CS540 Introduction to Artificial Intelligence
Lecture 10

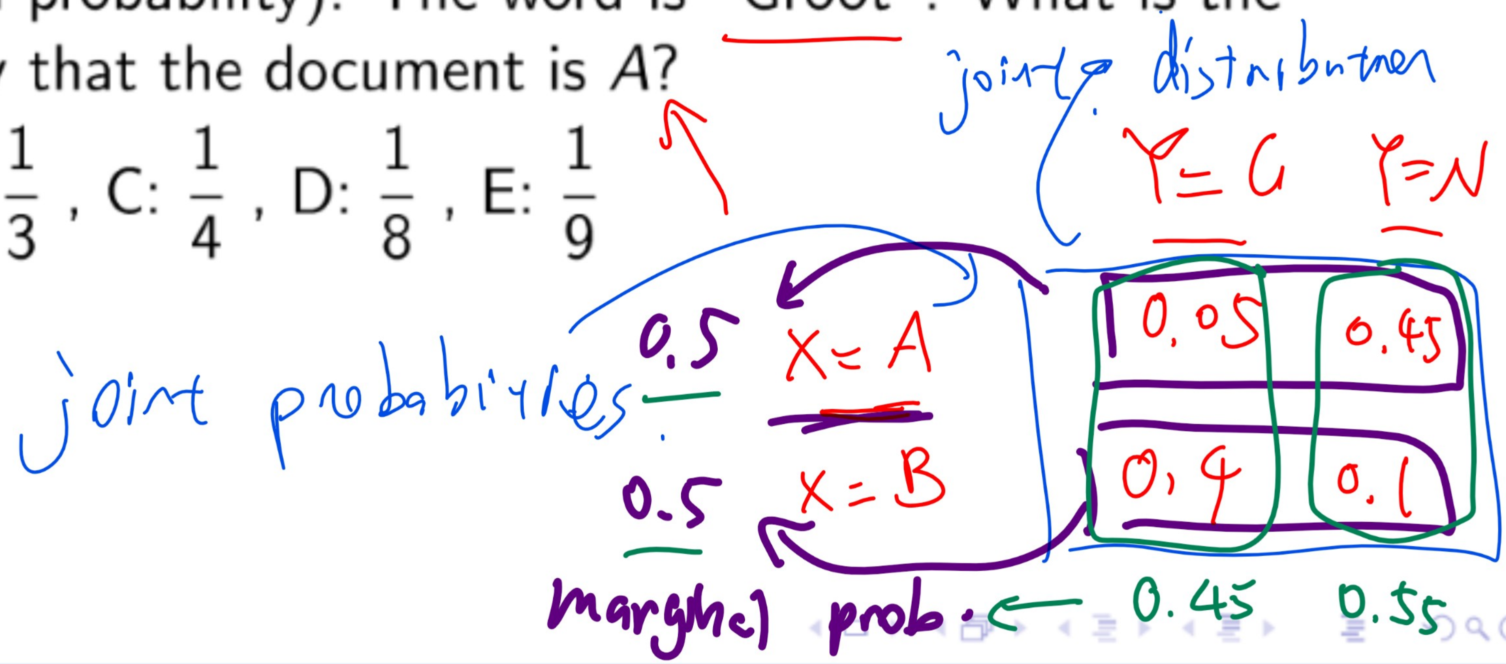
Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 15, 2020

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

$$Pr\{G \mid A\} = 0.1 = \frac{Pr\{G, A\}}{Pr\{A\}} = \frac{0.05}{0.5} = 0.1$$

$$Pr\{A \mid G\} = \frac{1}{9} = \frac{Pr\{A, G\}}{Pr\{G\}} = \frac{0.05}{0.45} = \frac{1}{9}$$

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 $\frac{1}{3}$ and B with probably $\frac{2}{3}$, 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: $\frac{1}{9}$, B: $\frac{1}{10}$, C: $\frac{1}{16}$, D: $\frac{1}{17}$, E: $\frac{1}{25}$

Bayes Rule Example 2 Distribution

Quiz

$$\Pr\{A \mid G\} = \frac{\Pr\{A, G\}}{\Pr\{G\}}$$

$$= \frac{\Pr\{G \mid A\} \cdot \Pr\{A\}}{\Pr\{A, G\} + \Pr\{B, G\}}$$

Bayes Rules.

$$= \frac{\Pr\{G \mid A\} \cdot \Pr\{A\}}{\Pr\{G \mid A\} \cdot \Pr\{A\} + \Pr\{G \mid B\} \cdot \Pr\{B\}}$$

$$= \frac{0.1 \cdot \frac{1}{3}}{0.1 \cdot \frac{1}{3} + 0.8 \cdot \frac{2}{3}} = \frac{1}{1+16} = \frac{1}{17}$$

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 $P(X_j)$.

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

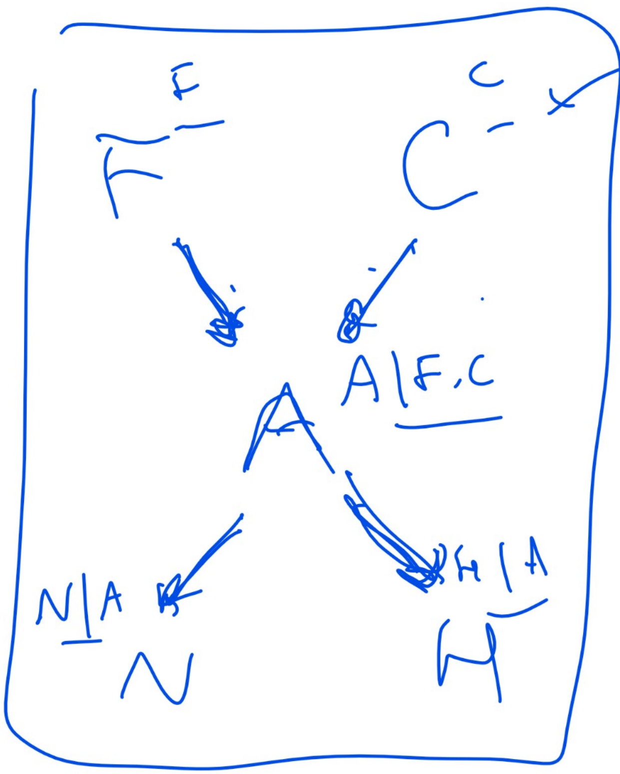
- It can be done by maximum likelihood estimation given a training set.

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

Bayesian Network Diagram

Quiz

- Story: You are travelling 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

Bayes Net Training Example, Training 1

Quiz

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

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

Bayes Net Training Example, Training 2

Quiz

$$\frac{C_{N=1, A=1}}{C_{A=1}} = \frac{3}{4}$$

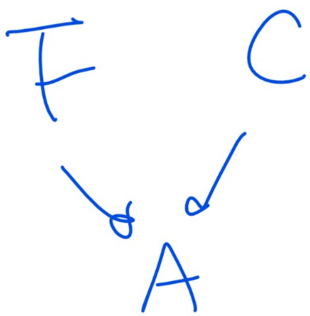
- Compute $\hat{\mathbb{P}}\{N = 1 | A = 1\}$.

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

Bayes Net Training Example, Training 3

Quiz

- Compute $\hat{\mathbb{P}}\left\{A = 1 | F = 0, C = 1\right\}$.



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

$$\frac{C_{A=1, F=0, C=1}}{C_{F=0, C=1}} = \frac{0}{1} = 0$$

Bayes Net Training Example, Training 4

Quiz

- What is the conditional probability $\hat{\mathbb{P}}\left\{A = 1 | F = 1, C = 0\right\}$?
- A: 0, B: $\frac{1}{3}$, C: $\frac{1}{2}$, D: $\frac{2}{3}$, E: 1

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

Bayes Net Training Example, Training 5

Quiz

- What is the conditional probability $\hat{\mathbb{P}}\{A = 0 | F = 0, C = 1\}$?
- A: 0 , B: $\frac{1}{3}$, C: $\frac{1}{2}$, D: $\frac{2}{3}$, E: 1

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

Q10

Laplace Smoothing

Definition

- Recall that the MLE estimation can incorporate Laplace smoothing.

$$\hat{\mathbb{P}}\left\{x_j | p\left(X_j\right)\right\} = \frac{c_{x_j, p(X_j)} + 1}{c_{p(X_j)} + |X_j|}$$

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

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

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

Bayes Net Inference 2

Definition

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

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

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

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

Bayes Net Inference Example 1

Quiz

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

$$\hat{\mathbb{P}}\{F = 1\} = 0.001, \hat{\mathbb{P}}\{C = 1\} = 0.001$$

$$\hat{\mathbb{P}}\{A = 1 | F = 1, C = 1\} = 0.95, \hat{\mathbb{P}}\{A = 1 | F = 1, C = 0\} = 0.94$$

$$\hat{\mathbb{P}}\{A = 1 | F = 0, C = 1\} = 0.29, \hat{\mathbb{P}}\{A = 1 | F = 0, C = 0\} = 0.00$$

$$\hat{\mathbb{P}}\{H = 1 | A = 1\} = 0.9, \hat{\mathbb{P}}\{H = 1 | A = 0\} = 0.05$$

$$\hat{\mathbb{P}}\{N = 1 | A = 1\} = 0.7, \hat{\mathbb{P}}\{N = 1 | A = 0\} = 0.01$$

10

Bayes Net Inference Example 1 Computation 1

Quiz

$\Pr\{F=1, C=1 \mid H=0, N=0\}$

$$= \frac{n = \Pr\{F=1, C=1, H=0, N=0\}}{\Pr\{H=0, N=0\}}$$

marginalization

$n = \Pr\{F=1, C=1, A=0, H=0, N=0\}$

$+ \Pr\{F=1, C=1, A=1, H=0, N=0\}$

$0.001 \quad 0.001 \quad 0.95 \quad (1-0.9) \quad (1-0.7)$

$\Pr\{F=1\} \cdot \Pr\{C=1\} \cdot \Pr\{A=0 \mid F=1, C=1\}$

$\Pr\{H=0 \mid A=0\} \cdot \Pr\{N=0 \mid A=0\}.$

$= 0.001 \cdot 0.001 \cdot (1-0.95)(1-0.05)(1-0.01)$

Bayes Net Inference Example 1 Computation 2

Quiz

$$\Pr\{N=0, H=0\} = \Pr\{F, C, A = 0,0,0, N, H = 0\}$$

0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

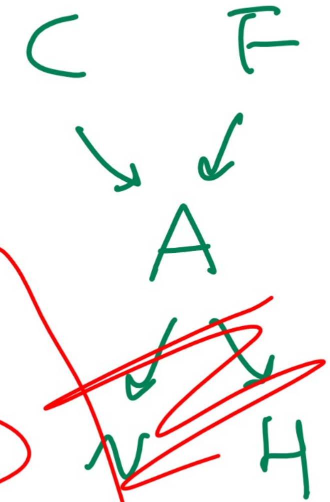
Bayes Net Inference Example 1 Computation 3

Quiz

Bayes Net Inference Example 2

Quiz

Q11



- Compute $\hat{\mathbb{P}}\{C = 1 | F = 1\}$? $= \frac{P_r\{C=1, F=1\}}{P_r\{F=1\}}$

$$\hat{\mathbb{P}}\{F\} = 0.001, \hat{\mathbb{P}}\{C\} = 0.001$$
$$\hat{\mathbb{P}}\{A|F, C\} = 0.95, \hat{\mathbb{P}}\{A|F, \neg C\} = 0.94$$
$$\hat{\mathbb{P}}\{A|\neg F, C\} = 0.29, \hat{\mathbb{P}}\{A|\neg F, \neg C\} = 0.00$$

$\frac{P_r\{C=1\} \cdot P_r\{F=1\}}{P_r\{F=1\}}$

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

Bayes Net Inference Example 2 Computation

Quiz

- Compute $\hat{\mathbb{P}}\{C = 1 | F = 1\}$?

$$\hat{\mathbb{P}}\{F\} = 0.001, \hat{\mathbb{P}}\{C\} = 0.001$$

$$\hat{\mathbb{P}}\{A | F, C\} = 0.95, \hat{\mathbb{P}}\{A | F, \neg C\} = 0.94$$

$$\hat{\mathbb{P}}\{A | \neg F, C\} = 0.29, \hat{\mathbb{P}}\{A | \neg F, \neg C\} = 0.00$$

$$Pr\{C=1 | A=1\} = Pr\{C=1\} \cdot Pr\{A=1 | C=1, F=?\}$$

$$Pr\{C=1, A=1, F=1\} + Pr\{C=1, A=1, F=0\}$$

Bayes Net Inference Example 3

Quiz

- Compute $\hat{\mathbb{P}}\{C = 1, F = 1 | A = 1\}$?

$$\hat{\mathbb{P}}\{F\} = 0.001, \hat{\mathbb{P}}\{C\} = 0.001$$
$$\hat{\mathbb{P}}\{A|F, C\} = 0.95, \hat{\mathbb{P}}\{A|F, \neg C\} = 0.94$$
$$\hat{\mathbb{P}}\{A|\neg F, C\} = 0.29, \hat{\mathbb{P}}\{A|\neg F, \neg C\} = 0.00$$

- 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 : $\dfrac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot (0.94 + 0.29)}$

Q12 (last)

$$\frac{Pr\{C=1, F=1, A=1\}}{Pr\{A=1\}}$$

Bayes Net Inference Example 3 Computation

Quiz

- Compute $\hat{\mathbb{P}}\left\{C = 1, F = 1 | A = 1\right\}$?

$$\hat{\mathbb{P}}\left\{F\right\} = 0.001, \hat{\mathbb{P}}\left\{C\right\} = 0.001$$

$$\hat{\mathbb{P}}\left\{A | F, C\right\} = 0.95, \hat{\mathbb{P}}\left\{A | F, \neg C\right\} = 0.94$$

$$\hat{\mathbb{P}}\left\{A | \neg F, C\right\} = 0.29, \hat{\mathbb{P}}\left\{A | \neg F, \neg C\right\} = 0.00$$

$$\frac{CFA}{CFA + \neg CFA + C\neg FA + \neg C\neg FA} = \frac{0.001 \cdot 0.001 \cdot 0.95}{0.001 \cdot 0.001 \cdot 0.95 + 0.999 \cdot 0.001 \cdot 0.94 + 0.001 \cdot 0.999 \cdot 0.29 + 0}$$

Chow Liu Algorithm

Discussion

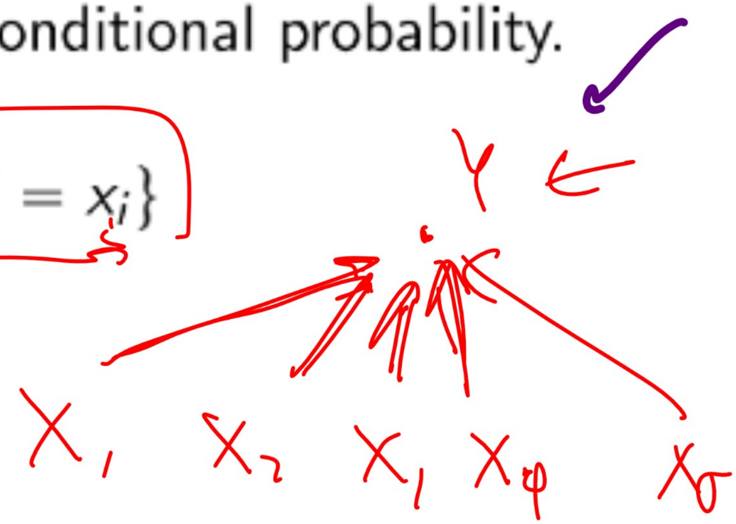
- Add an edge between features X_j and $X_{j'}$ with edge weight equal to the information gain of X_j given $X_{j'}$ for all pairs j, j'.
- Find the maximum spanning tree given these edges. The spanning tree is used as the structure of the Bayesian network.

Classification Problem

Discussion

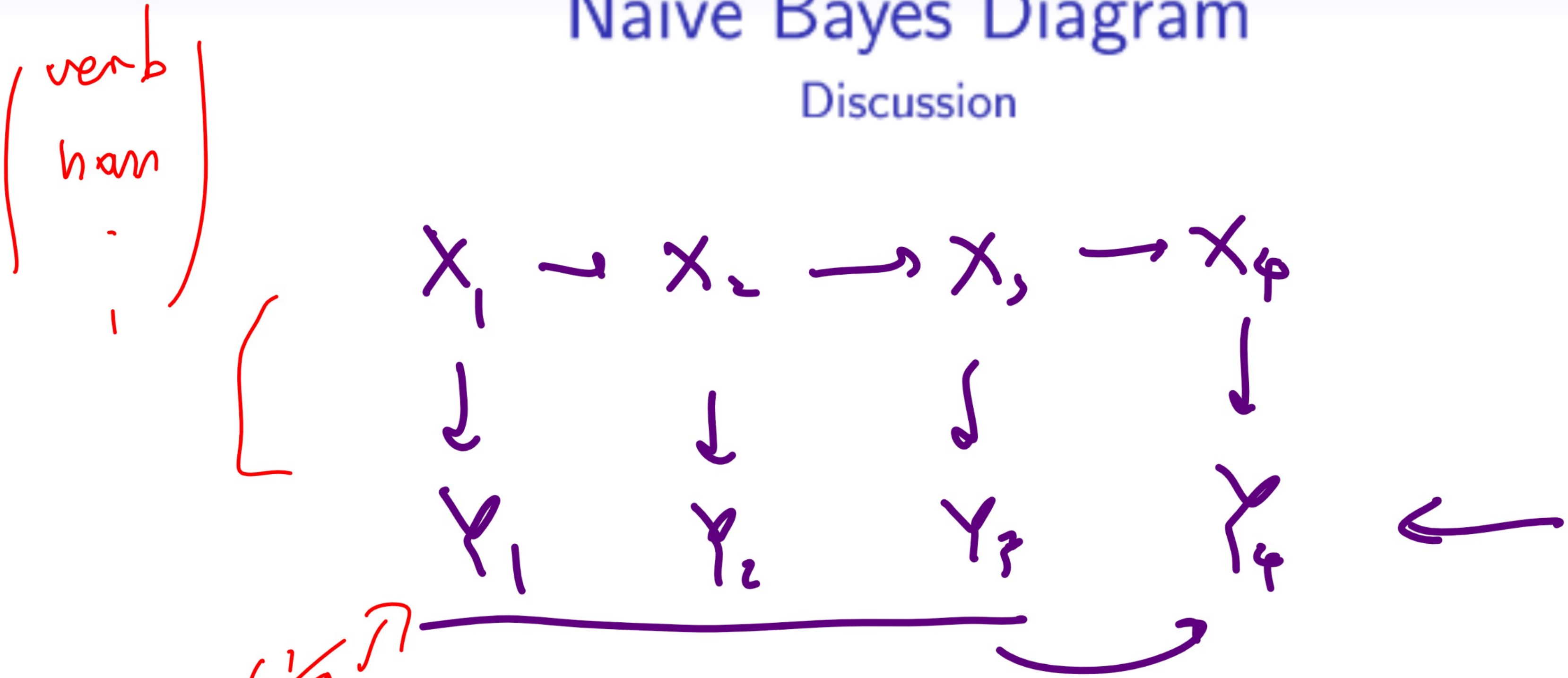
- Bayesian networks do not have a clear separation of the label Y and the features $X_1, X_2, ..., X_m$.
- The Bayesian network with a tree structure and Y as the root and $X_1, X_2, ..., X_m$ as the leaves is called the Naive Bayes classifier.
- Bayes rules is used to compute $\mathbb{P}\{Y = y | X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

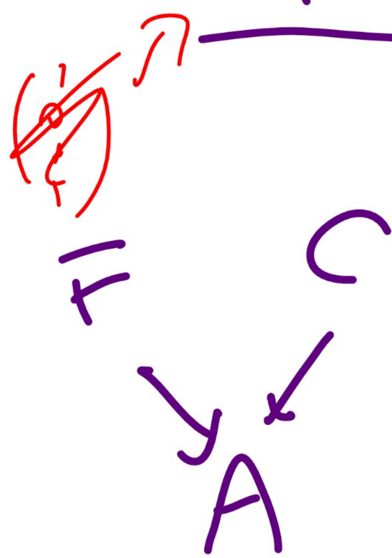
$$\hat{y}_i = \arg\max_y \mathbb{P}\{Y = y | X = x_i\}$$



Naive Bayes Diagram

Discussion





$\Pr\{F=1, C=1\} = \Pr\{F=1\} \cdot \Pr\{C=1\}$

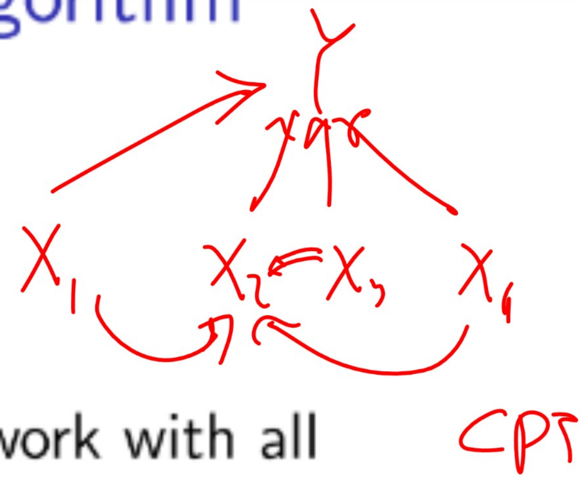
common effect

$\Pr\{F=1, C=1 \mid A=1\}$

$\rightarrow \Pr\{F=1 \mid A=1\} \Pr\{C=1 \mid A=1\}$

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