CS540 Introduction to Artificial Intelligence
Lecture 23

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

July 5, 2021

Midterm Reivew Session

- June 29 Dan will go through selected Homework questions and Past Exam questions, not recorded, notes will be posted.
- Dandi will go through the same questions this Thursday and Friday (June 18 and 19)$12:30$ to $1:45$ for section 1, you can use the guest link to attend too.

Special Bayesian Network for Sequences

Motivation

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

HMM Applications Part 2

Motivation

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

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

Other HMM Applications

Motivation

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

Hidden Markov Model Diagram

Motivation

Transition and Likelihood Matrices

Motivation

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

1. Initial state vector: π.

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

2. State transition matrix: A.

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

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

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

Markov Property

Motivation

- The Markov property implies the following conditionally independent property.

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

Evaluation and Training

Motivation

- There are three main tasks associated with an HMM.

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

Expectation-Maximization Algorithm

Description

- Start with a random guess of π, A, B.
- Compute the forward probabilities: the joint probability of an observed sequence and its hidden state.
- Compute the backward probabilities: the probability of an 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

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

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

- This is also called the Forward Algorithm.

Evaluation Problem Example, Part 1

Definition

- Fall 2018 Final Q28 and Q29
- Compute $\mathbb{P}\left\{X_4 = Y, X_5 = Z | X_3 = X\right\}$.
- Compute $\mathbb{P}\left\{X_1 = X, X_2 = Z | Y_1 = A, Y_2 = B\right\}$.

Evaluation Problem Example, Part 2

Definition

Evaluation Problem Example, Part 3

Definition

Evaluation Problem Example, Part 4

Definition

Decoding Problem

Definition

- The task is to find $x_1, x_2, ..., x_T$ that maximizes $\mathbb{P}\left\{x_1, x_2, ..., x_T \middle| y_1, y_2, ..., y_T, \pi, A, B\right\}$.
- Direct computation is too expensive.
- Dynamic programming needs to be used to save computation.
- This is called the Viterbi Algorithm.

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.

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

Viterbi Algorithm Policy Function

Definition

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

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

- 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

Viterbi Algorithm Diagram

Definition

Expectation-Maximization Algorithm (for HMM), Part 1

Algorithm

- Initialize the hidden Markov model.

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

- Perform the forward pass.

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

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

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

Expectation-Maximization Algorithm (for HMM), Part 2

Algorithm

- Perform the backward pass.

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

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

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

Expectation-Maximization Algorithm (for HMM), Part 3

Algorithm

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

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

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

Expectation-Maximization Algorithm (for HMM), Part 4

Algorithm

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

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

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

Expectation-Maximization Algorithm (for HMM), Part 5

Algorithm

- Update the model.

$$\pi'_i = \frac{\sum_{n=1}^{N} \gamma_{n,i,1}}{N}$$

$$A'_{ij} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \xi_{n,i,j,t}}{\sum_{n=1}^{N} \sum_{t=1}^{T-1} \gamma_{n,i,t}}$$

Expectation-Maximization Algorithm (for HMM), Part 6

Algorithm

- Update the model, continued.

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

- Repeat until π, A, B converge.