CS540 Introduction to Artificial Intelligence
Lecture 10

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

June 20, 2019

Joint Distribution

Motivation

- The joint distribution of X_j and $X_{j'}$ provides the probability of $X_j = x_j$ and $X_{j'} = x_{j'}$ occur at the same time.

$$\mathbb{P}\left\{X_j = x_j, X_{j'} = x_{j'}\right\}$$

training

$$\hat{P}_r\{z_t \mid z_{t-1}\} = \frac{c_{t,t-1}}{c_{t-1}}$$

classication → $\hat{z}_t = \arg\max_z P_r\{z \mid z_{t-1}\}$

- The marginal distribution of X_j can be found by summing over all possible values of $X_{j'}$.

$z_0, \to z_1 \to z_2 \to z_4.$

$$\mathbb{P}\left\{X_j = x_j\right\} = \sum_{x \in X_{j'}} \mathbb{P}\left\{X_j = x_j, X_{j'} = x\right\}$$

0, 1

$$P_r\{X_j = x_j\} = P_r\{X_j = x_j, X_{j'} = 0\} + P_r\{X_j = x_j, X_{j'} = 1\}$$

Conditional Distribution

Motivation

- Suppose the joint distribution is given.

$$\mathbb{P}\left\{X_j = x_j, X_{j'} = x_{j'}\right\}$$

- The conditional distribution of X_j given $X_{j'} = x_{j'}$ is ratio between the joint distribution and the marginal distribution.

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

Notation

Motivation

- The notations for joint, marginal, and conditional distributions will be shortened as the following.

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

- When the context is not clear, for example when $x_j = a, x_{j'} = b$ with specific constants a, b, subscripts will be used under the probability sign.

$$\mathbb{P}_{X_j, X_{j'}}\left\{a, b\right\}, \mathbb{P}_{X_j}\left\{a\right\}, \mathbb{P}_{X_j | X_{j'}}\left\{a | b\right\}$$

Bayes Rule

Quiz (Graded)

Q7

- Fall 2017 Final Q20
- Two documents A and B. Suppose $\hat{\mathbb{P}}\{H|A\} = 0.1$ in A and $\hat{\mathbb{P}}\{H|B\} = 0.8$ in B without Laplace smoothing. 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 H. 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: $\frac{1}{9}$

$$P_r\{A|H\} = \frac{P_r\{AH\}}{P_r\{H\}} = \frac{P_r\{H|A\} \cdot P_r\{A\}}{P_r\{HA\} + P_r\{HB\}} = \frac{\overset{0.1}{P_r\{H|A\}} \cdot \overset{\frac{1}{2}}{\cancel{P_r\{A\}}}}{\underset{0.1}{P_r\{H|A\}} \cdot \underset{\frac{1}{2}}{\cancel{P_r\{A\}}} + \overset{0.8}{P_r\{H|B\}} \cdot \overset{\frac{1}{2}}{\cancel{P_r\{B\}}}} = \frac{0.1}{0.1+0.8} = \frac{1}{9}$$

word

half of words in A is H

80% of B is H

$$P_r\{\underbrace{X_1 = \text{"H"}}_{\text{over words}} \mid \underbrace{X_2 = \text{"A"}}_{\text{over documents}}\}$$

$$P_r\{A|H\} = \frac{P_r\{HA\}}{P_r\{H\}} = \frac{P_r\{H|A\} \cdot P_r\{A\}}{P\{H\}}$$

$$= \frac{P_r\{H|A\} \cdot P_r\{A\}}{}$$

$P_r\{X_1 = \text{"B"}\} = \frac{1}{2}$, $P_r\{X_2 = \text{"A"}\} = \frac{1}{2}$

$\Pr\{H|A\}\cdot\Pr\{A\} + \Pr\{H|B\}\cdot\Pr\{B\}$.

Bayesian Network Diagram

Definition

- 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 or Happy or both might call you because of the alarm or for other reasons.

binary features

$\Pr\{H \mid \text{no } A\}$

$\Pr\{F \mid \{H, N\}\}$?

training set

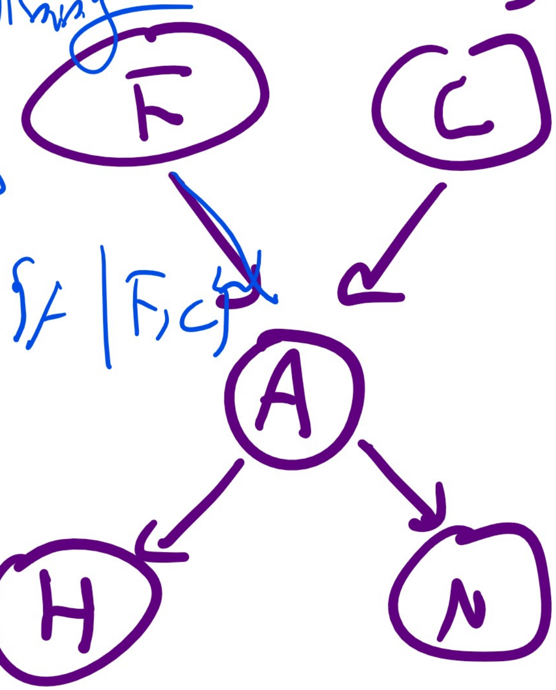
$\Pr\{C\}$

	x_1	x_2	x_3	x_4	x_5
	F	C	A	H	N
day 1	0	0	0	1	0
day 2	0	1	0	0	0
day 3	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

training

$\hat{\Pr}\{A \mid F, C\}$

inference



generate new data

simulation

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.

Conditional Independence

Definition

- Recall two events A, B are independent if:

$$\mathbb{P}\{A, B\} = \mathbb{P}\{A\}\,\mathbb{P}\{B\} \text{ or } \mathbb{P}\{A|B\} = \mathbb{P}\{A\}$$

- In general, two events A, B are conditionally independent, conditional on event C if:

$$\mathbb{P}\{A, B|C\} = \mathbb{P}\{A|C\}\,\mathbb{P}\{B|C\} \text{ or } \mathbb{P}\{A|B, C\} = \mathbb{P}\{A|C\}$$

Causal Chain

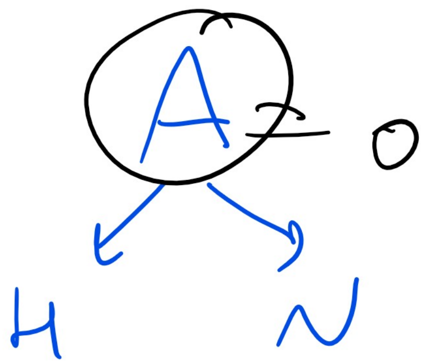
Definition

$C \to A \to H$

- For three events A, B, C, the configuration $A \to B \to C$ is called causal chain.
- In this configuration, A is not independent of C, but A is conditionally independent of C given information about B.
- Once B is observed, A and C are independent.

Common Cause

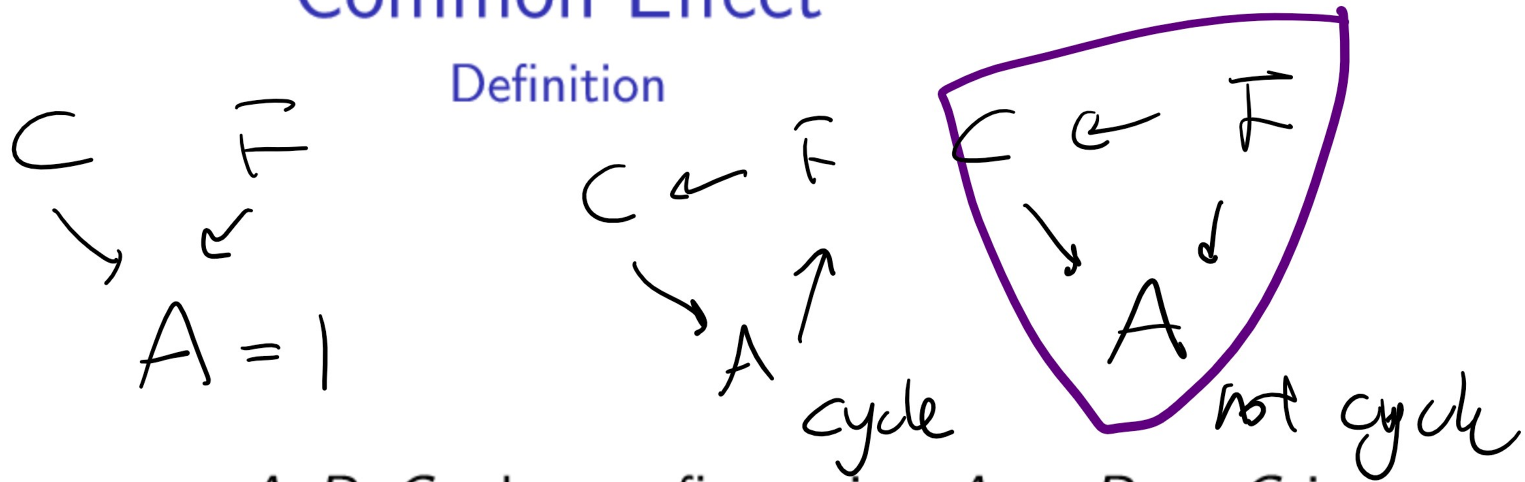
Definition



- For three events A, B, C, the configuration $A \leftarrow B \rightarrow C$ is called common cause.
- In this configuration, A is not independent of C, but A is conditionally independent of C given information about B.
- Once B is observed, A and C are independent.

Common Effect

Definition



- For three events A, B, C, the configuration $A \rightarrow B \leftarrow C$ is called common effect.
- In this configuration, A is independent of C, but A is not conditionally independent of C given information about B. Correct.
- Once B is observed, A and C are not independent.

$Pr\{A|B\}$

A is not conditionally indep of C

$$Pr\{A|C,B\} = \frac{Pr\{A,C,B\}}{Pr\{B,C\}} = \frac{Pr\{C\}Pr\{A,B\}}{Pr\{B,C\}} \neq \frac{Pr\{C\}Pr\{A,B\}}{Pr\{B\}Pr\{C\}}$$

Storing Distribution

Definition

$$\Pr\{A \mid B, C\} = \Pr\{A \mid C\} \text{ if } A, B \text{ cond indep}$$

- If there are m binary variables with k edges, there are 2^m joint probabilities to store.
- There are significantly less conditional probabilities to store. For example, if each node has at most 2 parents, then there are less than $4m$ conditional probabilities to store.
- Given the conditional probabilities, the joint probabilities can be recovered.

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

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

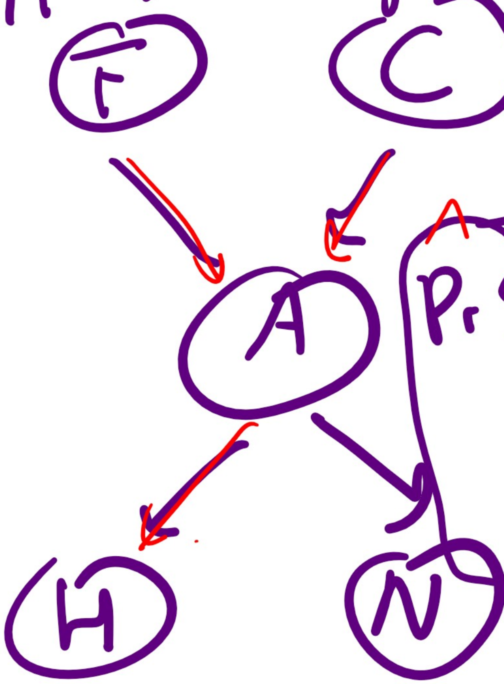
$$= \Pr\{F\} \cdot \Pr\{C \mid \cancel{F}\} \cdot \Pr\{A \mid C, F\} \cdot \Pr\{H \mid A, \cancel{C}, \cancel{F}\}$$

$Pr\{N \mid \cancel{H}, A, \cancel{C}, \cancel{F}\}$

Conditional Probability Table Diagram

Definition

$\hat{Pr}\{F=1\}$ $\hat{Pr}\{C=1\}$



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

___	0	1
	1	0
___	0	0
___		

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

. . . . .

2^m

$2^5 - 1 = 31$

$\hat{Pr}\{H=1 \mid A=0\}$ $\hat{Pr}\{N=1 \mid A=0\}$

$\hat{Pr}\{H=1 \mid A=1\}$ $\hat{Pr}\{N=1 \mid A=1\}$.

10 constant · m

$$Pr\{F=1, C=1, A=1, H=1, N=1\} = \prod_{j=1}^{m} Pr\{x_j \mid Parent(x_j)\}$$

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.

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

Bayes Net Training Example, Training, Part I

Definition

- Given a network and the training data.
  $F \rightarrow A, C \rightarrow A, A \rightarrow H, A \rightarrow N.$

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, Part II

Definition

- Compute $\hat{\mathbb{P}}\{F = 1\}$ $= \frac{c_{F=1}}{n} = \frac{1}{8}$

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, Part III

Definition

- Compute $\hat{\mathbb{P}}\{H = 1|A = 0\}$

$$\frac{C_{H=1,A=0}}{C_{A=0}} = \frac{2}{4} = \frac{1}{2}$$

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, Part V

Definition

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

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

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

$$= \frac{0}{1}$$

$$= 0$$

Bayes Net Training Example, Training, Part VI

Quiz (Graded)

Q8

- What is the conditional probability $\hat{\mathbb{P}}\{A = 1 | F = 0, C = 0\}$?
- 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

$\frac{4}{6} = \frac{2}{3}$

Laplace Smoothing

Definition

- Recall that the MLE estimation can incorporate Laplace smoothing.

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

$$\Pr\{A \mid F, C\} = \frac{c_{AFC} + 1}{c_{FC} + 4}$$

- 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 Example, Part I

Definition

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

$$\hat{\mathbb{P}}\left\{F = 1\right\} = 0.001, \hat{\mathbb{P}}\left\{C = 1\right\} = 0.001$$
$$\hat{\mathbb{P}}\left\{A = 1 | F = 1, C = 1\right\} = 0.95, \hat{\mathbb{P}}\left\{A = 1 | F = 1, C = 0\right\} = 0.94$$
$$\hat{\mathbb{P}}\left\{A = 1 | F = 0, C = 1\right\} = 0.29, \hat{\mathbb{P}}\left\{A = 1 | F = 0, C = 0\right\} = 0.00$$
$$\hat{\mathbb{P}}\left\{H = 1 | A = 1\right\} = 0.9, \hat{\mathbb{P}}\left\{H = 1 | A = 0\right\} = 0.05$$
$$\hat{\mathbb{P}}\left\{N = 1 | A = 1\right\} = 0.7, \hat{\mathbb{P}}\left\{N = 1 | A = 0\right\} = 0.01$$

Bayes Net Inference Example, Part II

Definition

$$\frac{\Pr\{F\,H\,N\}}{\Pr\{H\,N\}} = \frac{\Pr\{F,H,N\}}{\Pr\{F,H,N\} + \Pr\{\text{not}F,H,N\}}$$

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

- Compute $\hat{\mathbb{P}}\{F = 1 | H = 1, N = 1\}$?

$$\Pr\{F,H,N,\text{not } A, \text{not } C\} = \prod_{j=1}^{m} \Pr\{x_j / \text{Parents}(x_j)\}$$

$$= \Pr\{F\} \cdot \Pr\{H \mid \text{not } A\} \cdot \Pr\{N \mid \text{not} A\} \cdot \Pr\{\text{not } A \mid F, C\} \cdot \Pr\{\text{not } C\}.$$

Bayes Net Inference Example, Part III

Definition

0,001

$\Pr\{F, H, N, \text{not } A, C\} = 0.001 \cdot 0.05 \cdot 0.01 \cdot 0.05$

$$\begin{tikzcd}
F \arrow[dr] & & C \arrow[dl] \\
& A \arrow[dl] \arrow[dr] & \\
N & & H
\end{tikzcd}$$

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

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

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

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

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

Bayes Net Inference Example, Part IV

Definition

- Which of the following probabilies (multiple) are not required to compute $\hat{\mathbb{P}}\{C = 1|H = 1, N = 1\}$?
- A: $\hat{\mathbb{P}}\{A = 1|F = 1, C = 1\} = 0.95$
- B: $\hat{\mathbb{P}}\{A = 1|F = 1, C = 0\} = 0.94$
- C: $\hat{\mathbb{P}}\{A = 1|F = 0, C = 1\} = 0.29$
- D: $\hat{\mathbb{P}}\{A = 1|F = 0, C = 0\} = 0.00$
- E: none of the above.

Common Cause Example, Part I

Quiz (Graded)

- 2005 Fall Final Q20, 2006 Fall Final Q20
- Suppose A is the common cause of B and C. All variables are binary. What is $\mathbb{P}\{C = 1|B = 1\}$?

$$\mathbb{P}\{A = 1\} = 0.4, \mathbb{P}\{B = 1|A = 1\} = 0.9, \mathbb{P}\{B = 1|A = 0\} = 0.8$$

$$\mathbb{P}\{C = 1|A = 1\} = 0.5, \mathbb{P}\{C = 1|A = 0\} = 0.2$$

Common Cause Example, Part II

Quiz (Graded)

- What is $\mathbb{P}\{B = 1 | C = 1\}$?

$$\mathbb{P}\{A = 1\} = 0.4, \mathbb{P}\{B = 1 | A = 1\} = 0.9, \mathbb{P}\{B = 1 | A = 0\} = 0.8$$
$$\mathbb{P}\{C = 1 | A = 1\} = 0.5, \mathbb{P}\{C = 1 | A = 0\} = 0.2$$

- A: $\dfrac{0.9 \cdot 0.4 \cdot 0.5 \cdot 0.4 + 0.8 \cdot 0.6 \cdot 0.2 \cdot 0.6}{0.4 \cdot 0.5 + 0.6 \cdot 0.2}$
- B: $\dfrac{0.9 \cdot 0.4 \cdot 0.5 + 0.8 \cdot 0.6 \cdot 0.2}{0.4 \cdot 0.5 + 0.6 \cdot 0.2}$
- C: $\dfrac{0.9 \cdot 0.5 + 0.8 \cdot 0.2}{0.5 + 0.2}$
- D: $0.9 \cdot 0.4 + 0.8 \cdot 0.6$, E: none of the above

Bayesian Network

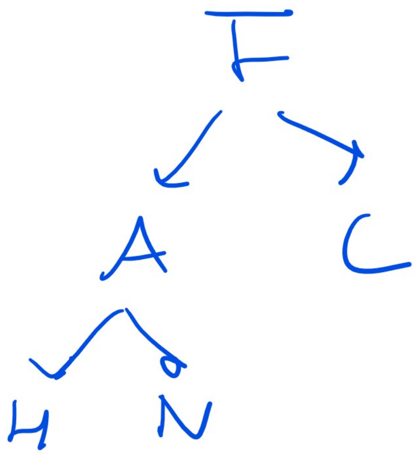
Algorithm

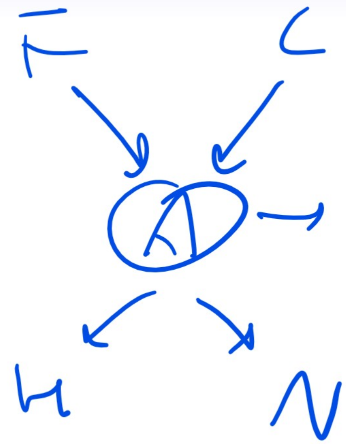
- Input: instances: $\{x_i\}_{i=1}^{n}$ and a directed acyclic graph such that feature X_j has parents $P(X_j)$.
- Output: conditional probability tables (CPTs): $\hat{\mathbb{P}}\{x_j | p(X_j)\}$ for $j = 1, 2, ..., m$.
- Compute the transition probabilities using counts and Laplace smoothing.

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

Network Structure

Discussion





two parents

↓

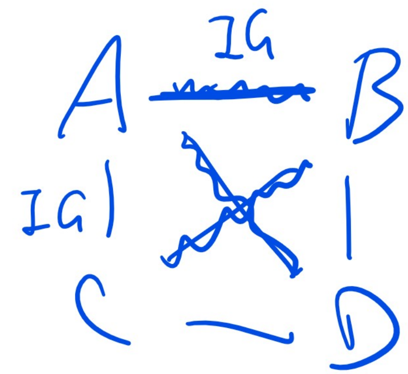
not tree

- 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.

one parent

Chow Liu Algorithm

Discussion



decision tree

- 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.

Aside: Prim's Algorithm

Discussion

- To find the maximum spanning tree, start with an arbitrary vertex, a vertex set containing only this vertex, V, and an empty edge set, E.
- Choose an edge with the maximum weight from a vertex $v \in V$ to a vertex $v' \notin V$ and add v' to V, add an edge from v to v' to E
- Repeat this process until all vertices are in V. The tree (V, E) is the maximum spanning tree.

Aside: Prim's Algorithm Diagram

Discussion

Classification Problem

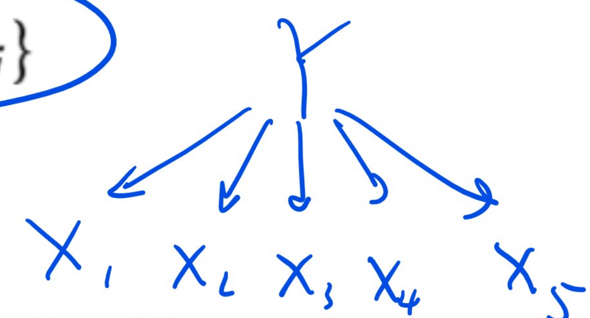
Discussion

- 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.

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

$Pr\{Y=1|x\}$, $Pr\{Y=0|x\}$?

Naive Bayes



$Pr\{X_1|Y=0\}$, $Pr\{X_1|Y=1\}$

Naive Bayes Diagram

Discussion

Multinomial Naive Bayes

Discussion

- The implicit assumption for using the counts as the maximum likelihood estimate is that the distribution of $X_j | Y = y$, or in general, $X_j | P(X_j) = p(X_j)$ has the multinomial distribution.

$$\mathbb{P}\left\{X_j = x | Y = y\right\} = p_x$$

$$\hat{p}_x = \frac{c_{x,y}}{c_y}$$

Gaussian Naive Bayes

Discussion

- If the features are not categorical, continuous distributions can be estimated using MLE as the conditional distribution.
- Gaussian Naive Bayes is used if $X_j | Y = y$ is assumed to have the normal distribution.

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{P} \left\{ x < X_j \leqslant x + \varepsilon \middle| Y = y \right\} = \frac{1}{\sqrt{2\pi} \sigma_y^{(j)}} \exp \left(-\frac{\left(x - \mu_y^{(j)} \right)^2}{2 \left(\sigma_y^{(j)} \right)^2} \right)$$

Gaussian Naive Bayes Training

Discussion

- Training involves estimating $\mu_y^{(j)}$ and $\sigma_y^{(j)}$ since they completely determines the distribution of $X_j | Y = y$.
- The maximum likelihood estimates of $\mu_y^{(j)}$ and $\left(\sigma_y^{(j)}\right)^2$ are the sample mean and variance of the feature j.

$$\hat{\mu}_y^{(j)} = \frac{1}{n_y} \sum_{i=1}^{n} x_{ij} \mathbb{1}_{\{y_i = y\}}, n_y = \sum_{i=1}^{n} \mathbb{1}_{\{y_i = y\}}$$

$$\left(\hat{\sigma}_y^{(j)}\right)^2 = \frac{1}{n_y} \sum_{i=1}^{n} \left(x_{ij} - \hat{\mu}_y^{(j)}\right)^2 \mathbb{1}_{\{y_i = y\}}$$

$$\text{sometimes } \left(\hat{\sigma}_y^{(j)}\right)^2 \approx \frac{1}{n_y - 1} \sum_{i=1}^{n} \left(x_{ij} - \hat{\mu}_y^{(j)}\right)^2 \mathbb{1}_{\{y_i = y\}}$$

Gaussian Naive Bayes Diagram

Discussion

Tree Augmented Network Algorithm

Discussion

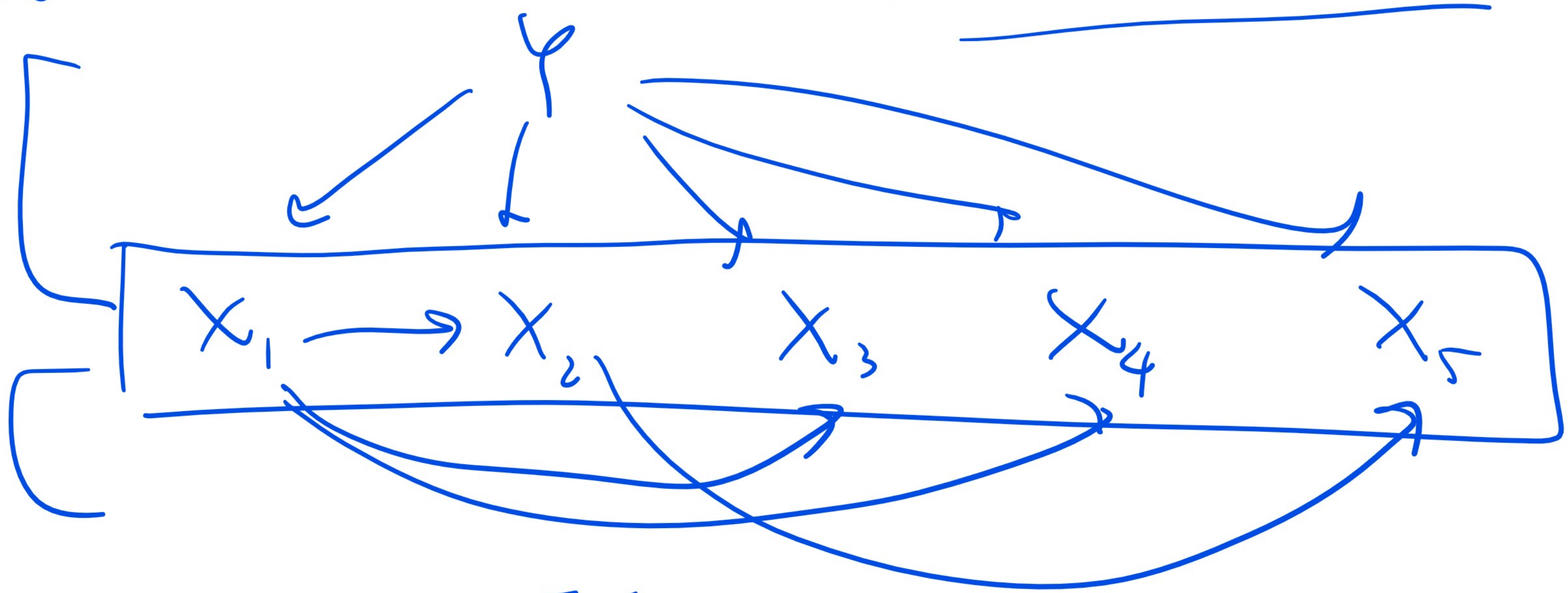
- It is also possible to create a Bayesian network with all features $X_1, X_2, ..., X_m$ connected to Y (Naive Bayes edges) and the features themselves form a network, usually a tree (MST edges).
- Information gain is replaced by conditional information gain (conditional on Y) when finding the maximum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

Tree Augmented Network Algorithm Diagram

Discussion

TAN

$P_r\{Y \mid X_1, \ldots X_5\}$



tree

max spanning tree

IG → conditional information on Y

end of supervised learning

end of midterm coverage