CS540 Introduction to Artificial Intelligence
Lecture 23

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

June 18, 2020

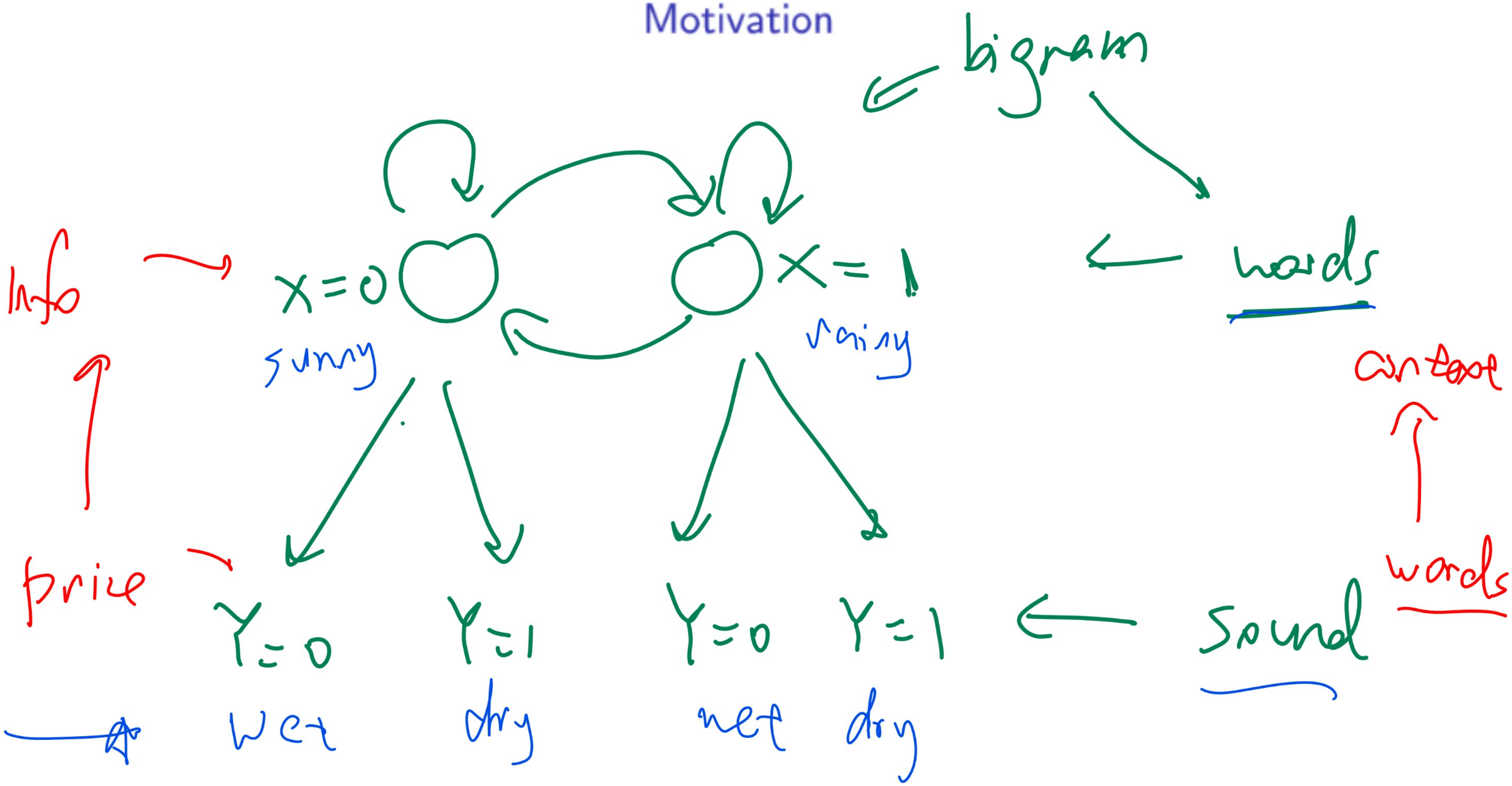
Special Bayesian Network for Sequences

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.

Hidden Markov Model Diagram

Motivation



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: $\pi, A, B, \ldots$

Evaluation Problem

Definition

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

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

- This is also called the Forward Algorithm.

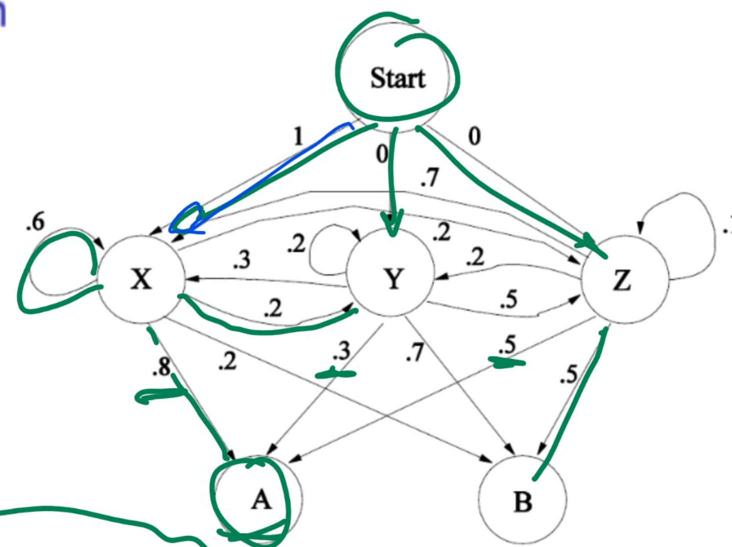
Evaluation Problem Example, Part 1

Definition

training HMM

↳ EM algorithm.

2018?



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

$$\pi = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \begin{matrix} \leftarrow X_1 = X \\ \leftarrow X_2 = Y \\ \leftarrow X_3 = Z \end{matrix} \qquad A = \begin{bmatrix} 0.6 & \boxed{0.2} & 0.2 \\ 0.3 & 0.2 & \boxed{0.5} \\ 0.7 & 0.2 & 0.1 \end{bmatrix}$$

$$B_A = \begin{pmatrix} 0.8 \\ 0.3 \\ 0.5 \end{pmatrix} \longleftrightarrow B_B = \begin{pmatrix} 0.2 \\ 0.7 \\ 0.5 \end{pmatrix} \qquad B = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.5 & 0.5 \end{pmatrix}$$

Evaluation Problem Example, Part 2

Definition

① $\Pr\{X_4 = Y, X_5 = Z \mid X_3 = X\}$

$= \Pr\{X_4 = Y \mid X_3 = X\} \cdot \Pr\{X_5 = Z \mid X_4 = Y\}$

$= 0.2 \cdot 0.5$

② $\Pr\{X_1 = X, X_2 = Z \mid Y_1 = A, Y_2 = B\}$

$= \Pr\{Y_1 = A \mid X_1 = X\} \cdot \Pr\{Y_2 = B \mid X_2 = Z\}$

$\Pr\{X_1 = X\} \cdot \Pr\{X_2 = Z \mid X_1 = X\}$

$\Pr\{Y_1 = A, Y_2 = B\}$

✓

Evaluation Problem Example, Part 3

Definition

$0.8 \cdot 0.5 \cdot 1 \cdot 0.2$

$P\{X_1 = X, X_2 = \boxed{\begin{matrix} X \\ Y \\ Z \end{matrix}}, Y_1 = A, Y_2 = B\}$

HMM

Bayesian

$$\begin{array}{ccccc} X_1 & \longrightarrow & X_2 & \longrightarrow & X_3 \\ \uparrow & & \uparrow & & \uparrow \\ Y_1 & & Y_2 & & Y_3 \end{array}$$

Evaluation Problem Example, Part 4

Definition

$\Pr\{D \mid a, b, c\}$

$\Pr\{a, b, c\}$

$= \Pr\{a\} \Pr\{b\} \cdot \Pr\{c\}$

unigram count.

$\boxed{\Pr\{a, b, c \mid D\}} \leftarrow \cdot \Pr\{D\}$

$\Pr\{a \mid D\} \cdot \Pr\{b \mid D\} \cdot \Pr\{c \mid D\}$

unigram

Decoding Problem

Definition

$$\Pr(D \mid a) = \frac{\Pr(a \mid D) \cdot \cancel{\Pr(D)}}{\Pr(a \mid D)\,\cancel{\Pr(D)} + \Pr(a \mid \neg D) \cdot \cancel{\Pr(\neg D)}}$$

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

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

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

Viterbi Algorithm Diagram

Definition

Expectation-Maximization Algorithm (for HMM), Part 1

EM 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.

$$\alpha_{i,t} \text{ represents } \mathbb{P}\left\{y_1, y_2, ..., y_t, X_t = i | \pi, A, B\right\}$$
$$\alpha_{i,1} = \pi_i B_{y_1, i}$$
$$\alpha_{i,t+1} = \sum_{j=1}^{|X|} \alpha_{j,t} A_{ji} B_{y_{t+1} i}$$

Expectation-Maximization Algorithm (for HMM), Part 2

Algorithm

- Perform the backward pass.

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

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

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

Expectation-Maximization Algorithm (for HMM), Part 3

Algorithm

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

C_i

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

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

Expectation-Maximization Algorithm (for HMM), Part 4

Algorithm

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

C_{ij}

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

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

Expectation-Maximization Algorithm (for HMM), Part 5

Algorithm

- Update the model.

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

$$\frac{C_{x_0=i}}{N}$$

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

$$\frac{C_{ij}}{C_i}$$

Expectation-Maximization Algorithm (for HMM), Part 6

Algorithm

- Update the model, continued.

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

$$\frac{C_{Y=j \mid X=i}}{C_{X=i}}$$

- Repeat until π, A, B converge.