CS540 Introduction to Artificial Intelligence
Lecture 10

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

June 19, 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.

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

$$\Pr\{X_j = 0, X_{j'} = 0\} =$$

1	0	=
0	1	=
1	1	=

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

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

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

Conditional Distribution

Motivation

- Suppose the joint distribution is given.

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

$$\Pr\{x_j = 0 \mid x_{j'} = 0\} = \frac{\Pr\{x_j = 0, x_{j'} = 0\}}{\Pr\{x_{j'} = 0\}}$$

joint

marginal

- 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.

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

$Pr_{x_1, x_2}\{0, 1\}$

- 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.

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

Conditional Probability Example

Quiz (Graded)

if they are all binary

$\Pr\{X_3 = 1 \mid X_1 = 0, X_2 = 0\}$

- 2017 Fall Final Q3
- Given the counts, find the MLE (no smoothing) of $\mathbb{P}\{\text{ saw sheep} \mid \neg \text{ rainy }, \neg \text{ warm }\}$.

$\Pr\{X_3 \mid \neg X_1, \neg X_2\}$

rainy	warm	sheep	c	rainy	warm	sheep	c
N	N	N	1	Y	N	N	1
N	N	Y	0	Y	N	Y	1
N	Y	N	0	Y	Y	N	1
N	Y	Y	4	Y	Y	Y	2

- A: 0, B: $\frac{1}{4}$, C: $\frac{1}{3}$, D: $\frac{1}{2}$, E: 1

$$\frac{\Pr\{s, \neg r, \neg w\}}{\Pr\{\neg r, \neg w\}} \rightarrow \frac{c_{s=1, r=0, w=0}}{c_{r=0, w=0}}$$

$= 0$

$c_{s=1,r=0,w=0} = 0$, $c_{r=0,w=0} = 1$

Bayesian Network Diagram

Definition

- Story: You are travelling. There may be a Fire problem or a Cat problem at home. Either problem might trigger an Alarm. In case of Alarm, your neighbors Nick or Happy or both might call you.

Inference

x_1 x_2 x_3 binary

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

day1

day2

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

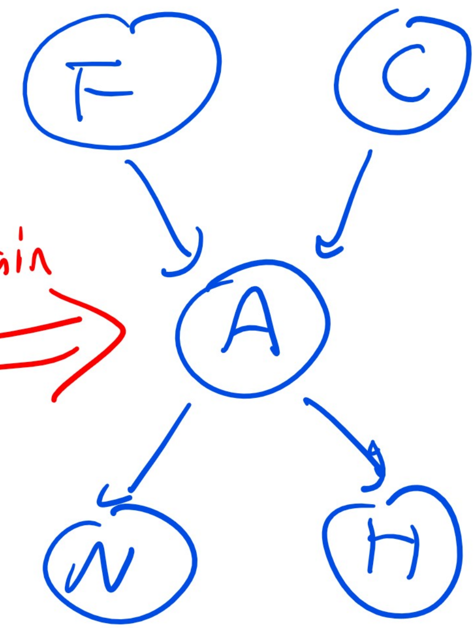
$\Pr\{F \mid N, H\}$

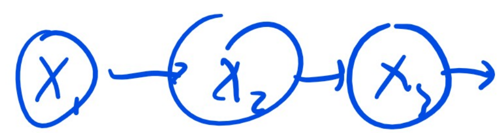
$\Pr\{N \mid F, C\}$

simulate

generate all event, for new day.

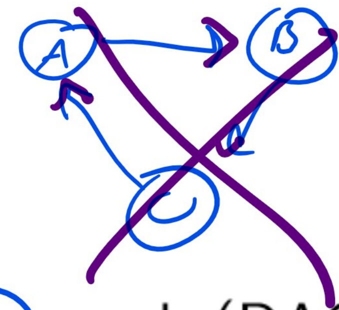
train





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:

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

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

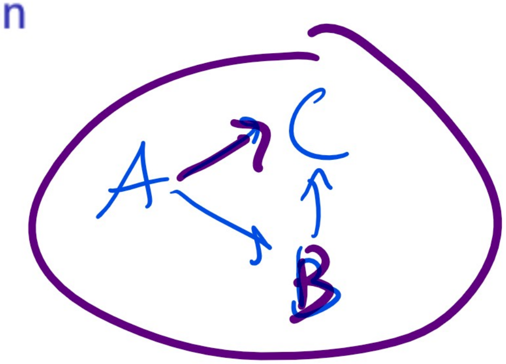
- In general, two events A, B are conditionally independent, conditional on event C if:

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

$$\frac{Pr\{A,B,C\}}{Pr\{B,C\}} = \checkmark$$

Causal Chain

Definition



- For three events A, B, C, the configuration $A \rightarrow B \rightarrow C$ 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

$$
\begin{tikzcd}
 & B \arrow[dl] \arrow[dr] & \\
A & & C
\end{tikzcd}
$$

- 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

$$\begin{tikzcd} A \arrow[rr] \arrow[dr] & & C \arrow[dl] \\ & B & \end{tikzcd}$$

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

Storing Distribution

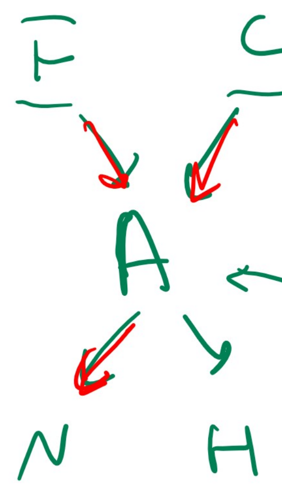
Definition

- 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~~ vertex 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.

Conditional Probability Table Diagram

Definition

① $\Pr\{F=1\}$ $\Pr\{C=1\}$



④ $\Pr\{A=1 \mid F=1, C=1\}$

$\Pr\{\_\}$ 0 1; 1 0; 0 0

② $\Pr\{N=1 \mid A=1\}$

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

③ $\Pr\{H=1 \mid A=1\}$

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

①

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

$2^5 = \boxed{32-1}$

2^N too large

$\boxed{10}$

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

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

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.
  $F \to A, C \to A, A \to H, A \to N.$

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

$$\frac{C_{F=1}}{n} = \frac{1}{8}$$

- Compute $\hat{\mathbb{P}}\{F = 1\} \Rightarrow$

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

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

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

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 IV

Quiz (Graded)

- What is the conditional probability $\hat{\mathbb{P}}\{H = 1 | A = 1\}$?
- A: 0 , B: $\frac{1}{4}$, C: $\frac{1}{2}$, D: $\frac{3}{4}$, E: 1

Q1

$$\frac{\cancel{\Pr}\{H=1, A=1\}}{\cancel{\Pr}\{A=1\}}$$

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

$$\frac{3}{4}$$

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

Bayes Net Training Example, Training, Part VI

Quiz (Graded)

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

$\sim ?$

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

$\dfrac{\quad}{C_{F=1, C=1}}$

Laplace Smoothing

Definition

- Recall that the MLE estimation can incorporate Laplace smoothing.

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

categories of 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 Example, Part I

Definition

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

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

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

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

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

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

Bayes Net Inference Example, Part II

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

$\Pr\{H=1, N=1\}$

Definition

$$\Pr\{F, H, N\} \Rightarrow \underline{\Pr\{F, H, N, A=0/1, C=0/1\}}$$

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

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

$\to \Pr\{F=1\} \cdot \Pr\{C=0\}$

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

$\Pr\{H=1 | A=0\}$

$\Pr\{N=1 | A=0\}$

Bayes Net Inference Example, Part III

$Pr\{F=1, N=1, H=1\}$

Definition

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

$0.001 \cdot 0.999 \cdot \cancel{0.94}\ 0.06 \cdot 0.01 \cdot 0.05$

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

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

$Pr\{A=0|F=1, C=0\} = 0.06$

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

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

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

Bayes Net Inference Example, Part IV

Definition

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

Bayes Net Inference Example, Part V

Definition

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

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

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

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

$$\begin{tikzcd} & A \arrow[dl] \arrow[dr] & \\ B & & C \end{tikzcd}$$

Common Cause Example, Part II

Quiz (Graded)

- Suppose A is the common cause of B and C. All variables are binary. 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

given

- 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.

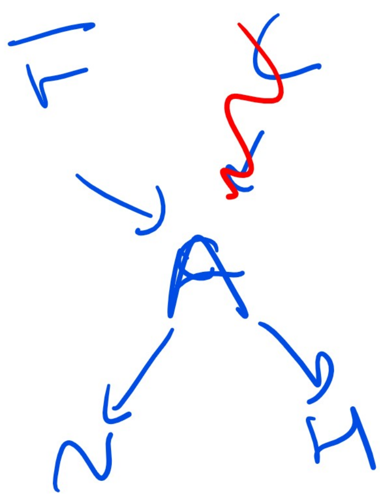
$$\hat{\mathbb{P}}\{x_j | p(X_j)\} = \frac{c_{x_j, p(X_j)} + 1}{c_{p(x_j)} + |X_j|}$$

Network Structure

Discussion

- 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.

only one parent



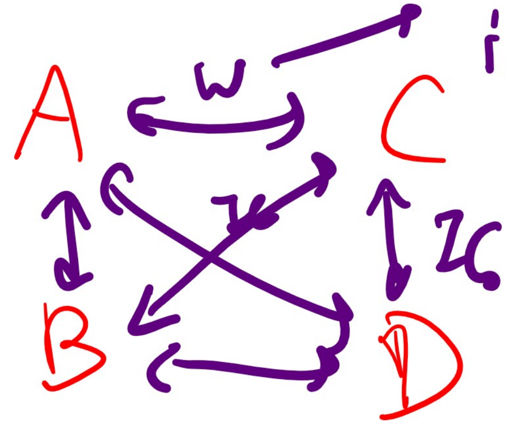
Chow Liu Algorithm

Discussion

$$IG(X_j \mid X_{j'}) = \underline{H(X_{j'})} - H(X_j \mid X_{j'})$$

$$= -\sum_{x=1}^{\#\text{category}} \Pr\{X_j = x\} \cdot \log_2 \Pr\{X_j = x\} + \sum_{x=1}^{\#} \sum_{x'=1}^{\#} \Pr\{X_j = x \mid X_{j'} = x'\} \log_2 \Pr\{X_j = x \mid X_{j'} = x'\}$$

- 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.



info gain of A given C

Decision Tree.

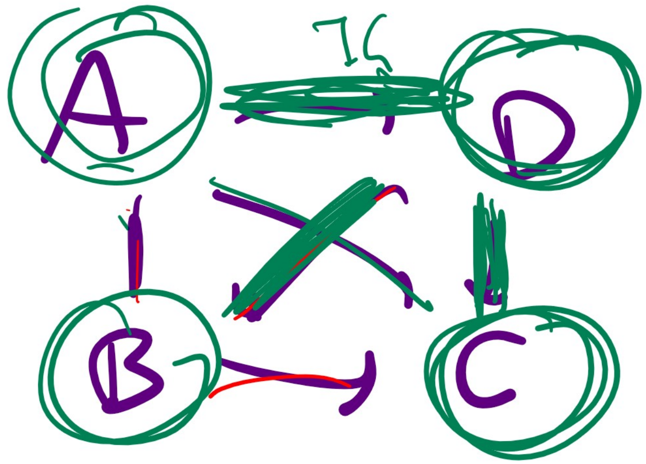
Aside: Prim's Algorithm

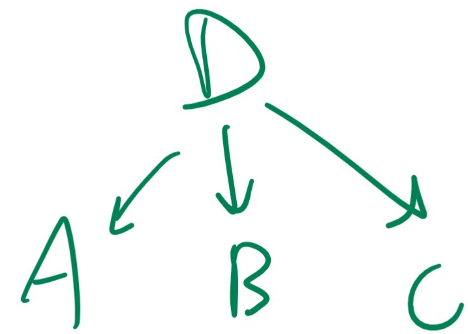
Discussion

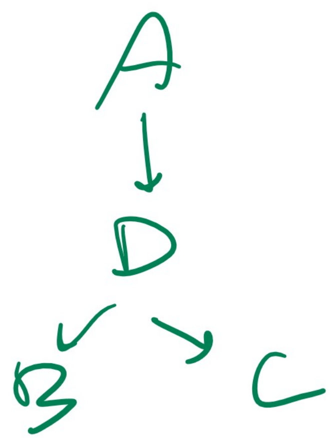
- To find the maximum spanning tree, start with an arbitrary vertex, a vertex set containing only this vertex, V, and an empty edge set, E.
- Choose an edge with the maximum weight from a vertex $v \in V$ to a vertex $v' \notin V$ and add v' to V, add an edge from v to v' to E
- Repeat this process until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

Aside: Prim's Algorithm Diagram

Discussion







Classification Problem

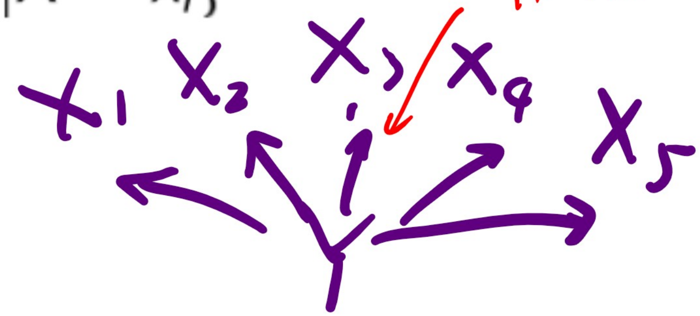
Discussion

- 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.

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

wrong direction in lecture

Naive Bayes $\longrightarrow$



Naive Bayes Diagram

Discussion

train on $Y, X_1, X_2 \ldots X_J$

test now $X = 1 \; 0 \; 1 \; 0 \; 1$

compare

$\Pr\{Y = 0 \mid 1\,0\,1\,0\,1\}$

$\Pr\{Y = 1 \mid 1\,0\,1\,0\,1\}$

$$\hat{y} = \underset{y \in \{0,1\}}{\operatorname{argmax}} \Pr\{Y = y \mid 1\,0\,1\,0\,1\},$$

X_1 is categorical

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.

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

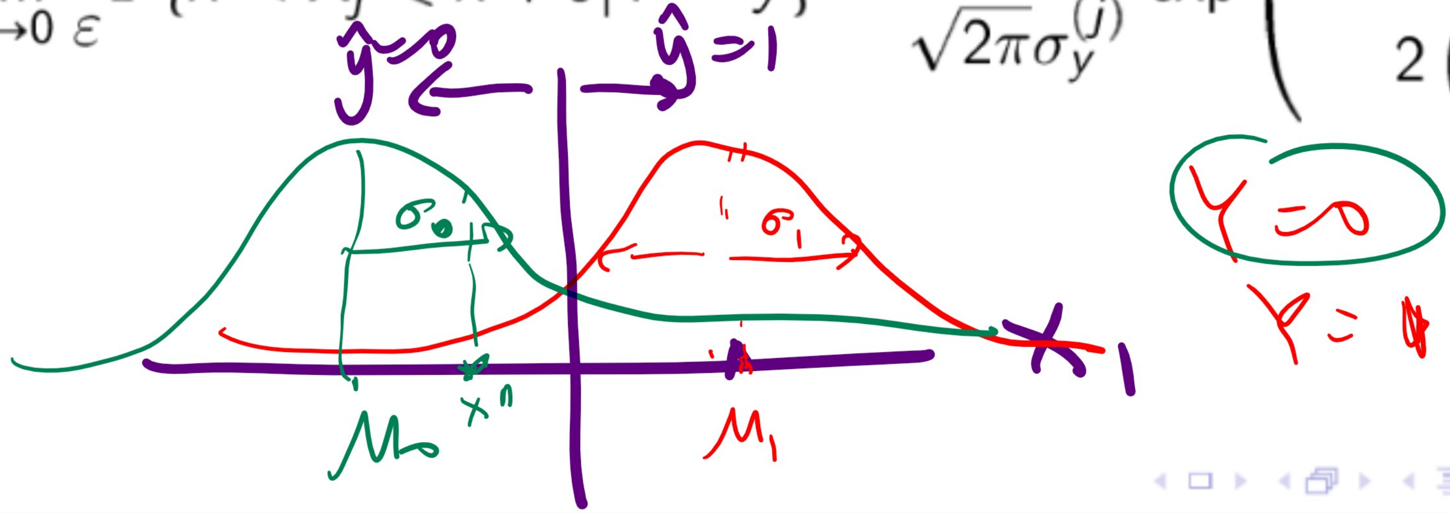
$$\hat{p}_x = \frac{c_{x,y}}{c_y}$$

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.

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



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.

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

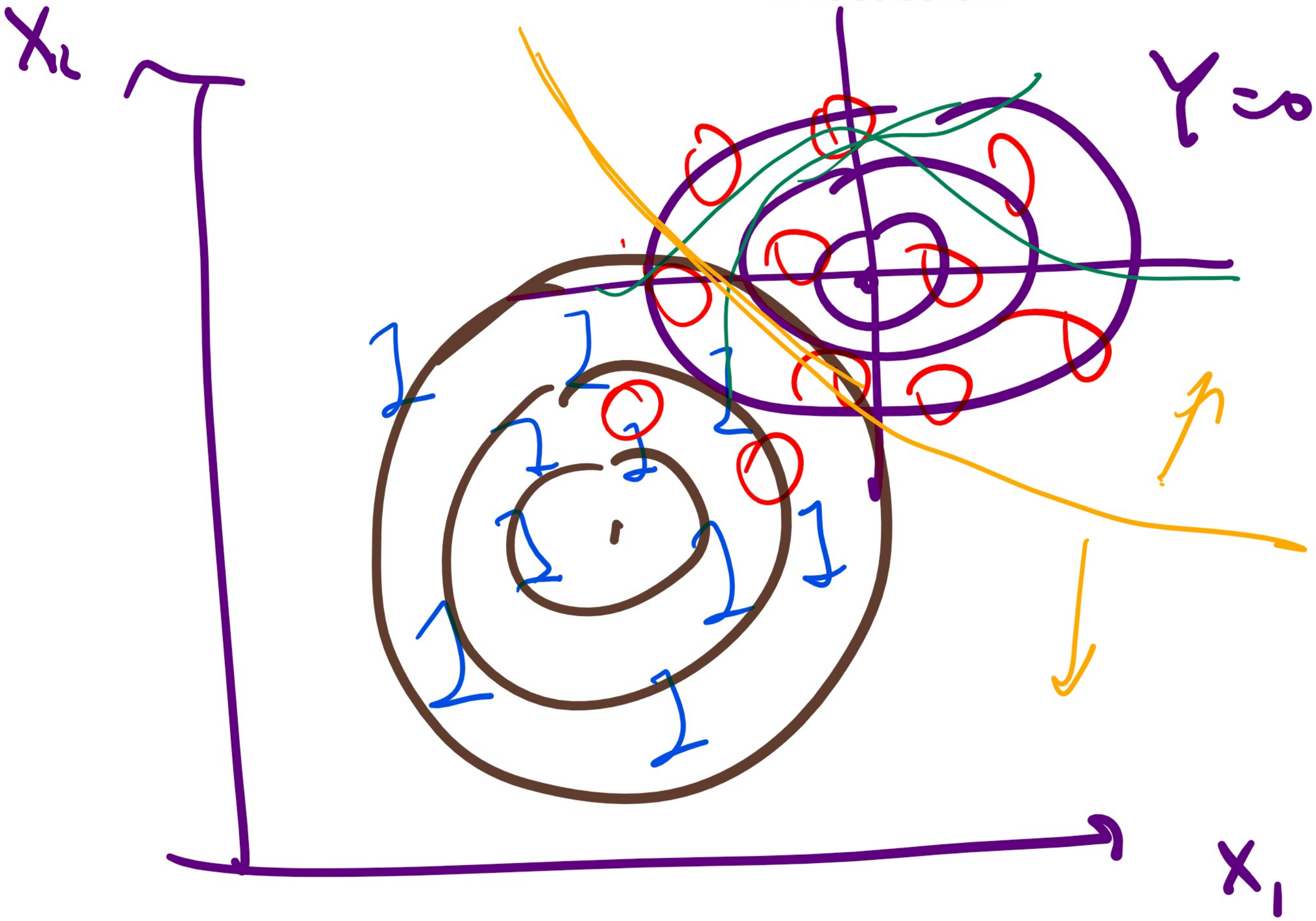
MLE

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

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

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