y \mid X)\ P(X)$$

$$P(X, Y, Z) = P(X, Y \mid Z)\ P(Z)$$

$$P(X, Y, Z) = P(X \mid Y, Z)\ P(Y \mid Z)\ P(Z)$$

Forward and backward messages

The key is to compute two things:

- **forward messages**: for each s and t

$$\alpha_s(t) = P(X_t = s, O_{1:t} = o_{1:t})$$

  The intuition is, if we observe up to time t, what is the likelihood of the Markov chain in state s?

- **backward messages**: for each s and t

$$\beta_s(t) = P(O_{t+1:T} = o_{t+1:T} \mid X_t = s)$$

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

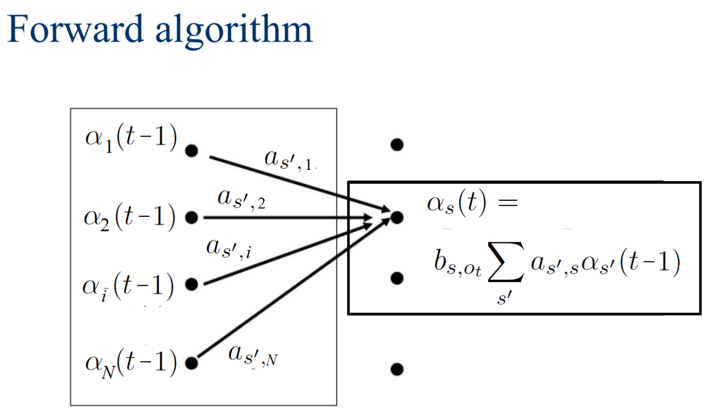
Computing forward messages

Key: *establish a recursive formula*

$$\begin{aligned}
\alpha_s(t) &= P(X_t = s, O_{1:t}) = P(X_t = s, O_{1:t-1}, O_t) \\
&= P(O_t \mid X_t = s, O_{1:t-1}) P(X_t = s, O_{1:t-1}) \\
&= P(O_t \mid X_t = s) P(X_t = s, O_{1:t-1}) && \text{(independence)} \\
&= b_{s,o_t} \sum_{s'} P(X_t = s, X_{t-1} = s', O_{1:t-1}) && \text{(marginalizing)} \\
&= b_{s,o_t} \sum_{s'} P(X_t = s | X_{t-1} = s', O_{1:t-1}) P(X_{t-1} = s', O_{1:t-1}) \\
&= b_{s,o_t} \sum_{s'} P(X_t = s | X_{t-1} = s') P(X_{t-1} = s', O_{1:t-1}) \\
&= b_{s,o_t} \sum_{s' \in [N]} a_{s',s} \alpha_{s'}(t-1)
\end{aligned}$$

Base case: $\alpha_s(1) = P(X_1, O_1) = P(O_1 | X_1) P(X_1) = \pi_s b_{s,o_1}$

Forward procedure



Forward procedure

Forward algorithm

For all $s \in [N]$, compute $\alpha_s(1) = \pi_s b_{s,o_1}$.

For $t = 2, \ldots, T$

- for each $s \in [N]$, compute

$$\alpha_s(t) = b_{s,o_t} \sum_{s' \in [N]} a_{s',s}\, \alpha_{s'}(t-1)$$

It takes $O(N^2T)$ time and $O(NT)$ space using dynamic programming.

Oh, no, CSCI-570 again..

Computing backward messages

Again establish a recursive formula

$$\beta_s(t) = P(O_{t+1:T} \mid X_t = s) = P(O_{t+1:T}, X_t = s)/P(X_t = s) =$$

$$= \sum_{s'} P(O_{t+1:T}, X_{t+1} = s', X_t = s)/P(X_t = s) \qquad \text{(marginalizing)}$$

$$= \sum_{s'} P(O_{t+1:T} \mid X_{t+1} = s', X_t = s)P(X_{t+1} = s' \mid X_t = s)$$

$$= \sum_{s'} a_{s,s'} P(O_{t+1:T} \mid X_{t+1} = s') = \sum_{s'} a_{s,s'} P(O_{t+1}, O_{t+2:T} \mid X_{t+1} = s')$$

$$= \sum_{s'} a_{s,s'} P(O_{t+1} \mid O_{t+2:T}, X_{t+1} = s')P(O_{t+2:T} \mid X_{t+1} = s')$$

$$= \sum_{s'} a_{s,s'} b_{s',o_{t+1}} \beta_{s'}(t+1)$$

Base case: $\beta_s(T) = 1$ (prove it!)

Backward procedure

Backward algorithm

For all $s \in [N]$, set $\beta_s(T) = 1$.

For $t = T-1, \ldots, 1$

- for each $s \in [N]$, compute

$$\beta_s(t) = \sum_{s' \in [N]} a_{s,s'}\, b_{s',o_{t+1}} \beta_{s'}(t+1)$$

Again it takes $O(N^2T)$ time and $O(NT)$ space.

Solving Problem 1

With forward messages $\alpha_s(t) = P(X_t = s, O_{1:t})$, we can compute $P(O_{1:T})$.

Indeed,
$$P(O_{1:T}) = \sum_s P(O_{1:T}, X_T = s) = \sum_s \alpha_s(T)$$

Solving Problem 2

Given the model and a sequence of observations, our goal is to find the most likely sequence of states that maximizes $P(X_{1:T}, O_{1:T})$.

This is called Viterbi decoding. We solve this using Dynamic Programming.

We define DP subproblems in the following way – the highest probable state sequence that ends at $X_t = s$ given observations $O_1, O_2, \dots, O_t$

$$\delta_s(t) = \max_{X_{1:t-1}} P(X_{1:t-1}, X_t = s, O_{1:t})$$

In the next slide we compute $\delta_s(t)$ recursively.

Computing $\delta_s(t)$

The goal is to get a recurrence. We will use $X_t = s, X_{t-1} = s'$.

$$\begin{aligned}
\delta_s(t) &= \max_{X_{1:t-1}} P(X_t = s, X_{1:t-1}, O_{1:t}) \\
&= \max_{X_{1:t-1}} P(X_t = s, O_t, X_{1:t-1}, O_{1:t-1}) \\
&= \max_{s'} P(X_t = s, O_t \mid X_{1:t-1}, O_{1:t-1}) \max_{X_{1:t-2}} P(X_{1:t-1}, O_{1:t-1}) \\
&= \max_{s'} \delta_{s'}(t-1) P(X_t, O_t \mid X_{1:t-1}, O_{1:t-1}) \\
&= \max_{s'} \delta_{s'}(t-1) P(O_t, X_t \mid X_{1:t-1}) \\
&= \max_{s'} \delta_{s'}(t-1) P(O_t \mid X_t, X_{1:t-1}) P(X_t \mid X_{1:t-1}) \\
&= \max_{s'} \delta_{s'}(t-1) P(O_t \mid X_t) P(X_t \mid X_{t-1}) \\
&= b_{s,o_t} \max_{s'} a_{s',s} \delta_{s'}(t-1)
\end{aligned}$$

Base case: $\delta_s(1) = P(X_1 = s, O_1 = o_1) = \pi_s b_{s,o_1}$

The optimal path

Note that this only gives the optimal probability, not the optimal path itself.

$$\delta_s(t) = b_{s,o_t} \max_{s'} a_{s',s} \delta_{s'}(t-1)$$

We need to keep a track of each preceding state where the maximum occurs. Thus we create a table to record the highest-scoring state at each possible state at each time-stamp.

$$\Delta_s(t) = \operatorname*{argmax}_{s'} a_{s',s} \delta_{s'}(t-1)$$

This must remind you Dijkstra's shortest path algorithm from CS570.

Viterbi Algorithm

Viterbi Algorithm

For each $s \in [N]$, compute $\delta_s(1) = \pi_s b_{s,o_1}$.

For each $t = 2, \ldots, T$,

- for each $s \in [N]$, compute

$$\delta_s(t) = b_{s,o_t} \max_{s'} a_{s',s} \delta_{s'}(t-1)$$

$$\Delta_s(t) = \operatorname{argmax}_{s'} a_{s',s} \delta_{s'}(t-1)$$

Backtracking: let $o_T^* = \operatorname{argmax}_s \delta_s(T)$.
For each $t = T, \ldots, 2$: set $o_{t-1}^* = \Delta_{o_t^*}(t)$.

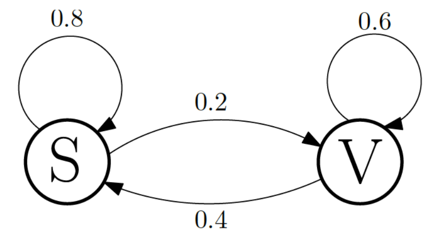
Output the most likely path $o_1^*, \ldots, o_T^*$.

Example

Consider the HMM below. In this world, every time step (say every few minutes), you can either be Studying or playing Video games. You're also either Grinning or Frowning while doing the activity.

E	$p(E \mid X=S)$
grin	0.5
frown	0.5

E	$p(E \mid X=V)$
grin	0.8
frown	0.2



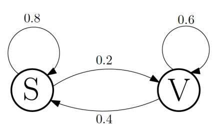
Suppose that we believe that the initial state distribution is 50/50. We observe: Grin, Grin, Frown, Grin. What is the most likely path for this sequence of observations?

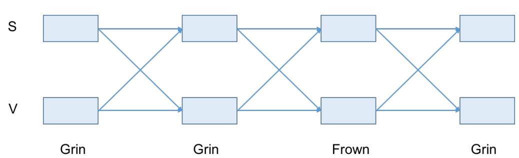
$t = 1$, the initial time

$\delta_s(1) = \pi_s b_{s,o_1}$. Compute $\delta_S(1)$ and $\delta_V(1)$.

E	$p(E \mid X=S)$
grin	0.5
frown	0.5

E	$p(E \mid X=V)$
grin	0.8
frown	0.2





$$\delta_S(1) = P(O_1 = Grin \mid X_1 = S)\pi(X_1 = S) = 0.5 \times 0.5 = 0.25$$

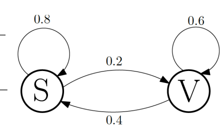
$$\delta_V(1) = P(O_1 = Grin \mid X_1 = V)\pi(X_1 = V) = 0.8 \times 0.5 = 0.4$$

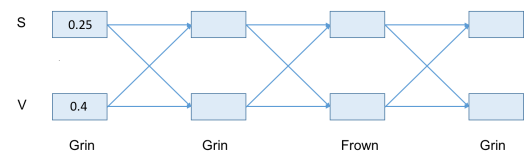
$t = 2$

$\delta_s(t) = b_{s,o_t} \max_{s'} a_{s',s} \delta_{s'}(t-1)$. Compute $\delta_S(2)$ and $\delta_V(2)$.

E	$p(E \mid X=S)$
grin	0.5
frown	0.5

E	$p(E \mid X=V)$
grin	0.8
frown	0.2

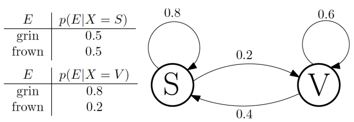


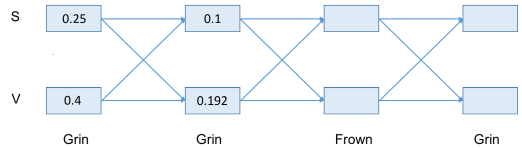


$$\begin{aligned}
\delta_S(2) &= P(O_2 = Grin \mid X_2 = S) \times \\
&\quad \max\{P(X_2 = S \mid X_1 = S)\delta_S(1), P(X_2 = S \mid X_1 = V)\delta_V(1)\} \\
&\quad = 0.5 \times \max\{0.8 \times 0.25, 0.4 \times 0.4\} = 0.01 \\
\delta_V(2) &= P(O_2 = Grin \mid X_2 = V) \times \\
&\quad \max\{P(X_2 = V \mid X_1 = S)\delta_S(1), P(X_2 = V \mid X_1 = V)\delta_V(1)\} \\
&\quad 0.8 \times \max\{0.2 \times 0.25, 0.6 \times 0.4\} = 0.192 \\
\Delta_S(2) &= S, \Delta_V(2) = S
\end{aligned}$$

$t = 3$

$\delta_s(t) = b_{s,o_t} \max_{s'} a_{s',s}\delta_{s'}(t-1)$. Compute $\delta_S(3)$ and $\delta_V(3)$.

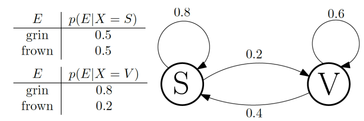


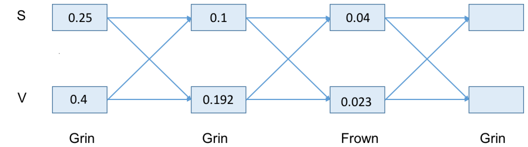


$$
\begin{aligned}
\delta_S(3) &= P(O_3 = Frown|X_3 = S)\times \\
&\quad \max\{P(X_3 = S|X_2 = S)\delta_S(2), P(X_3 = S|X_2 = V)\delta_V(2)\} \\
&\quad = 0.5 \times \max\{0.8 \times 0.1, 0.4 \times 0.192\} = 0.04 \\
\delta_V(3) &= P(O_3 = Frown|X_3 = V)\times \\
&\quad \max\{P(X_3 = V|X_2 = S)\delta_S(2), P(X_3 = V|X_2 = V)\delta_V(2)\} \\
&\quad = 0.2 \times \max\{0.2 \times 0.1, 0.6 \times 0.192\} = 0.023 \\
\Delta_S(3) &= S, \Delta_V(3) = V
\end{aligned}
$$

$t = 4$

Observation at $t = 4$ is 'Grin'



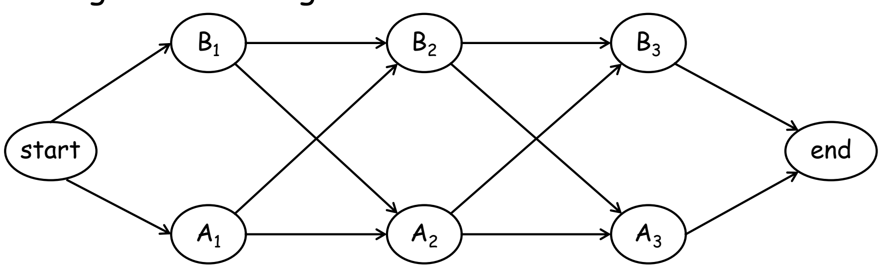


$$
\begin{aligned}
\delta_S(4) &= P(O_4 = Grin|x_4 = S)\times \\
&\quad \max\{P(X_4 = S|X_3 = S)\delta_S(3), P(X_4 = S|X_3 = V)\delta_V(3)\} \\
&\quad = 0.5 \times \max\{0.8 \times 0.04, 0.4 \times 0.023\} = 0.016 \\
\delta_V(4) &= P(O_4 = Grin|X_4 = V)\times \\
&\quad \max\{P(X_4 = V|X_3 = S)\delta_S(3), P(X_4 = V|X_3 = V)\delta_V(3)\} \\
&\quad = 0.8 \times \max\{0.2 \times 0.04, 0.6 \times 0.023\} = 0.011 \\
\Delta_S(4) &= S, \Delta_V(4) = V
\end{aligned}
$$

Then the path is $S(4) \leftarrow S(3) \leftarrow S(2) \leftarrow S(1)$. Please verify!

Problem 2

Assuming the following HMM



with the following transition and emission probabilities

	A	B	End
Start	0.7	0.3	0
A	0.2	0.7	0.1
B	0.7	0.2	0.1

	S	x	y
Start	1	0	0
A	0	0.4	0.6
B	0	0.3	0.7

What is the most likely sequence of states that produced the input sequence *xyy* ?

Solution

Solution

Problem 3: Learning HMM

Given a sequence of observations, our goal is to adjust the model parameters $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ to best fit the observations (to maximize the probability of $P(O_1, O_2, \ldots, O_T)$).

First, we define

$$\gamma_s(t) = P(s \mid O_{1:T})$$

as a probability of being at state $X_t = s$ at time t. e.g. given Bob's activities for one week, how was the weather like on Wed?

$\gamma_s(t)$ is computed using forward and backward messages:

$$\gamma_s(t) = \frac{\alpha_s(t)\beta_s(t)}{P(O_{1:T})}$$

Here the denominator is the solution to Problem 1.

Computing $\gamma_s(t)$

$$\begin{aligned}\gamma_s(t) &= P(X_t = s \mid O_{1:T}) \propto P(X_t = s, O_{1:T}) \\ &= P(X_t = s, O_{1:t}, O_{t+1:T}) \\ &= P(X_t = s, O_{1:t})P(O_{t+1:T} \mid X_t = s, O_{1:t}) \\ &= P(X_t = s, O_{1:t})P(O_{t+1:T} \mid X_t = s) \\ &= \alpha_s(t)\beta_s(t)\end{aligned}$$

What constant are we omitting in "$\propto$"? It is exactly

$$P(O_{1:T}) = \sum_{s\in[N]} P(O_{1:T}, X_T = s) = \sum_{s\in[N]} \alpha_s(T) = \sum_{s\in[N]} \alpha_s(t)\beta_s(t)$$

This is true for any t; a good way to check correctness of your code.

Problem 3

Next, we define

$$\xi_{s,s'}(t) = P(s, s' \mid O_{1:T})$$

a probability of being at state $X_t = s$ at time t and at state $X_{t+1} = s'$ at time $t+1$, e.g. given Bob's activities for one week, how was the weather like on Wed and Thu?

This probability is computed using forward and backward messages:

$$\xi_{s,s'}(t) = \frac{\alpha_s(t)\, a_{s,s'}\, b_{s',O_{t+1}}\, \beta_{s'}(t+1)}{P(O_{1:T})}$$

Here $\gamma_s(t)$ and $\xi_{s,s'}(t)$ are related by

$$\sum_{s'} \xi_{s,s'}(t) = \gamma_s(t)$$

Computing $\xi_{s,s'}(t)$

$$
\begin{aligned}
\xi_{s,s'}(t) &= P(X_t = s, X_{t+1} = s' \mid O_{1:T}) \\
&\propto P(X_t = s, X_{t+1} = s', O_{1:T}) \\
&= P(X_t = s, O_{1:t}, X_{t+1} = s', O_{t+1:T}) \\
&= P(X_t = s, O_{1:t}) P(X_{t+1} = s', O_{t+1:T} \mid X_t = s, O_{1:t}) \\
&= \alpha_s(t)\, P(X_{t+1} = s', O_{t+1:T} \mid X_t = s) \\
&= \alpha_s(t)\, P(O_{t+1:T} \mid X_{t+1} = s', X_t = s) P(X_{t+1} = s' \mid X_t = s) \\
&= \alpha_s(t) P(O_{t+1:T} \mid X_{t+1} = s')\, a_{s,s'} \\
&= \alpha_s(t)\, a_{s,s'}\, P(O_{t+1:T}, O_{t+2:T} \mid X_{t+1} = s') \\
&= \alpha_s(t)\, a_{s,s'}\, P(O_{t+2:T} \mid X_{t+1} = s', O_{t+1:T}) P(O_{t+1} \mid X_{t+1} = s') \\
&= \alpha_s(t)\, a_{s,s'}\, P(O_{t+2:T} \mid X_{t+1} = s') P(O_{t+1} \mid X_{t+1} = s') \\
&= \alpha_s(t)\, a_{s,s'}\, b_{s',o_{t+1}}\, \beta_{s'}(t+1)
\end{aligned}
$$

The normalization constant is in fact again $P(O_{1:T})$

Outline

The Baum–Welch algorithm

The algorithm trains both the transition probabilities A and the emission probabilities B of the HMM.

The Baum–Welch algorithm (1972) is a special case of the more general Expectation-Maximization (EM) algorithm (1977).

EM is an iterative algorithm, computing an initial estimate for the probabilities, then using those estimates to computing a better estimate, and so on, iteratively improving the probabilities that it learns.