CSCI-567: Machine Learning

Prof. Victor Adamchik

U of Southern California

July 30, 2020

Your model is only as good as your data.

Outline

Outline

Definition

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

We denote the transition and initial probabilities as

$$a_{s,s'} = P(X_{t+1} = s'|X_t = s), \quad \pi_s = P(X_1 = s)$$

Each state $X_t \in 1, 2, \ldots, S$ also "emits" some **outcome** O_t based on the following model

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

independent of anything else.

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

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 very simple.

However, *most often we do not observe the states!* Think about the speech recognition example. This is called **Hidden Markov Model (HMM)**.

There are three fundamental problems that we solve.

HMM problems

- **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 $P(O_1, O_2, \ldots, O_T)$?

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

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

HMM problems

- **Problem 2**: Decoding (the 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?

We solve this using Dynamic Programming.

Let $\delta_s(t)$ be the optimal probability of a sequence that ends at $X_t = s$ given observations $O_1, O_2, \ldots, O_t$

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

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

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

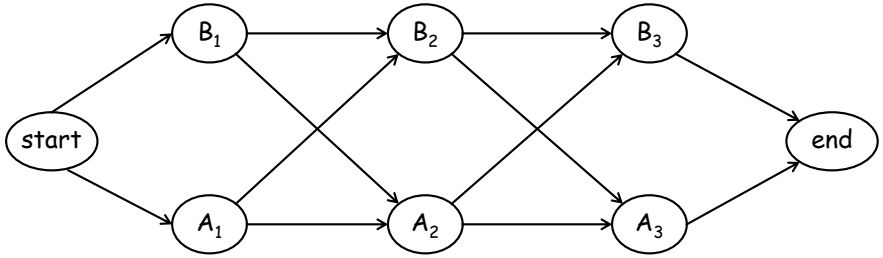
$$\Delta_s(t) = \operatorname*{argmax}_{s'} a_{s',s} \delta_{s'}(t-1)$$

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 $o_1^*, \ldots, o_T^*$.

Exercise 1

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

Outline

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

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

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:

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

Computing $\gamma_s(t)$

Exercise 2

Using $\gamma_s(t) = P(X_t = s|O_{1:T})$, prove that $\beta_s(T) = 1$.

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

Next, we define

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

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:

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

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

Exercise 3

Compute

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

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.

The Baum–Welch algorithm

The solution to Problem 3 can be summarized as follows:

- Initialize the parameters $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$.
- Compute $\alpha_s(t), \beta_s(t), \gamma_s(t)$ and $\xi_{s,s'}(t)$.
- Update the model parameters $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$.
- If $P(O_{1:T})$ increases, goto to the second step.

Initialization

We initialize $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ with a best guess or randomly (uniformly)

$$\pi_s \sim 1/N, \quad a_{s,s'} \sim 1/N, \quad b_{s,O_t} \sim 1/N.$$

Note, the parameters must be row stochastic:

$$\sum_i \pi_i = 1, \quad \sum_j a_{i,j} = 1, \quad \sum_j b_{i,j} = 1.$$

Updating the model parameters

Compute new initial probability in state s by:

$$\pi_s = \gamma_s(1)$$

A new transition probability from state s to state s' is computed by:

$$a_{s,s'} = \sum_{t=1}^{T-1} \xi_{s,s'}(t) \Big/ \sum_{t=1}^{T-1} \gamma_s(t)$$

The numerator here is the expected number of transitions from state s to state s'.

The denominator is the expected number of transitions from s to any state.

Updating the model parameters

A new emission probability in state s observing $O_t = k$ is computed by:

$$b_{s,k} = \sum_{t=1}^{T} \mathbb{I}\left[O_t == k\right] \gamma_s(t) \Big/ \sum_{t=1}^{T} \gamma_s(t)$$

where $\mathbb{I}\left[x\right]$ denotes an indicator function.

The denominator here is the expected number of times the model is in state s.

The numerator is the expected number of times the model is in state s with observation $O_t = k$.

General EM algorithm

Step 0 Initialize $\theta = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$.

Step 1 (E-Step) update the posterior of latent variables

$$q = P(X_t = s \mid O_{1:T}\,; \theta)$$

and obtain expectation of complete likelihood

$$Q(\theta) = \mathbb{E}_{X_{1:T} \sim q}\left[\ln P(O_{1:T}, X_{1:T}\,; \theta)\right]$$

Step 2 (M-Step) update the model parameter via maximization

$$\operatorname*{argmax}_{\theta} Q(\theta)$$

Step 3 goto Step 1 if not converged

Applying EM: E-Step

In the E-Step we fix the parameters and find the posterior distributions q of the hidden states (for each sample)

$$q = P(X_t = s \mid O_{1:T}\,; \theta) = \gamma_s(t)$$

This leads to the complete log-likelihood:

$$Q(\theta) = \mathbb{E}_{X_{1:T} \sim q}\left[\ln P(X_{1:T}, O_{1:T})\right]$$

We showed in the previous lecture that

$$\ln P(X_{1:T}, O_{1:T}) = \ln \pi_{X_1} + \sum_{t=2}^{T} \ln a_{X_{t-1}, X_t} + \sum_{t=1}^{T} \ln b_{X_t, O_t}$$

Applying EM: E-Step

It follows,

$$\begin{aligned} Q(\theta) &= \mathbb{E}_{X_{1:T} \sim q}\left[\ln \pi_{X_1} + \sum_{t=2}^{T} \ln a_{X_{t-1}, X_t} + \sum_{t=1}^{T} \ln b_{X_t, O_t}\right] \\ &= \sum_{s} \gamma_s(1) \ln \pi_s + \sum_{t=1}^{T-1} \sum_{s,s'} \xi_{s,s'}(t) \ln a_{s,s'} + \sum_{t=1}^{T} \sum_{s} \gamma_s(t) \ln b_{s,O_t} \end{aligned}$$

In the first term we are repeatedly selecting the values of X_1, so it is just the marginal expression for time $t = 1$.

In the second term we are looking over all transitions from X_{t-1} to X_t and weighting that by the corresponding probability.

In the last term we are looking at the emissions for all states and weighting each possible emission by the corresponding probability, so that is just the sum of the marginal for time t.

Applying EM: E-Step

Let us maximize the first term.

Adding the Lagrange multiplier λ, using the constraint that $\sum_s \pi_s = 1$, and setting the derivative equal to zero, we get

$$\frac{\partial}{\partial \pi_s}\left(\sum_s \gamma_s(1) \ln \pi_s + \lambda\left(1 - \sum_s \pi_s\right)\right) = 0$$

Take the derivative

$$\frac{\gamma_s(1)}{\pi_s} = \lambda$$

then use the constraint $\sum_s \pi_s = 1$, to get $\sum_s \gamma_s(1) = \lambda$.

So we get

$$\pi_s = \frac{\gamma_s(1)}{\sum_s \gamma_s(1)} = \gamma_s(1)$$

Applying EM: E-Step

Let us maximize the third term:

$$\sum_{t=1}^{T} \sum_{s} \gamma_s(t) \ln b_{s,O_t}$$

where

$$\begin{aligned}\gamma_s(t) &= P(X_t = s \mid O_{1:T}) \\ &= P(X_t = s, O_t = k \mid O_{1:T}) + P(X_t = s, O_t \neq k \mid O_{1:T})\end{aligned}$$

Adding the Lagrange multiplier λ, using the constraint that $\sum_k b_{s,k} = 1$, and setting the derivative equal to zero, we get

$$\frac{\partial}{\partial b_{s,O_t}} \left(\sum_{t=1}^{T} \sum_{s} \gamma_s(t) \ln b_{s,O_t} + \lambda \left(1 - \sum_{k} b_{s,k} \right) \right) = 0$$

Applying EM: E-Step

Take the derivative to get

$$\frac{1}{b_{s,k}} \sum_{t=1}^{T} P(X_t = s, O_t = k \mid O_{1:T}) = \lambda$$

Next we sum it up over k

$$\sum_{k} \sum_{t=1}^{T} P(X_t = s, O_t = k \mid O_{1:T}) = \lambda \sum_{k} b_{s,k}$$

Using marginalization and the the constraint $\sum_k b_{s,k} = 1$, we get

$$\sum_{t=1}^{T} \gamma_s(t) = \lambda$$

Applying EM: E-Step

So, it follows

$$b_{s,k} = \frac{\sum_{t=1}^{T} P(X_t = s, O_t = k \mid O_{1:T})}{\sum_{t=1}^{T} \gamma_s(t)}$$

which could be also written as in slide 19

$$b_{s,k} = \frac{\sum_{t=1}^{T} \gamma_s(t) \mathbb{I}\left[O_t == k\right]}{\sum_{t=1}^{T} \gamma_s(t)}$$

Exercise 4

Compute a new transition probability from state s to state s' by maximizing the complete log-likelihood $\mathrm{Q}(\Theta)$ from the lecture slide 21.