CS540 Introduction to Artificial Intelligence
Lecture 23

Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 13, 2020

Special Bayesian Network for Sequences

Motivation

- A sequence of features $X_1, X_2, ...$ can be modelled by a Markov Chain but they are not observable.
- A sequence of labels $Y_1, Y_2, ...$ depends only on the current hidden features and they are observable.
- This type of Bayesian Network is call a Hidden Markov Model.

HMM Applications Part 1

Motivation

- Weather prediction.
- Hidden states: $X_1, X_2, \ldots$ are weather that is not observable by a person staying at home (sunny, cloudy, rainy).
- Observable states: $Y_1, Y_2, \ldots$ are Badger Herald newspaper reports of the condition (dry, dryish, damp, soggy).
- Speech recognition.
- Hidden states: $X_1, X_2, \ldots$ are words.
- Observable states: $Y_1, Y_2, \ldots$ are acoustic features.

sounds

re k a ni s b ch

wreck a nice beach

HMM Applications Part 2

Motivation

- Stock or bond prediction.
- Hidden states: $X_1, X_2, \ldots$ are information about the compnay (profitability, risk measures).
- Observable states: $Y_1, Y_2, \ldots$ are stock or bond prices.

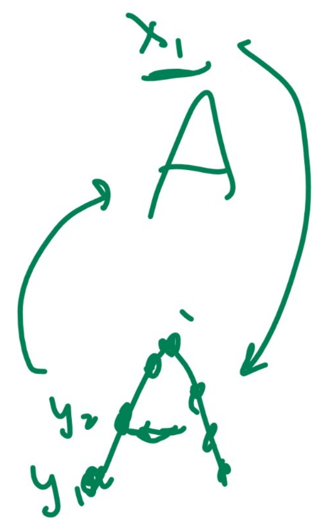
Y_{t+1}

- Speech synthesis: Chatbox.
- Hidden states: $X_1, X_2, \ldots$ are context or part of speech.
- Observable states: $\boxed{Y_1, Y_2, \ldots}$ are words.

Other HMM Applications

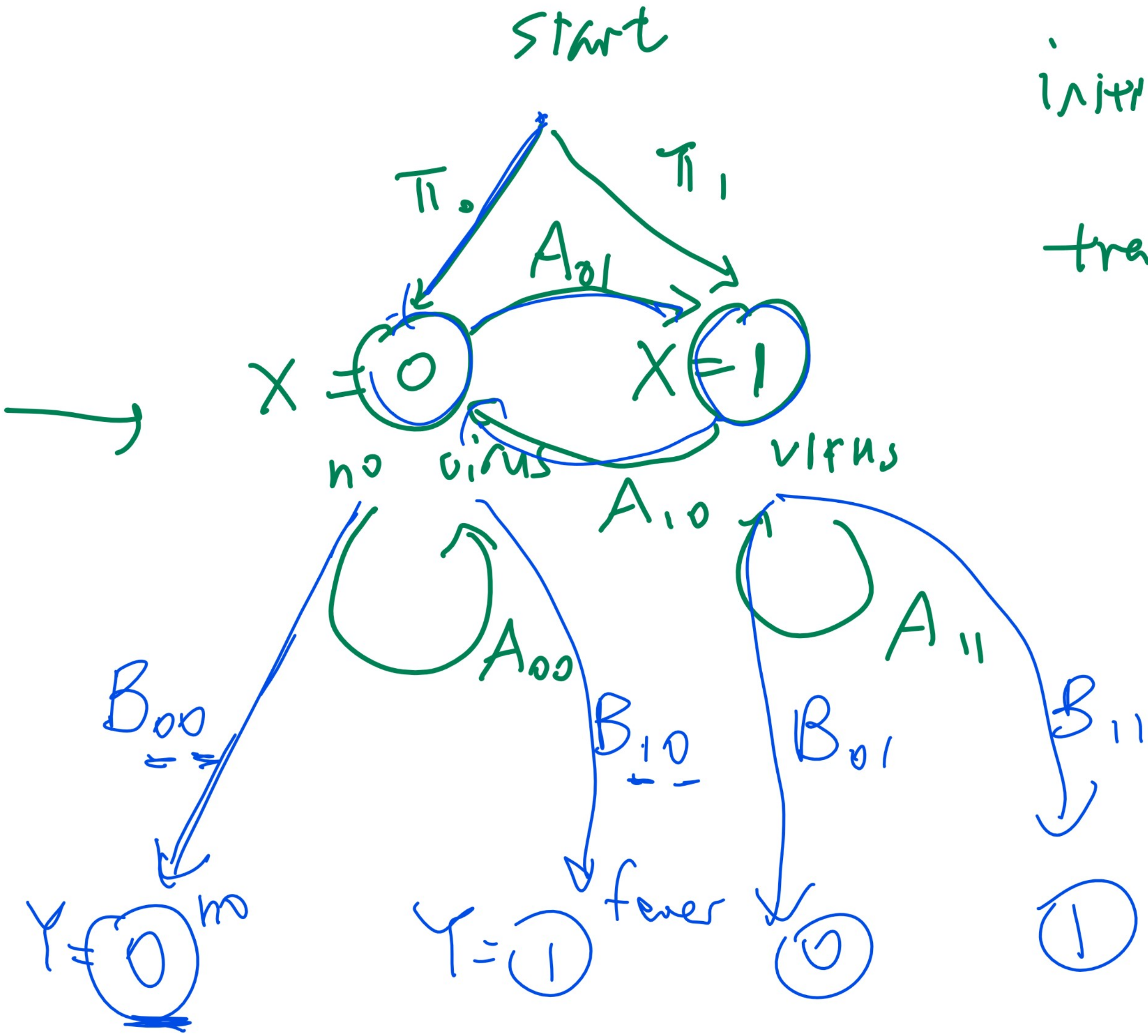
Motivation

- Machine translation.
- Handwritting recognition.
- Gene prediction.
- Traffic control.



Hidden Markov Model Diagram

Motivation



initial prob (π_0, π_1)

transition matrix $A = \begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix}$

Conditional prob

$B_0 = [\underline{B_{00}}, B_{01}]$

$B_1 = [B_{10}, B_{11}]$

Transition and Likehood Matrices

Motivation

- An initial distribution vector and two state transition matrices are used to represent a hidden Markov model.

1. Initial state vector: π.

$$\pi_i = \mathbb{P}\{X_1 = i\}, i \in 1, 2, ..., |X|$$

2. State transition matrix: A.

$$A_{ij} = \mathbb{P}\{X_t = j | X_{t-1} = i\}, i, j \in 1, 2, ..., |X|$$

3. Observation Likelihood matrix (or output probability distribution): B.

$$B_{ij} = \mathbb{P}\{Y_t = i | X_t = j\}, i \in 1, 2, ..., |Y|, j \in 1, 2, ..., |X|$$

Markov Property

Motivation

- The Markov property implies the following conditionally independence property.

$$\mathbb{P}\left\{x_t | x_{t-1}, x_{t-2}, ..., x_1\right\} = \mathbb{P}\left\{x_t | x_{t-1}\right\}$$
$$\mathbb{P}\left\{y_t | x_t, x_{t-1}, ..., x_1\right\} = \mathbb{P}\left\{y_t | x_t\right\}$$

Evaluation and Training

Motivation

- There are three main tasks associated with a HMM.

1. Evaluation problem: finding the probability of an observed sequence given an HMM: $y_1, y_2, \dots$
2. Decoding problem: finding the most probable hidden sequence given the observed sequence: $x_1, x_2, \dots$
3. Learning problem: finding the most probable HMM given an observed sequence: $\pi, A, B, \dots$

Training HMM

CPT

EM

Expectation Maximization Algorithm

Description

- Start with a random guess of π, A, B.
- Compute the forward probabilities: the joint probability of a observed sequence and its hidden state.
- Compute the backward probabilities: the probability of a observed sequence given its hidden state.
- Update the model π, A, B using Bayes rule.
- Repeat until convergence.
- Sometimes, it is called the Baum-Welch Algorithm.

Evaluation Problem

Definition

=0.1 =0.1 =0.1

- The task is to find the probability $\mathbb{P}\{y_1, y_2, ..., y_T | \pi, A, B\}$.

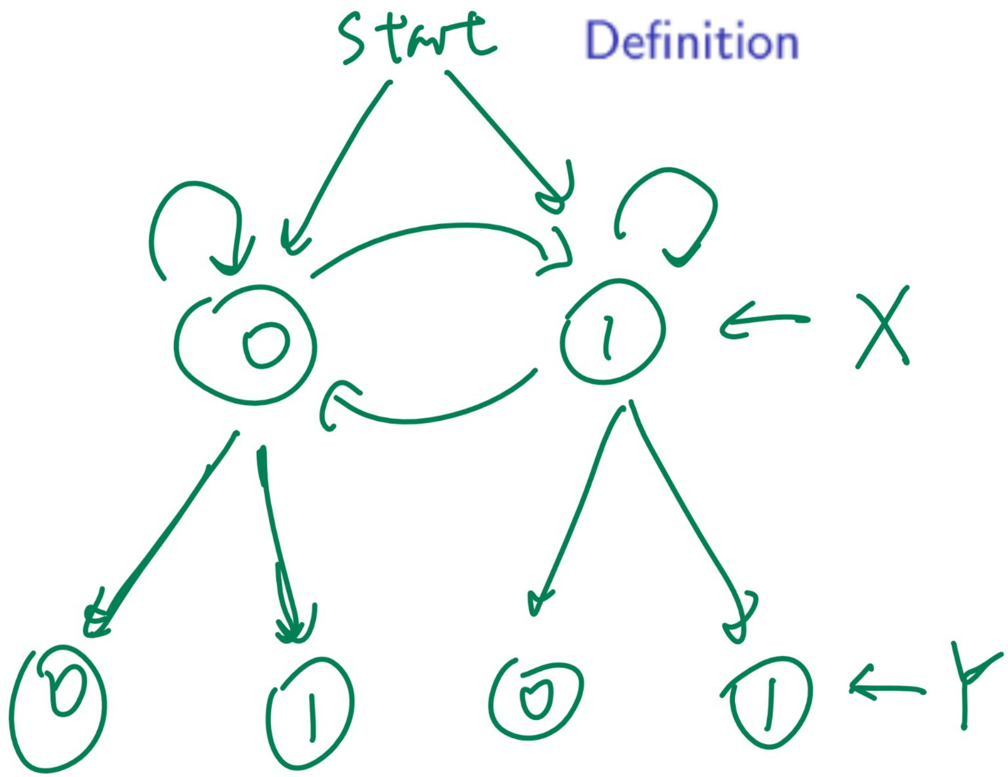
0 0 1 1 -

$$
\begin{aligned}
&\mathbb{P}\{y_1, y_2, ..., y_T | \pi, A, B\} \\
&= \sum_{x_1, x_2, ..., x_T} \mathbb{P}\{y_1, y_2, ..., y_T | x_1, x_2, ..., x_T\} \mathbb{P}\{x_1, x_2, ..., x_T\} \\
&= \sum_{x_1, x_2, ..., x_T} \left(\prod_{t=1}^{\cancel{T}\,2} B_{y_t x_t} \right) \left(\pi_{x_1} \prod_{t=2}^{\cancel{T}\,2} A_{x_{t-1} x_t} \right)
\end{aligned}
$$

- This is also called the Forward Algorithm.

Evaluation Problem Example, Part 1

Definition



$$\Pr\{\underbrace{Y_1 = 0}_{y_1},\ \underbrace{Y_2 = 1}_{y_2} \mid \underline{\pi, A, B}\}$$

$$= \Pr\{\underline{Y_1 = 0,\ Y_2 = 1},\ X_1 = 0,\ X_2 = 0 \mid \pi, A, B\} +$$

$$\begin{array}{cc} X_1 & X_2 \\ 0 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{array} \quad \begin{array}{c} + \\ + \\ + \end{array}$$

Evaluation Problem Example, Part 2

Definition

$$\Pr\{Y_1 = 0, Y_2 = 1, X_1 = 0, X_2 = 0 \mid \pi, A, B\}$$

$$= \underbrace{\Pr\{Y_1 = 0 \mid X_1 = 0\}}_{B_{00}} \cdot \underbrace{\Pr\{Y_2 = 1 \mid X_2 = 0\}}_{B_{10}} \cdot \underbrace{\Pr\{X_2 = 0 \mid X_1 = 0\}}_{A_{00}} \cdot \underbrace{\Pr\{X_1 = 0\}}_{\pi_0}$$

$$\Pr\{Y_1 = a, Y_2 = b, Y_3 = c, X_1 = d, X_2 = e, X_3 = f\}$$

$$= \underbrace{B_{ad} \cdot B_{be} \cdot B_{cf}} \cdot \pi_d \cdot \underbrace{A_{de} \cdot A_{ef}}$$

Evaluation Problem Example, Part 3

Definition

Decoding Problem

Definition

try all combination

- The task is to find $x_1, x_2, ..., x_T$ that maximizes $\mathbb{P}\{x_1, x_2, ..., x_T | y_1, y_2, ..., y_T, \pi, A, B\}$.

$$= \frac{\Pr\{y_1 \ldots y_T \mid x_1 \ldots x_T\}}{\Pr\{y_1 \ldots\}}$$

- Direct computation is too expensive.
- Dynamic programming needs to be used to save computation.
- This is called the Viterbi Algorithm.

577

Viterbi Algorithm Value Function

Definition

- Define the value functions to keep track of the maximum probabilities at each time t and for each state k.

$$
\begin{aligned}
V_{1,k} &= \mathbb{P}\left\{y_1 | X_1 = k\right\} \cdot \mathbb{P}\left\{X_1 = k\right\} = \Pr\{y_1, X_1\} \\
&= B_{y_1 k} \pi_k \\
V_{t,k} &= \max_x \mathbb{P}\left\{y_t | X_t = k\right\} \mathbb{P}\left\{X_t = k | X_{t-1} = x\right\} V_{1,k} \\
&= \max_x B_{y_t k} A_{kx} V_{1,k}
\end{aligned}
$$

$\Pr\{y_t, X_1 \ldots X_t\}$

Viterbi Algorithm Policy Function

Definition

- Define the policy functions to keep track of the x_t that maximizes the value function.

$$\text{policy}_{t,k} = \arg\max_x B_{y_t k} A_{kx} V_{1,k}$$

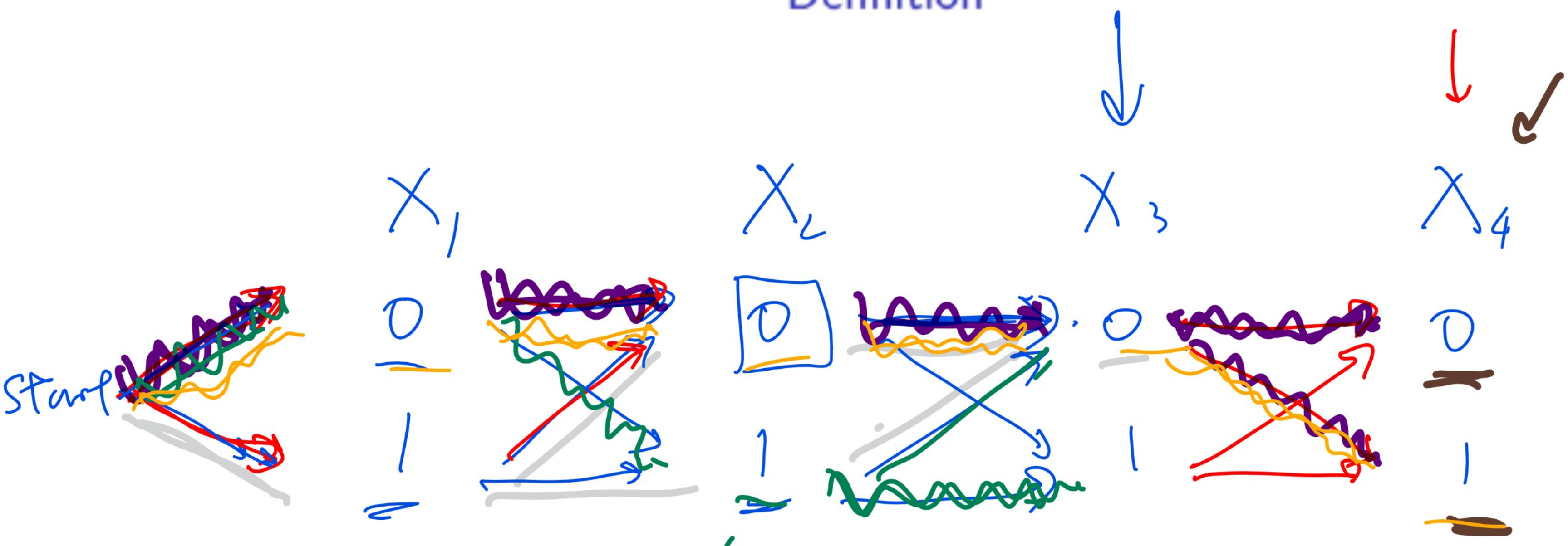
- Given the policy functions, the most probable hidden sequence can be found easily.

$$x_T = \arg\max_x V_{T,x}$$

$$x_t = \text{policy}_{t+1, x_{t+1}}$$

Dynamic Programming Diagram

Definition



V_1 $\quad Pr\{y_1, x_1\}$

$V_{2,1}$ $\quad Pr\{y_2, y_1, x_1, x_2\}$ $\quad V_{3,0}$

Viterbi Algorithm Diagram

Definition

Expectation Maximization Algorithm (for HMM), Part 1

Algorithm

- Initialize the hidden Markov model.

CPT

$$\pi \sim D\left(|X|\right), A \sim D\left(|X|, |X|\right), B \sim D\left(|Y|, |X|\right)$$

- Perform the forward pass.

count $X_t = i, X_{t+1} = j$

$X_t = i$

$$\alpha_{i,t} \text{ represents } \mathbb{P}\left\{y_1, y_2, ..., y_t, X_t = i | \pi, A, B\right\}$$

$$\alpha_{i,1} = \pi_i B_{y_1,i}$$

$\Pr\{y_1, X_1 = i\}$

$y_1, \ldots, y_t$ $\quad y_{t+1}, \ldots y_T$

$$\alpha_{i,t+1} = \sum_{j=1}^{|X|} \alpha_{j,t} A_{ji} B_{y_{t+1} i}$$

$\Pr\{y_1, y_2, X_1 = ?, X_2 = i\}$

$\Pr\{y_1 \ldots y_t, X_1^{=0,1} \ldots X_{t-1}^{=0,1}, X_t^{=j}\}$

Expectation Maximization Algorithm (for HMM), Part 2

Algorithm

- Perform the backward pass.

$$\beta_{i,t} \text{ represents } \mathbb{P}\left\{y_{t+1}, y_{t+2}, ..., y_T | X_t = i, \pi, A, B\right\}$$

$$\beta_{i,T} = 1$$

$\Pr\{1 \mid X_T = i\}$

$$\boxed{\beta_{i,t}} = \sum_{j=1}^{|X|} A_{ij} B_{y_{t+1} j} \beta_{j,t+1}$$

$\Pr\{y_T \mid X_{T-1} = i, \}$

$\Pr\{y_T, X_T \mid X_{T-1} = i\}$

$0, 1$

$\Pr\{y_T, y_{T-1} \mid X_{T-2} = i\}$

X_T, X_{T-1}

j

Expectation Maximization Algorithm (for HMM), Part 3

Algorithm

- Define the conditional hidden state probabilities for each training sequence n.

count $X_t = i$

$$\gamma_{n,i,t} = \text{ represents } \mathbb{P}\{X_t = i | y_1, y_2, ..., y_T, \pi, A, B\}$$

$$\gamma_{n,i,t} = \boxed{\frac{\alpha_{i,t}\beta_{i,t}}{\sum_{j=1}^{|X|} \alpha_{j,t}\beta_{j,t}}}$$

$$\frac{\Pr\{X_t = i, y_1 \ldots y_T\}}{\sum_{j=1}^{|X|} \Pr\{X_t = j, y_1 \ldots y_T\}}$$

$\Pr\{y_1 \ldots y_t, X_t = i\} \Pr\{y_{t+1} \ldots y_T | X_t = i\}$

Expectation Maximization Algorithm (for HMM), Part 4

Algorithm

- Define the conditional hidden state probabilities for each training sequence n.

count $X_t = i, X_{t+1} = j$

$\xi_{n,i,j,t}$ represents $\mathbb{P}\left\{X_t = i, X_{t+1} = j | y_1, y_2, ..., y_T, \pi, A, B\right\}$

$$\xi_{n,i,j,t} = \dfrac{\alpha_{i,t} A_{ij} \beta_{j,t+1} B_{y_{t+1} j}}{\displaystyle\sum_{k=1}^{|X|} \sum_{l=1}^{|X|} \alpha_{k,t} A_{kl} \beta_{l,t+1} B_{y_{t+1} w}}$$

$$\rightarrow \dfrac{\Pr\{X_t = i, X_{t+1} = j, y_1 ... y_t, y_{t+1}, y_{t+2} ... y_T\}}{\Pr\{y_1 ... y_T\}}$$

$\rightarrow \alpha_{i,t} \, A_{ij} \quad \beta_{j,t+1} \, B_{y_{t+1} j}$

Expectation Maximization Algorithm (for HMM), Part 5

Algorithm

- Update the model.

$$\pi_i' = \frac{\sum_{n=1}^{N} \gamma_{n,i,1}}{N} \qquad \frac{\# X_1 = i}{\sim}$$

$$A_{ij}' = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \xi_{n,i,j,t}}{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \gamma_{n,i,t}} \qquad \frac{\# X_t = i, X_{t+1} = j}{\# X_t = i}$$

Expectation Maximization Algorithm (for HMM), Part 6

Algorithm

- Update the model, continued.

$$B'_{ij} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T} \mathbb{1}_{\{y_{n,t}=j\}}\gamma_{n,i,t}}{\sum_{n=1}^{N}\sum_{t=1}^{T}\gamma_{n,i,t}}$$

$$\frac{(\# X_t = i,\ Y_t = j)}{\# X_t = i}$$

- Repeat until π, A, B converge.