CS540 Introduction to Artificial Intelligence
Lecture 10

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

June 8, 2020

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.

$$\mathbb{P}\{X_j = \overset{0}{x_j}, X_{j'} = \overset{0}{x_{j'}}\} \in [0, 1]$$

0, 1, 2   0, 1, 2   1

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

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

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.

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

joint

marginal

Distribution Example

Motivation

	first coin T	H	
$X_1 =$	0	1	X_2
$X_2 = 0$	$\boxed{1/4}$	0	1/4
#H 2 coins 1	1/4	$\boxed{1/4}$	$\boxed{1/2}$
2	0	1/4	1/4
$X_1 \rightarrow$	$\boxed{1/2}$	1/2	

HH	HT	$\boxed{TH}$	TT
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$\Pr\{X_1 = 1 \mid X_2 = 1\} = \frac{1/4}{1/2} = \frac{1}{2}$$

$$\Pr\{X_2 = 1 \mid X_1 = 1\} = \frac{1/4}{1/2} = \frac{1}{2}$$

Notation

Motivation

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

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

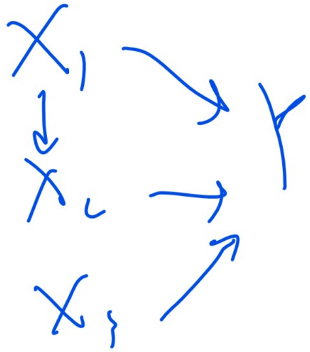
$Pr\{X_1 = x_1, X_2 = x_2\}$

$Pr\{1, 1\}$ $Pr\{x_1, x_2\}$

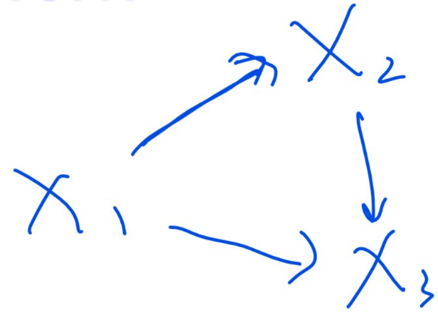
- When the context is not clear, for example when $x_j = a, x_{j'} = b$ with specific constants a, b, subscripts will be used under the probability sign.

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

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

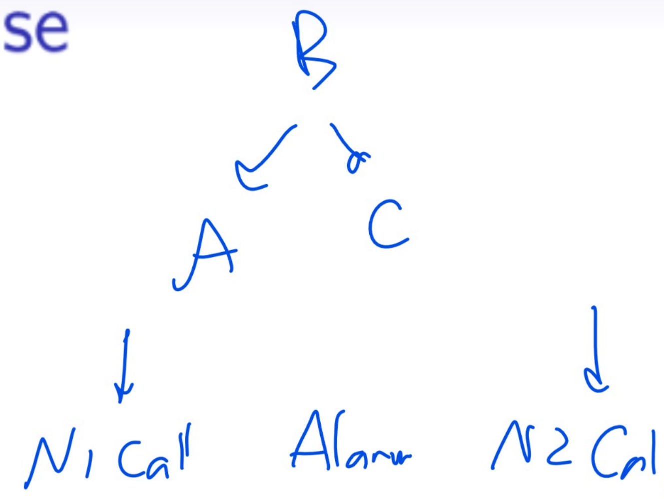
Definition

Fire Alarm Call

- 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



- 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

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

Fire Alarm Cat

- 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.
- 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 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

Fire

Cat
(Flerken)

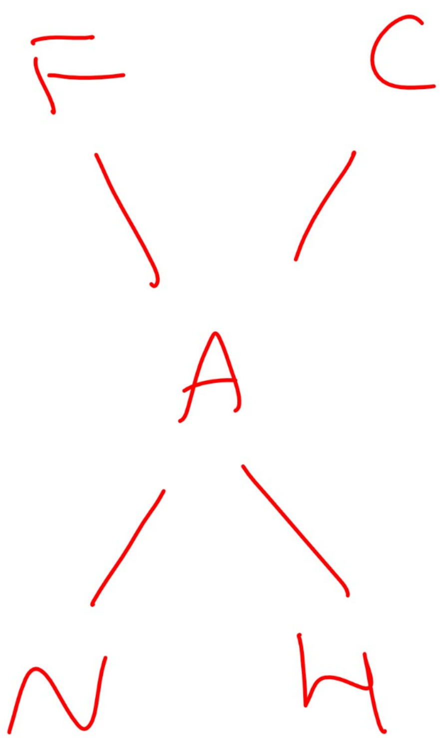
Alarm

Nick

Happy

Conditional Probability Table Example

Definition



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

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

$2^5 = 32 - 1$

$$\left.\begin{array}{l} \Pr\{F\} \quad \cancel{\Pr\{\neg F\}} \quad \Pr\{C\} \\ \Pr\{A \mid FC\}, \ \Pr\{A \mid \neg FC\} \\ \Pr\{A \mid F\neg C\}, \ \Pr\{A \mid \neg F \neg C\} \\ \Pr\{N \mid A\}, \ \Pr\{N \mid \neg A\}, \ \Pr\{H \mid A\}, \ \Pr\{H \mid \neg A\} \end{array}\right\} 10$$

$\cancel{\Pr\{\neg A \mid FC\}}$

Conditional Probability Table Larger Example

Definition

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 $P(X_j)$.

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

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

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

Bayes Net Training Example, Part I

Definition

	F	C	A	H	N	
day 1	0	0	0	1	0	$\leftarrow x_1$
day 2	0	1	0	0	0	$\leftarrow x_2$
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	

$$\Pr\{F\} = \frac{1}{8}$$

$$\Pr\{H \mid A\} = \frac{3}{4}$$

$$\Pr\{H \mid \neg A\} = \frac{2}{4} = \frac{1}{2}$$

$$\Pr\{N \mid \neg A\} = \frac{2}{4} = \frac{1}{2}$$

$$\Pr\{A \mid \neg F, \neg C\} = \frac{4}{6} = \frac{2}{3}$$

$$\Pr\{A \mid F, C\} = \frac{0}{0} = ?$$

Bayes Net Training Example, Part II

Definition

Laplace Smoothing

Definition

- Recall that the MLE estimation can incorporate Laplace smoothing.

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

$P_r\{A | F, C\}$

$= \frac{0+1}{0+2} = \frac{1}{2}$

- 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.

$\frac{\emptyset 1}{0}$

Bayes Net Inference

Definition

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

$$\mathbb{P}\left\{x_1, x_2, ..., x_m\right\} = \cancel{\text{proj}}_{j=1}^{m} \mathbb{P}\left\{x_j | x_{j+1}, x_{j+2}, ..., j_m\right\}$$

$\prod$

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

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

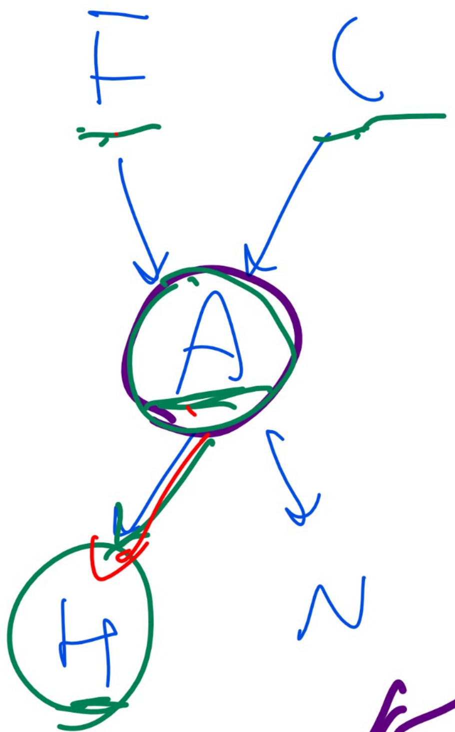
$$\mathbb{P}\left\{x_j | x_{j'}, x_{j''}, ...\right\} = \frac{\mathbb{P}\left\{x_j, x_{j'}, x_{j''}, ...\right\}}{\mathbb{P}\left\{x_{j'}, x_{j''}, ...\right\}}$$

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

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

Bayes Net Inference Example, Part I

Definition



$$\Pr\{H \mid \neg F, \neg C\}$$

$$= \frac{\Pr\{H, \neg F, \neg C\}}{\Pr\{\neg F, \neg C\}}$$

$$\Pr\{\neg F\} \cdot \Pr\{\neg C\}$$

$$= (1 - \underbrace{\Pr\{F\}}_{\text{in CPT}})(1 - \underbrace{\Pr\{C\}}_{\text{in CPT}})$$

$$\Pr\{H, A, \neg F, \neg C\} + \underbrace{\Pr\{H, \neg A, \neg F, \neg C\}}_{\Pr\{H \mid \neg A\} \cdot \Pr\{\neg A \mid \neg F, \neg C\} \cdot \Pr\{\neg F\} \cdot \Pr\{\neg C\}}$$

Bayes Net Inference Example, Part II

Definition

$$\underbrace{\Pr\{H \mid A}_{\text{in CPT}}, \cancel{\neg F}, \cancel{\neg C}\} \cdot \Pr\{A, \neg F, \neg C\}$$

$$\Pr\{A, \neg F, \neg C\} = \underbrace{\Pr\{A \mid \neg F, \neg C\}}_{\text{in CPT}} \cdot \Pr\{\neg F, \neg C\}$$

$$\Pr\{\neg F, \neg C\} \rightarrow (1 - \underbrace{\Pr\{F\}}_{\text{in CPT}})(1 - \underbrace{\Pr\{C\}}_{\text{in CPT}})$$

from training

Bayes Net Inference Example, Part III

Definition

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

- 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.

5??

Aside: Prim's Algorithm

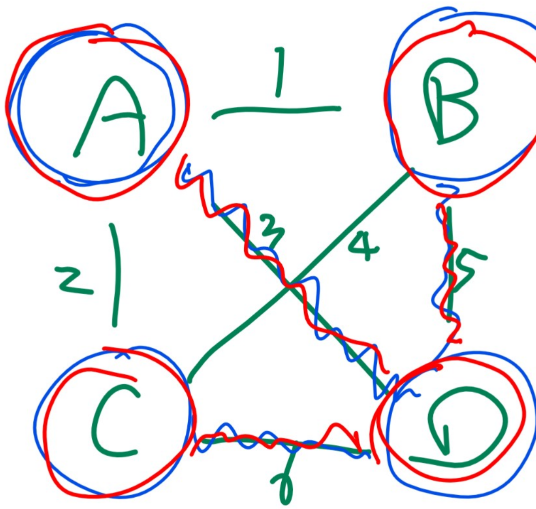
Discussion

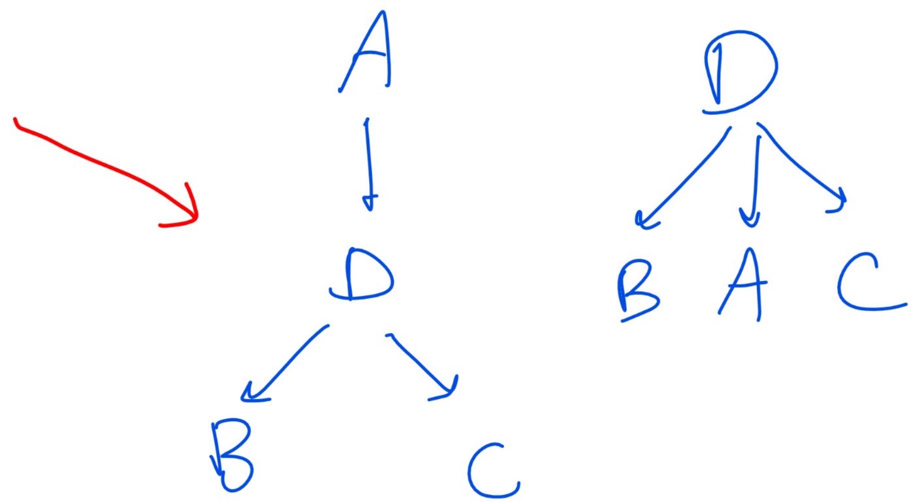
- To find the maximum spanning tree, start with an arbitrary vertex, a vertex set containing only this vertex, V, and an empty edge set, E.
- Choose an edge with the maximum weight from a vertex $v \in V$ to a vertex $v' \notin V$ and add v' to V, add an edge from v to v' to E
- Repeat this process until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

Aside: Prim's Algorithm Diagram

Discussion

$$\boxed{\begin{aligned}&\mathrm{InfoGain}(X_i, X_j)\\ &= \mathrm{InfoGain}(X_j, X_i)\end{aligned}}$$





Classification Problem

Discussion

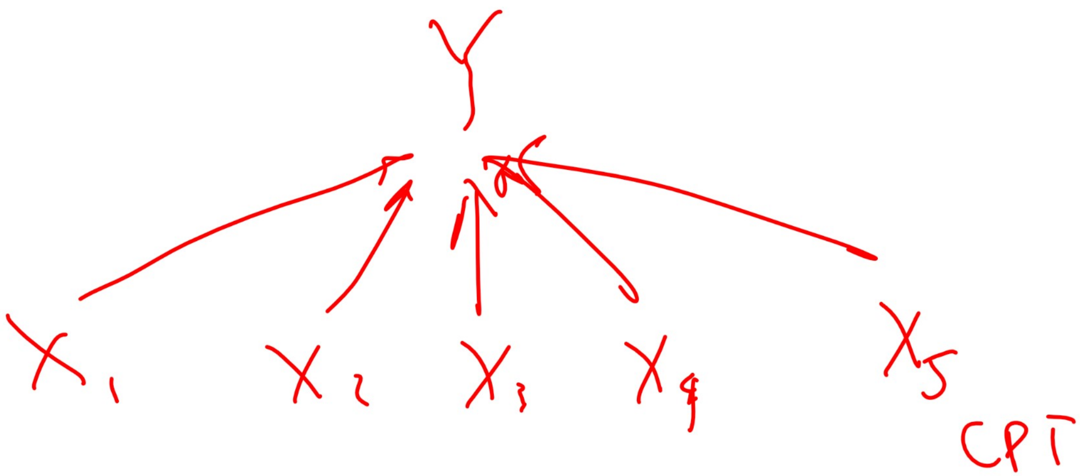
- Bayesian networks do not have a clear separation of the label Y and the features $X_1, X_2, ..., X_m$.
- The Bayesian network with a tree structure and Y as the root and $X_1, X_2, ..., X_m$ as the leaves is called the Naive Bayes classifier.
- Bayes rules is used to compute $\mathbb{P}\{Y = y | X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

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

$$\frac{Pr\{Y=y, X=x_i\}}{Pr\{X=x_i\}}$$

Naive Bayes Diagram

Discussion



P3.

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.

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

0
1
2

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

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.

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

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.

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

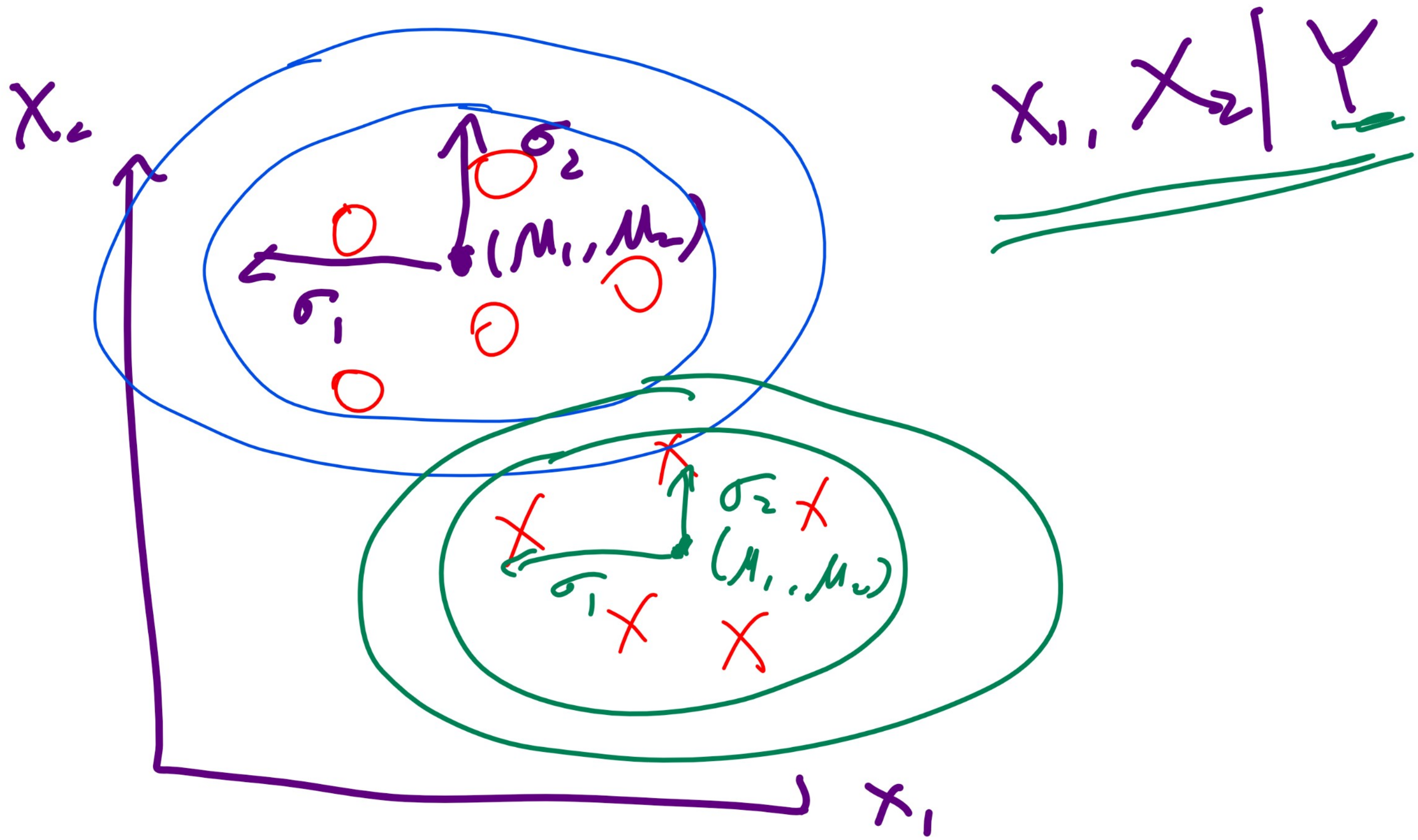
$$\left(\hat{\sigma}_y^{(j)}\right)^2 = \frac{1}{n_y} \sum_{i=1}^{n} \left(x_{ij} - \hat{\mu}_y^{(j)}\right)^2 \mathbb{1}_{\{y_i = y\}}$$

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

MLE

Gaussian Naive Bayes Diagram

Discussion



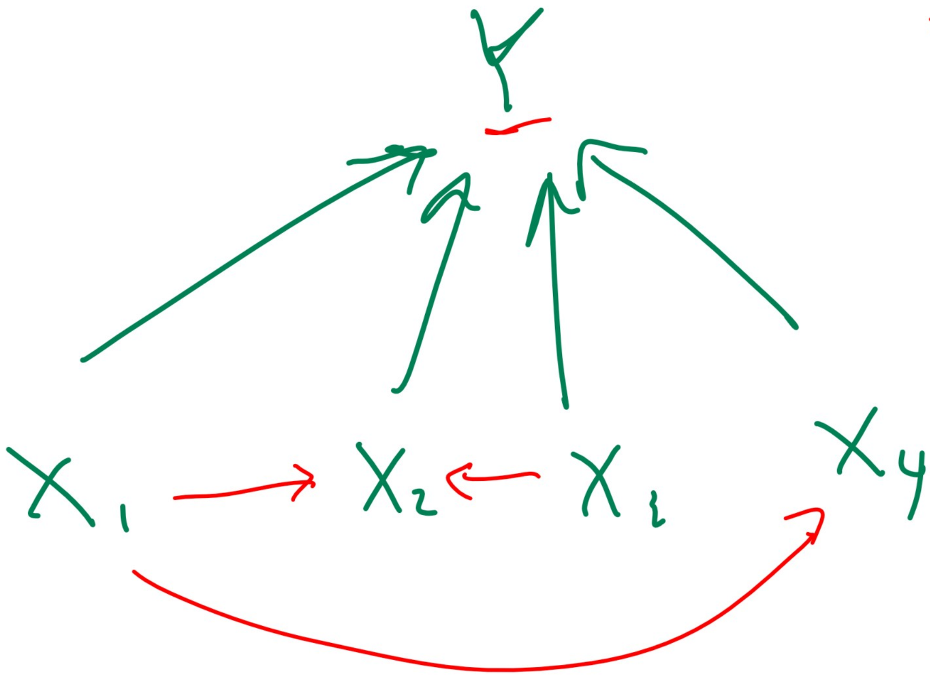
Tree Augmented Network Algorithm

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



$Y \mid X_1, X_2, X_3, X_4$