CS540 Introduction to Artificial Intelligence
Lecture 11

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

July 7, 2021

Guess the Box Office, Prior

Admin

Q 1

- Guess which one of the following movies will have the highest US box office?
- A: Fast and Furious 9 (June 25)
- B: Black Widow (July 9)
- C: Space Jam: A New Legacy (July 16)
- D: Jungle Cruise (July 30)
- E: Other (only movies opening before August 1, i.e. exclude The Suicide Squad)

Guess the Box Office, Likelihood

Admin

Q2

- Which ones of the following movies are you planning to see (or have seen) in the theaters? You do not have to select any.
- A: Fast and Furious 9 (June 25, 120+ million so far)
- B: Black Widow (July 9, opening tomorrow)
- C: Space Jam: A New Legacy (July 16)
- D: Jungle Cruise (July 30)
- E: Other (only movies opening before August 1)

Guess the Box Office, Posterior

Admin

Q3

- Guess again which one of the following movies will have the highest US box office?
- A: Fast and Furious 9 (June 25)
- B: Black Widow (July 9)
- C: Space Jam: A New Legacy (July 16)
- D: Jungle Cruise (July 30)
- E: Other (only movies opening before August 1)

Remind Me to Start Recording

Admin

- The messages you send in chat will be recorded: you can change your Zoom name (movie or TV show characters, animals or plants, nothing political or offensive please) now before I start recording.

N Gram Model

Algorithm

- Input: series $\{z_1, z_2, ..., z_{d_i}\}_{i=1}^{n}$.
- Output: transition probabilities $\hat{\mathbb{P}}\{z_t | z_{t-1}, z_{t-2}, ..., z_{t-N+1}\}$ for all $z_t = 1, 2, ..., m$.
- Compute the transition probabilities using counts and Laplace smoothing.

$$\hat{\mathbb{P}}\{z_t | z_{t-1}, z_{t-2}, ..., z_{t-N+1}\} = \frac{C_{z_{t-N+1}, z_{t-N+2}, ..., z_t} + 1}{C_{z_{t-N+1}, z_{t-N+2}, ..., z_{t-1}} + m}$$

Markov Chain Demo

Motivation

Smoothing Example

Quiz

$$\frac{c_i + 1}{\sum_{i=1}^{m} (c_i + 1)} = \frac{1}{m}$$

- Fall 2018 Midterm Q12.
- Given a vocabulary of 10^6, a document with 10^{12} tokens with $c_{\text{zoodles}} = 3$. What is the MLE estimation of $\mathbb{P}\{\text{zoodles}\}$ with and without Laplace smoothing?

# of unique words

# of tokens

word

$$\hat{Pr}\{\text{zoodles}\} = \frac{c_z}{c_{\text{total}}} = \frac{3}{10^{12}}$$

Smoothing

$$\frac{c_z + 1}{c_{\text{total}} + m} \qquad \frac{3+1}{10^{12} + 10^6}$$

Smoothing Example 2

Quiz

Q4

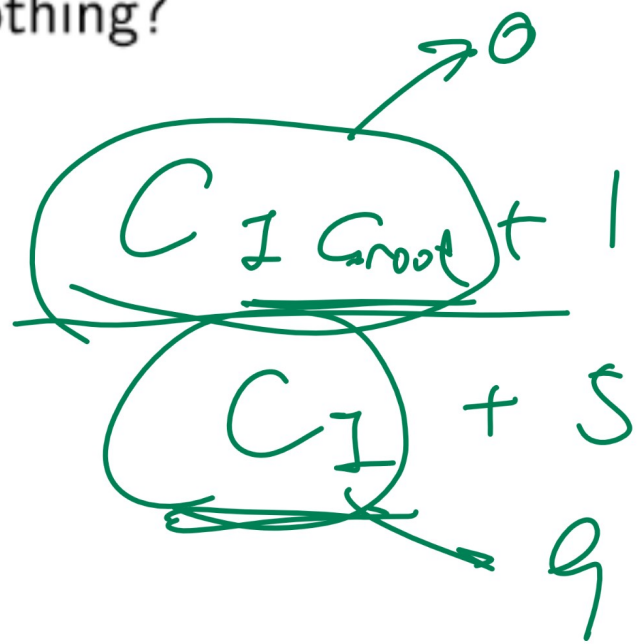
- Given the training instance with 9 "I am Groot" followed by 1 "We are Groot", what is the MLE estimation of $\mathbb{P}\{\text{Groot}\}$ with Laplace smoothing?
- A: $\frac{1}{2}$
- B: $\frac{11}{35}$
- C: $\frac{1}{3}$
- D: $\frac{11}{31}$
- E: $\frac{1}{4}$

$$P_r\{\text{Groot}\} = \frac{C_{\text{Groot}} + 1}{C_{\text{total}} + 5} = \frac{10+1}{30+5}$$

Smoothing Example 3

Quiz

- Given the training instance with 9 "I am Groot" followed by 1 "We are Groot", what is the MLE estimation of $\mathbb{P}\{\text{Groot} \mid \text{I}\}$ with Laplace smoothing?



- A: $\frac{1}{10}$
- B: $\frac{1}{11}$
- C: $\frac{1}{14}$
- D: $\frac{1}{15}$
- E: 0

Sampling from Discrete Distribution

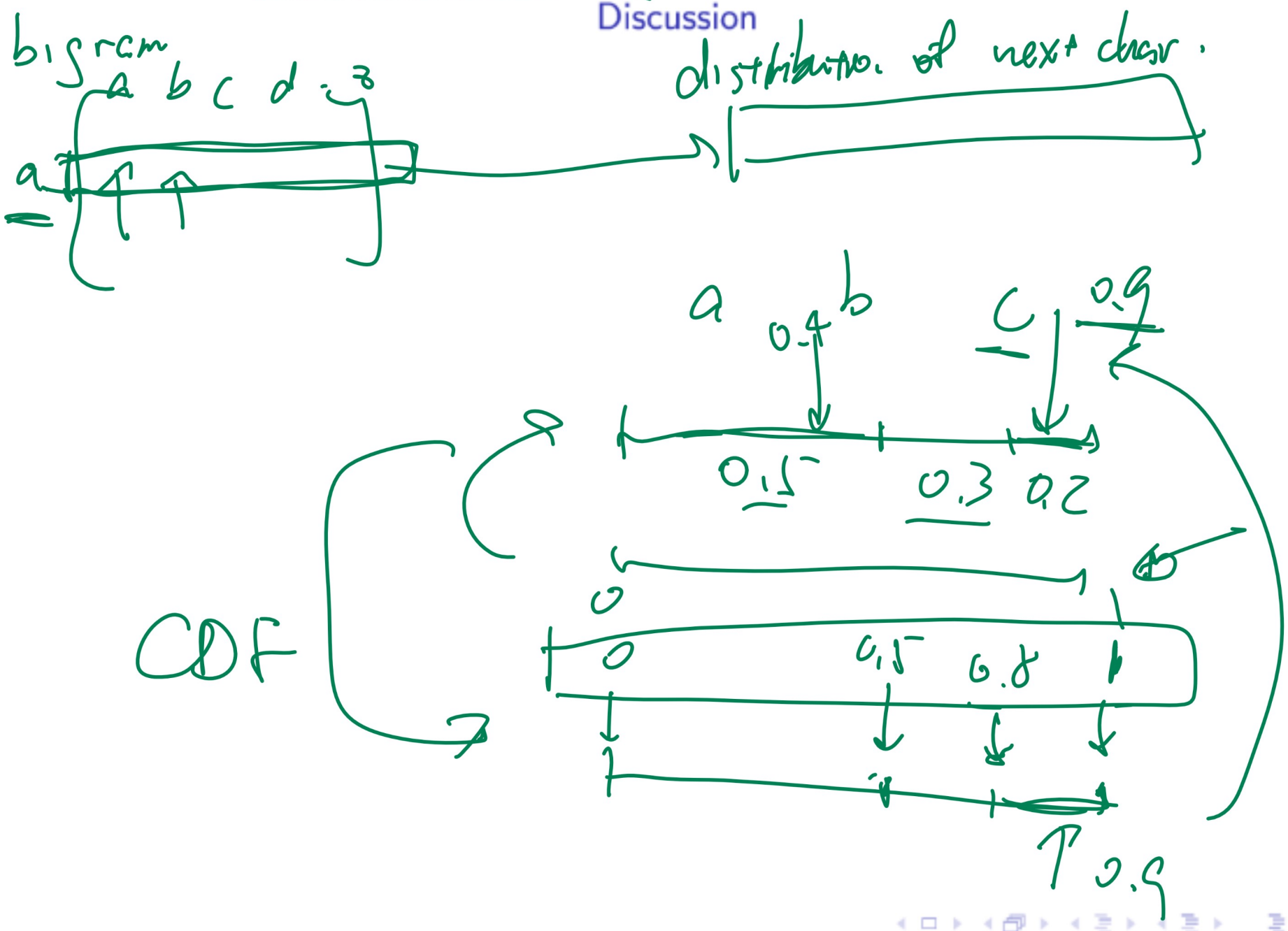
Discussion

P3

- To generate new sentences given an N gram model, random realizations need to be generated given the conditional probability distribution.
- Given the first $N-1$ words, $z_1, z_2, ..., z_{N-1}$, the distribution of next word is approximated by $p_x = \hat{\mathbb{P}}\{z_N = x | z_{N-1}, z_{N-2}, ..., z_1\}$. This process then can be repeated for on $z_2, z_3, ..., z_{N-1}, z_N$ and so on.

CDF Inversion Method Diagram

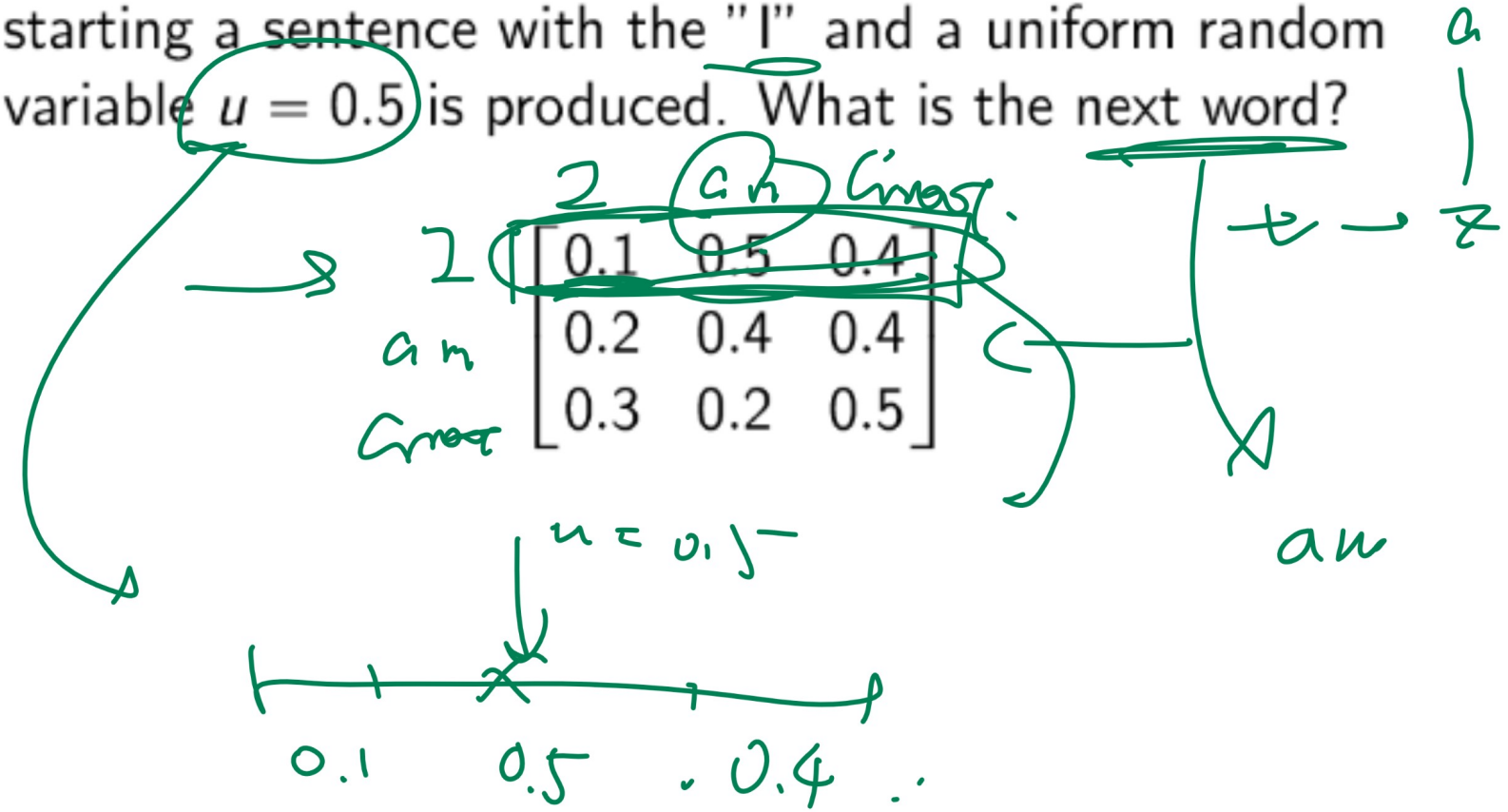
Discussion



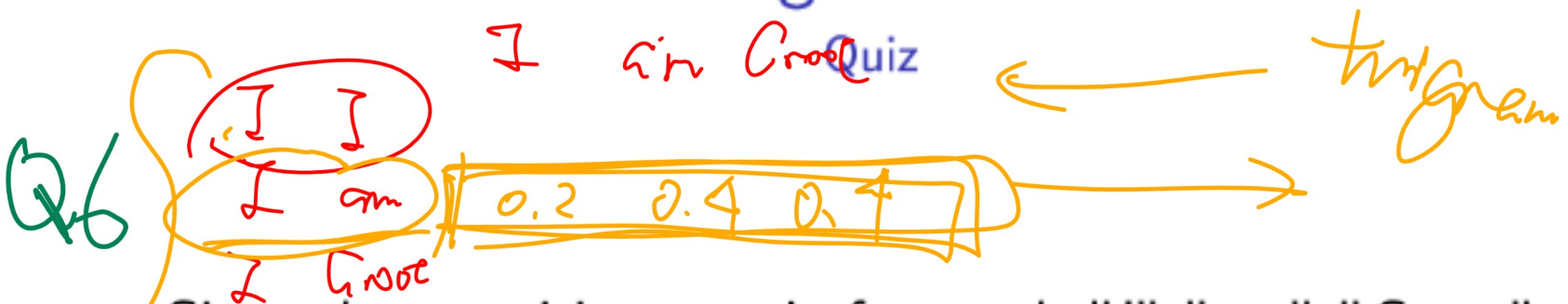
Generating New Words 1

Quiz

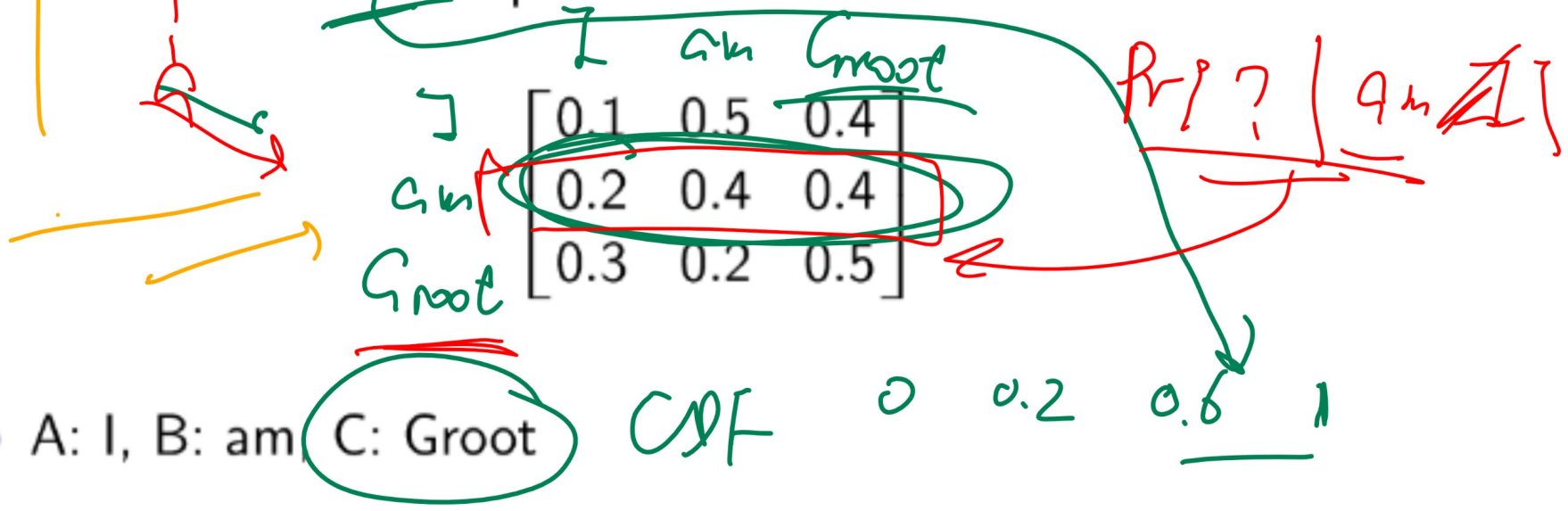
- Given the transition matrix for words "I" "am" "Groot", starting a sentence with the "I" and a uniform random variable $u = 0.5$ is produced. What is the next word?



Generating New Words 2



- Given the transition matrix for words "I" "am" "Groot", starting a sentence with the "I am" and a uniform random variable $u = 0.75$ is produced. What is the next word?



- A: I, B: am, C: Groot

Bayes Rule Example 1

Quiz

- Two documents A and B. Suppose A contains 1 "Groot" and 9 other words, and B contains 8 "Groot" and 2 other words. One document is taken out at random (with equal probability), and one word is picked out at random (all words with equal probability). The word is "Groot". What is the probability that the document is A?
- A: $\frac{1}{2}$, B: $\frac{1}{3}$, C: $\frac{1}{4}$, D: $\frac{1}{8}$, E: $\frac{1}{9}$

Bayes Rule Example 1 Distribution

Quiz

Bayes Rule Example 2

Quiz

- Two documents A and B. Suppose A contains 1 Groot and 9 other words, and B contains 8 "Groot" and 2 other words. One document is taken out A with probably $\frac{2}{3}$ and B with probably $\frac{1}{3}$, and one word is picked out at random with equal probabilities. The word is "Groot". What is the probability that the document is A?
- A: $\frac{1}{5}$, B: $\frac{1}{9}$, C: $\frac{1}{10}$, D: $\frac{1}{16}$, E: $\frac{1}{17}$

Bayes Rule Example 2 Distribution

Quiz

Bayesian Network

Definition

- A Bayesian network is a directed acyclic graph (DAG) and a set of conditional probability distributions.
- Each vertex represents a feature X_j.
- Each edge from X_j to $X_{j'}$ represents that X_j directly influences $X_{j'}$.
- No edge between X_j and $X_{j'}$ implies independence or conditional independence between the two features.

Training Bayes Net

Definition

- Training a Bayesian network given the DAG is estimating the conditional probabilities. Let $P(X_j)$ denote the parents of the vertex X_j, and $p(X_j)$ be realizations (possible values) of $P(X_j)$.

$$\mathbb{P}\{x_j | p(X_j)\}, p(X_j) \in P(X_j)$$

- It can be done by maximum likelihood estimation given a training set.

$$\hat{\mathbb{P}}\{x_j | p(X_j)\} = \frac{c_{x_j, p(X_j)}}{c_{p(X_j)}}$$

Bayesian Network Diagram

Quiz

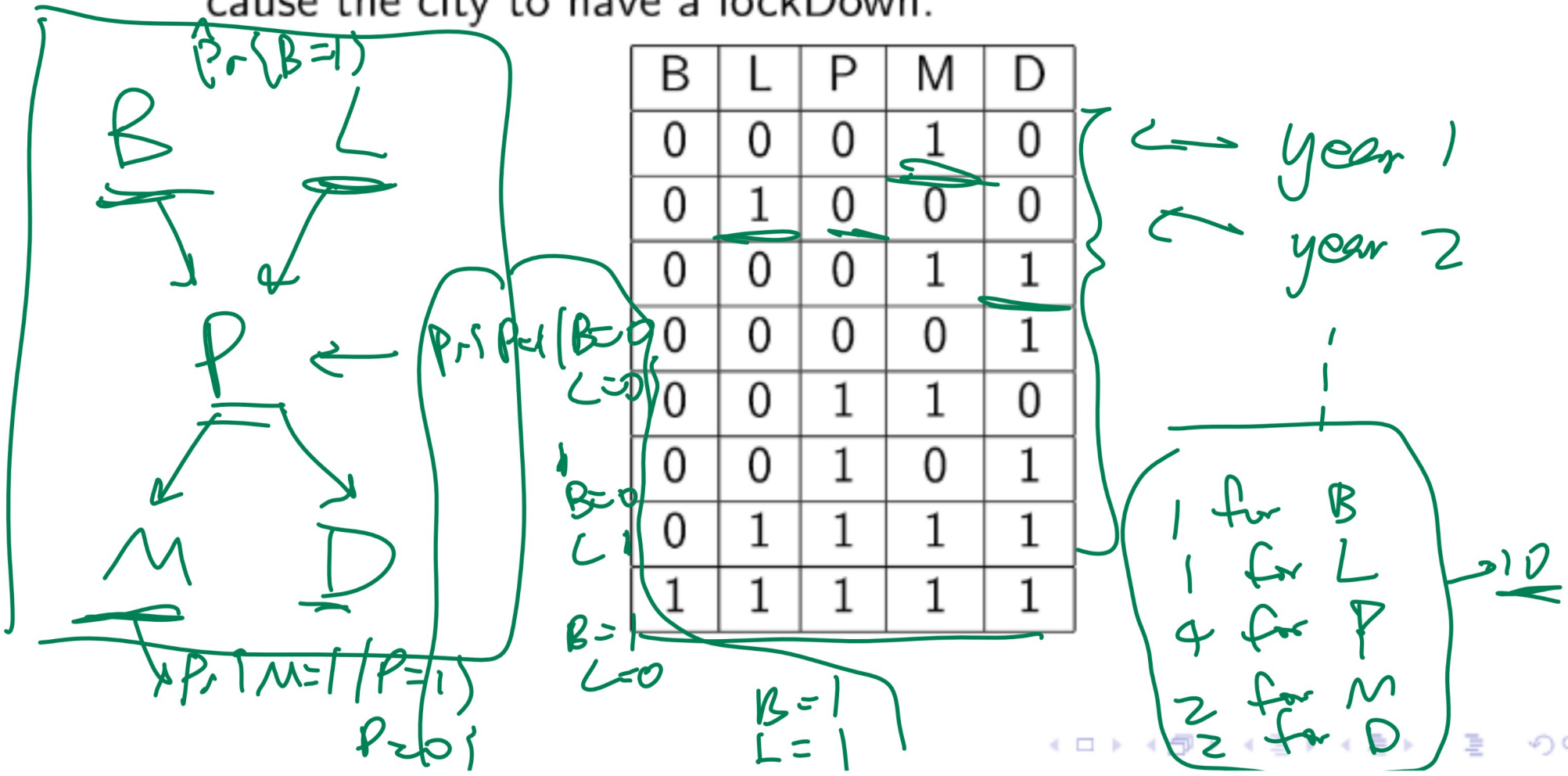
- Story: You are traveling far from home. There may be a Fire problem or a Cat problem at home. Either problem might trigger an Alarm. Then your neighbors Nick (Fury) or Happy or both might call you because of the alarm or for other reasons.

F	C	A	H	N
0	0	0	1	0
0	1	0	0	0
0	0	0	1	1
1	0	0	0	1
0	0	1	1	0
0	0	1	0	1
0	0	1	1	1
0	0	1	1	1

Bayesian Network Diagram

Quiz

- Story: There may be Bat virus or a Lab leak that result in a Pandemic, which in turn can cause people to wear Masks or cause the city to have a lockDown.



B	L	P	M	D
0	0	0	1	0
0	1	0	0	0
0	0	0	1	1
0	0	0	0	1
0	0	1	1	0
0	0	1	0	1
0	1	1	1	1
1	1	1	1	1

Bayesian Network Diagram CPTs

Quiz

Bayes Net Training Example, Training 1

Quiz

- Compute $\hat{\mathbb{P}}\{L = 1\}$.

B	L	P	M	D
0	0	0	1	0
0	1	0	0	0
0	0	0	1	1
0	0	0	0	1
0	0	1	1	0
0	0	1	0	1
0	1	1	1	1
1	1	1	1	1

MLE $\frac{C_{L=1}}{C_{total}} = \frac{3+1}{8+5} = \frac{4}{13}$

8 instances.

Bayes Net Training Example, Training 2

Quiz

- Compute $\hat{\mathbb{P}}\{M = 1 | P = 1\}$

B	L	P	M	D
0	0	0	1	0
0	1	0	0	0
0	0	0	1	1
0	0	0	0	1
0	0	1	1	0
0	0	1	0	1
0	1	1	1	1
1	1	1	1	1

Bayes Net Training Example, Training 3

Quiz

- Compute $\hat{\mathbb{P}}\{P = 1 | B = 0, L = 1\}$.
- A: 0 , B: $\frac{1}{3}$, C: $\frac{1}{2}$, D: $\frac{2}{3}$, E: 1

B	L	P	M	D
0	0	0	1	0
0	1	0	0	0
0	0	0	1	1
0	0	0	0	1
0	0	1	1	0
0	0	1	0	1
0	1	1	1	1
1	1	1	1	1

Bayes Net Training Example, Training 4

Quiz

- Compute $\hat{\mathbb{P}}\{P = 1 | B = 1, L = 0\}$.
- A: 0 , B: $\frac{1}{3}$, C: $\frac{1}{2}$, D: $\frac{2}{3}$, E: 1

Q8

no smoothing

B	L	P	M	D
0	0	0	1	0
0	1	0	0	0
0	0	0	1	1
0	0	0	0	1
0	0	1	1	0
0	0	1	0	1
0	1	1	1	1
1	1	1	1	1

0 / 1

0 / 2

if smoothing

Laplace Smoothing

Definition

- Recall that the MLE estimation can incorporate Laplace smoothing.

$$\hat{\mathbb{P}}\{x_j|p(X_j)\} = \frac{c_{x_j,p(X_j)} + \lambda}{c_{p(X_j)} + |X_j|\lambda}$$

- Here, $|X_j|$ is the number of possible values (number of categories) of X_j.
- Laplace smoothing is considered regularization for Bayesian networks because it avoids overfitting the training data.

Bayes Net Inference 1

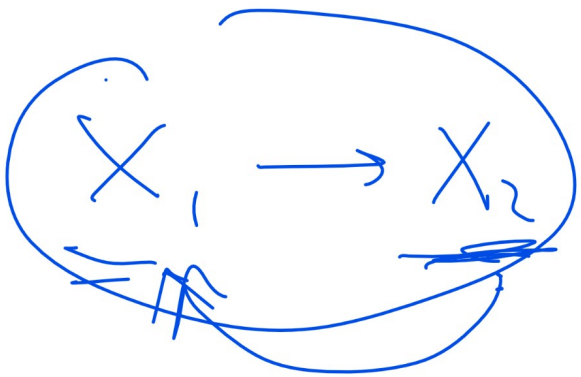
Definition

- Given the conditional probability table, the joint probabilities can be calculated using conditional independence.

$$\mathbb{P}\{x_1, x_2, \ldots, x_m\} = \prod_{j=1}^{m} \mathbb{P}\{x_j | x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m\}$$

$$= \prod_{j=1}^{m} \mathbb{P}\{x_j | p(X_j)\}$$

parents



$$Pr\{X_1 | X_2\} = \frac{Pr\{X_1, X_2\}}{Pr\{X_2\}}$$

$$Pr\{X_2 | X_1\} \cdot Pr\{X_1\}$$

Bayes Net Inference 2

Definition

- Given the joint probabilities, all other marginal and conditional probabilities can be calculated using their definitions.

$$\mathbb{P}\left\{x_j | x_{j'}, x_{j''}, \ldots\right\} = \frac{\mathbb{P}\left\{x_j, x_{j'}, x_{j''}, \ldots\right\}}{\mathbb{P}\left\{x_{j'}, x_{j''}, \ldots\right\}}$$

$$\mathbb{P}\left\{x_j, x_{j'}, x_{j''}, \ldots\right\} = \sum_{x_k : k \neq j, j', j'', \ldots} \mathbb{P}\left\{x_1, x_2, \ldots, x_m\right\}$$

$$\mathbb{P}\left\{x_{j'}, x_{j''}, \ldots\right\} = \sum_{x_k : k \neq j', j'', \ldots} \mathbb{P}\left\{x_1, x_2, \ldots, x_m\right\}$$

Bayes Net Inference Example 1

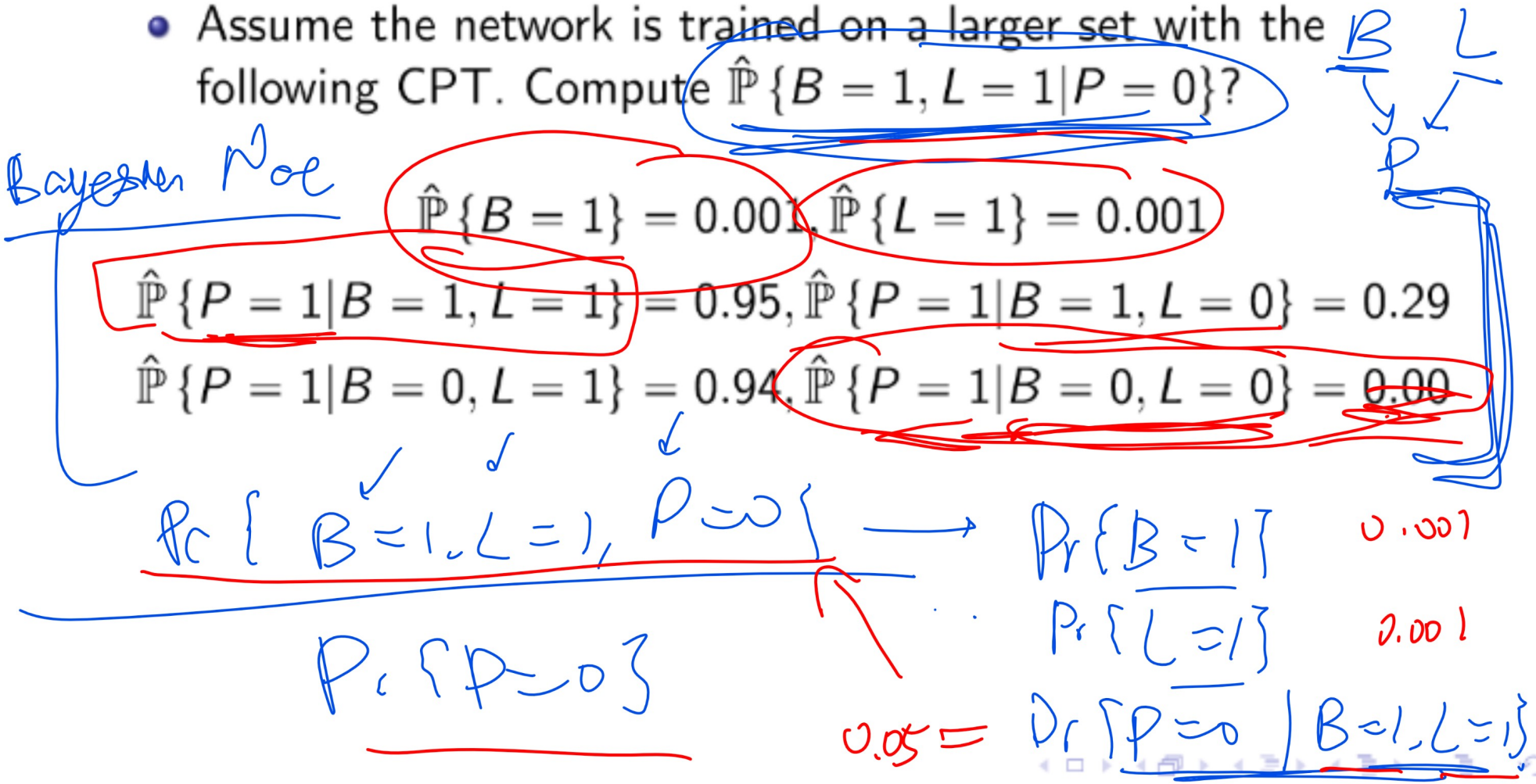
Quiz

- Assume the network is trained on a larger set with the following CPT. Compute $\hat{\mathbb{P}}\{B = 1, L = 1|P = 0\}$?

$$\hat{\mathbb{P}}\{B = 1\} = 0.001, \hat{\mathbb{P}}\{L = 1\} = 0.001$$

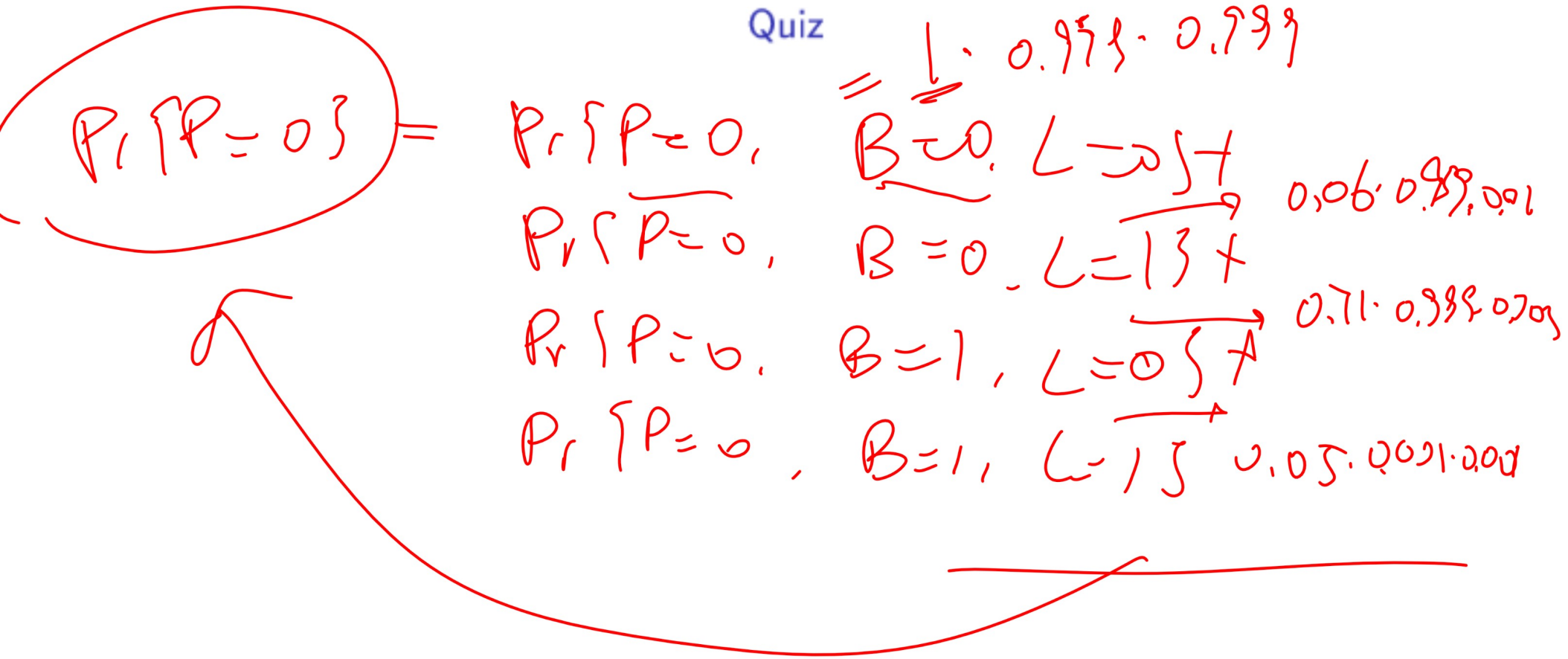
$$\hat{\mathbb{P}}\{P = 1|B = 1, L = 1\} = 0.95, \hat{\mathbb{P}}\{P = 1|B = 1, L = 0\} = 0.29$$

$$\hat{\mathbb{P}}\{P = 1|B = 0, L = 1\} = 0.94, \hat{\mathbb{P}}\{P = 1|B = 0, L = 0\} = 0.00$$



Bayes Net Inference Example 1 Computation 1

Quiz



Bayes Net Inference Example 1 Computation 2

Quiz

Bayes Net Inference Example 2

Quiz

- Compute $\hat{\mathbb{P}}\{B = 1 | L = 1\}$?

$$\hat{\mathbb{P}}\{B = 1\} = 0.001, \hat{\mathbb{P}}\{L = 1\} = 0.001$$

$$\hat{\mathbb{P}}\{P = 1 | B = 1, L = 1\} = 0.95, \hat{\mathbb{P}}\{P = 1 | B = 1, L = 0\} = 0.29$$

$$\hat{\mathbb{P}}\{P = 1 | B = 0, L = 1\} = 0.94, \hat{\mathbb{P}}\{P = 1 | B = 0, L = 0\} = 0.00$$

- A: 0, B: 0.001, C: 0.0094, D: 0.0095, E: 1

Bayes Net Inference Example 3

Quiz

- Compute $\hat{\mathbb{P}}\{B = 1, L = 1 | P = 1\}$?

$$\hat{\mathbb{P}}\{B = 1\} = 0.001, \hat{\mathbb{P}}\{L = 1\} = 0.001$$

$$\hat{\mathbb{P}}\{P = 1 | B = 1, L = 1\} = 0.95, \hat{\mathbb{P}}\{P = 1 | B = 1, L = 0\} = 0.29$$

$$\hat{\mathbb{P}}\{P = 1 | B = 0, L = 1\} = 0.94, \hat{\mathbb{P}}\{P = 1 | B = 0, L = 0\} = 0.00$$

- A: $0.001 \cdot 0.001$, B: $0.001 \cdot 0.001 \cdot 0.95$,
- C: $\dfrac{0.001}{0.001 \cdot 0.95 + 0.999 \cdot (0.94 + 0.29)}$
- D : $\dfrac{0.001 \cdot 0.001}{0.001 \cdot 0.95 + 0.999 \cdot (0.94 + 0.29)}$
- E : $\dfrac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot (0.94 + 0.29)}$

Network Structure

Discussion

- Selecting from all possible structures (DAGs) is too difficult.
- Usually, a Bayesian network is learned with a tree structure.
- Choose the tree that maximizes the likelihood of the training data.

Chow Liu Algorithm

Discussion

- Add an edge between features X_j and $X_{j'}$ with edge weight equal to the information gain of X_j given $X_{j'}$ for all pairs j, j'.
- Find the maximum spanning tree given these edges. The spanning tree is used as the structure of the Bayesian network.

Classification Problem

Discussion

- Bayesian networks do not have a clear separation of the label Y and the features $X_1, X_2, ..., X_m$.
- The Bayesian network with a tree structure and Y as the root and $X_1, X_2, ..., X_m$ as the leaves is called the Naive Bayes classifier.
- Bayes rules is used to compute $\mathbb{P}\{Y = y|X = x\}$, and the prediction $\hat{y}$ is y that maximizes the conditional probability.

$$\hat{y}_i = \arg\max_y \mathbb{P}\{Y = y|X = x_i\}$$

Naive Bayes Diagram

Discussion

Tree Augmented Network Algorithm

Discussion

- It is also possible to create a Bayesian network with all features $X_1, X_2, ..., X_m$ connected to Y (Naive Bayes edges) and the features themselves form a network, usually a tree (MST edges).
- Information gain is replaced by conditional information gain (conditional on Y) when finding the maximum spanning tree.
- This algorithm is called TAN: Tree Augmented Network.

Tree Augmented Network Algorithm Diagram

Discussion