UT DALLAS

CS 6347

Lecture 3

Bayesian Networks

Chain Rule

$$p(x_1, x_2) = p(x_1)p(x_2|x_1)$$

.

.

.

$$p(x_1, \dots, x_n) = p(x_1)p(x_2|x_1) \dots p(x_n|x_1, \dots, x_{n-1})$$

Structured Distributions

- Consider a general joint distribution $p(X_1, \dots, X_n)$ over binary valued random variables
- If $X_1, \dots, X_n$ are all independent given a different random variable Y, then

$$p(x_1, \dots, x_n|y) = p(x_1|y) \dots p(x_n|y)$$

and

$$p(y, x_1, \dots, x_n) = p(y)p(x_1|y) \dots p(x_n|y)$$

- How much storage is needed to represent this model?

Structured Distributions

- Consider a different joint distribution $p(X_1, \dots, X_n)$ over binary valued random variables
- Suppose, for $i > 2$, X_i is independent of $X_1, \dots, X_{i-2}$ given X_{i-1}

$$p(x_1, \dots, x_n) = p(x_1)p(x_2|x_1) \dots p(x_n|x_1, \dots, x_{n-1})$$
$$= p(x_1)p(x_2|x_1)p(x_3|x_2) \dots p(x_n|x_{n-1})$$

- How much storage is needed to represent this model?
- This distribution is chain-like

Bayesian Network

- A **Bayesian network** is a directed graphical model that captures independence relationships of a given probability distribution
  - Directed acyclic graph (DAG), $G = (V, E)$
  - One node for each random variable
  - One conditional probability distribution per node
  - Directed edge represents a direct statistical dependence

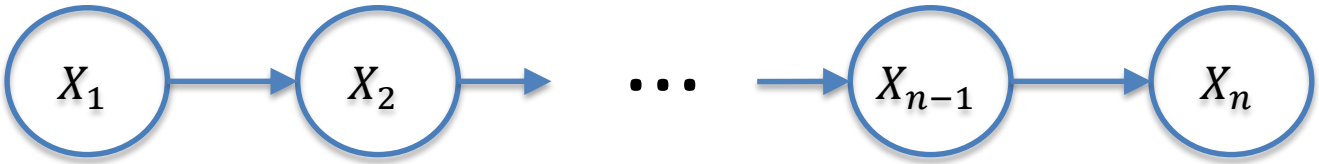
Bayesian Network

- A **Bayesian network** is a directed graphical model that captures independence relationships of a given probability distribution
  - Encodes **local Markov** independence assumptions that each node is independent of its non-descendants given its parents
  - Corresponds to a **factorization** of the joint distribution

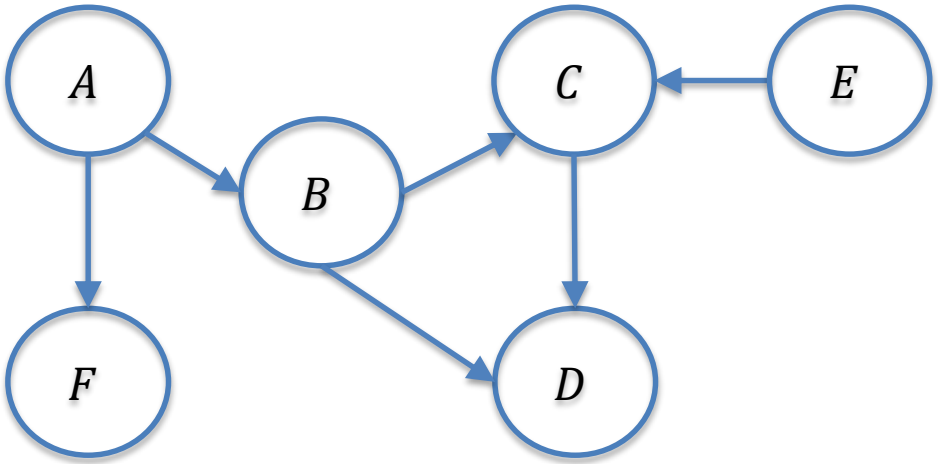
$$p(x_1, \ldots, x_n) = \prod_i p(x_i | x_{parents(i)})$$

Directed Chain

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \ldots p(x_n|x_{n-1})$$

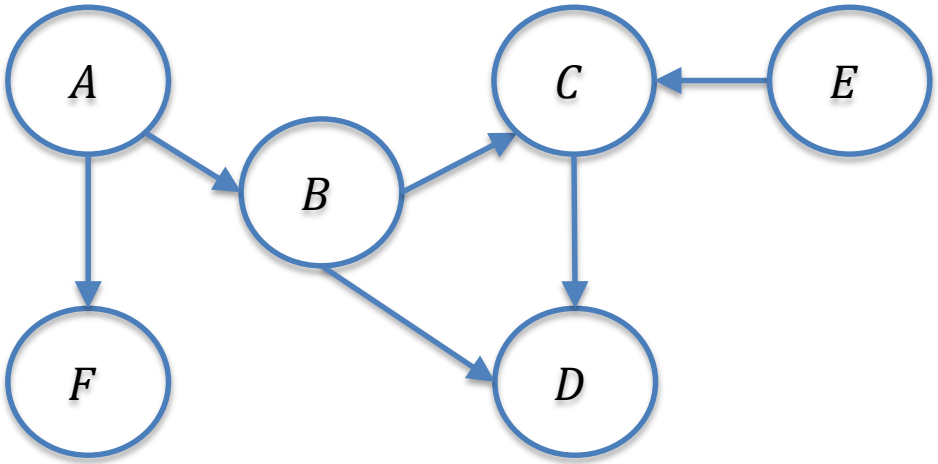


Example:



- Local Markov independence relations?
- Joint distribution?

Example:



- This list is not exhaustive:
  - How can we figure out which independence relationships the model represents?

D-separation

- Independence relationships can be figured out by looking at the graph structure!
  - Easier than looking at the tables and plugging into the definition
- We look at all of the paths from X to Y in the graph and determine whether or not they are **blocked**
  - $X \subset V$ is d-separated from $Y \subset V$ given $Z \subset V$ iff every path from X to Y in the graph is blocked by Z

D-separation

- Three types of situations can occur along any given path

(1) Sequential

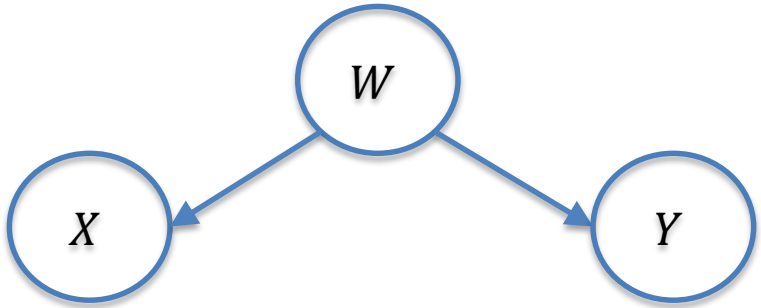
The path from X to Y is blocked if we condition on W

Intuitively, if we condition on W, then information about X does not affect Y and vice versa

D-separation

- Three types of situations can occur along any given path

(2) Divergent



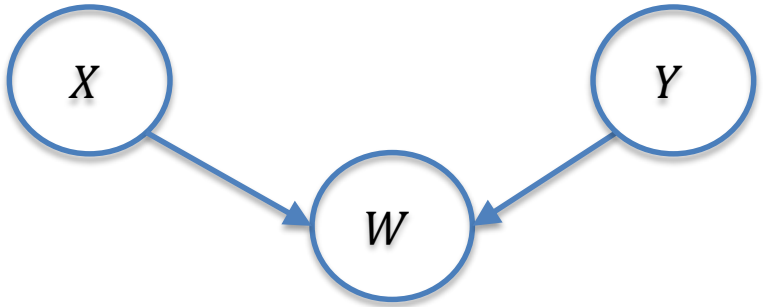
The path from X to Y is blocked if we condition on W

If we don't condition on W, then information about W could effect the probability of observing either X or Y

D-separation

- Three types of situations can occur along any given path

(3) Convergent



The path from X to Y is blocked if we **do not** condition on W or any of its descendants

Conditioning on W couples the variables X and Y: knowing whether or not X occurs impacts the probability that Y occurs

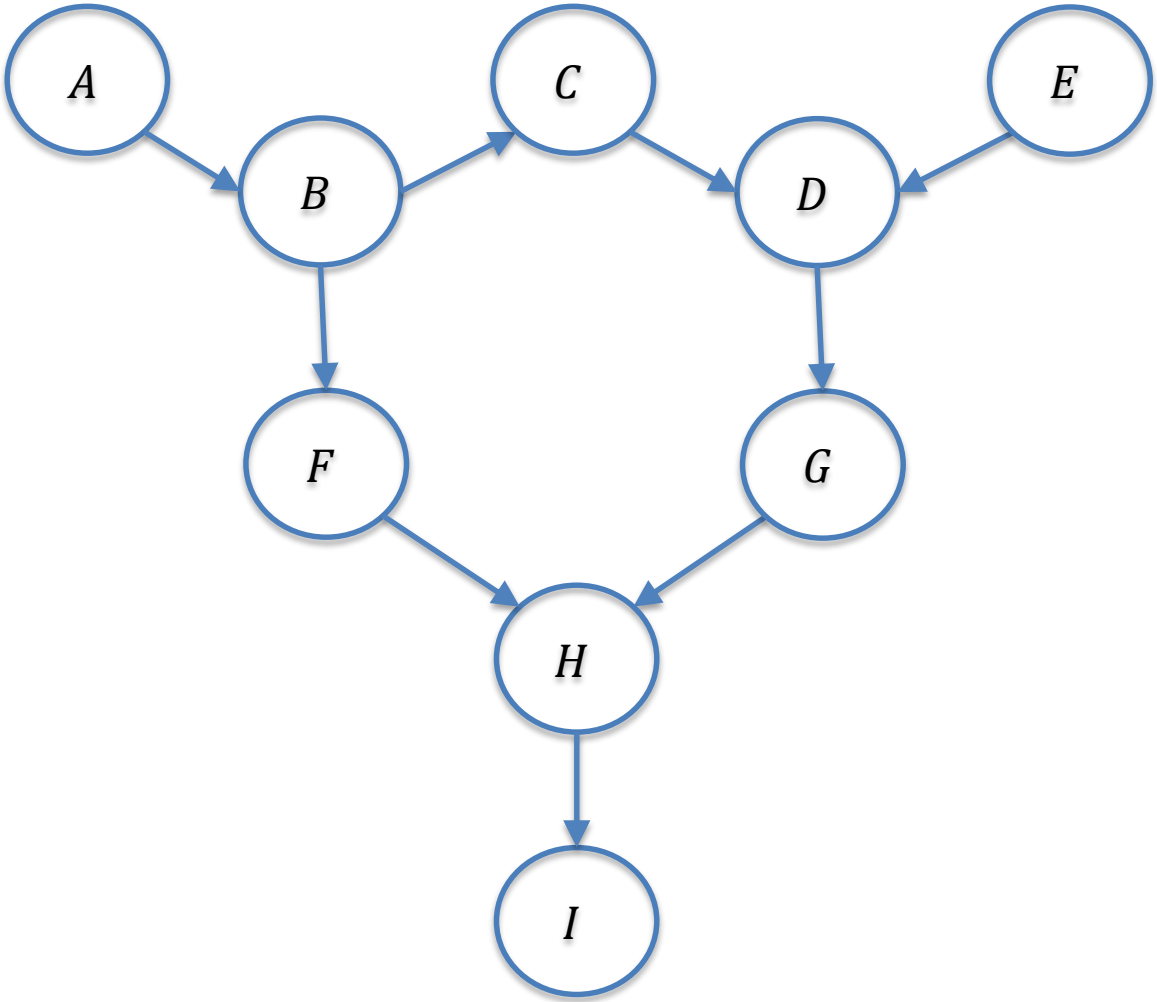
D-separation

- If the joint probability distribution factorizes with respect to the DAG $G = (V, E)$, then X is d-separated from Y given Z implies $X \perp Y \mid Z$
  - We can use this to quickly check independence assertions by using the graph
  - In general, these are only a subset of all independence relationships that are actually present in the joint distribution
  - If X and Y are not d-separated in G given Z, then there is some distribution that factorizes over G in which X and Y dependent

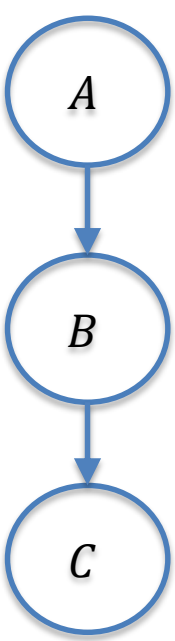
D-separation

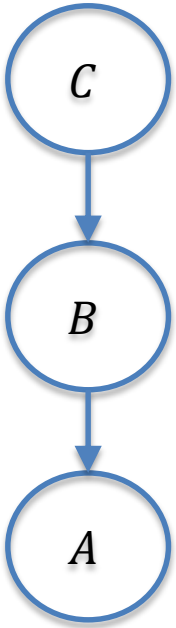
- Let $I(p)$ be the set of all independence relationships in the joint distribution p and $I(G)$ be the set of all independence relationships in the graph G
- We say that G is an I-map for $I(p)$ if $I(G) \subseteq I(p)$
- Theorem: the joint probability distribution, p, factorizes with respect to the DAG $G = (V, E)$ iff G is an I-map for $I(p)$
- An I-map is perfect if $I(G) = I(p)$
  - Not always possible to perfectly represent all of the independence relations with a graph

D-separation Example



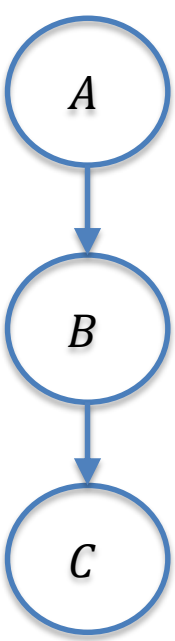
Equivalent Models?

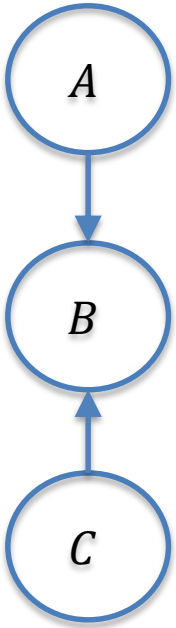




Do these models represent the same independence relations?

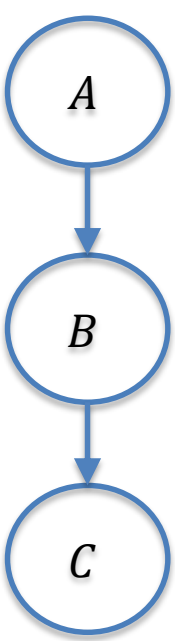
Equivalent Models?

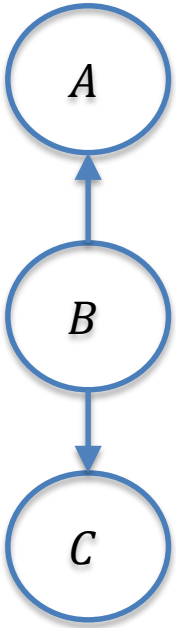




Do these models represent the same independence relations?

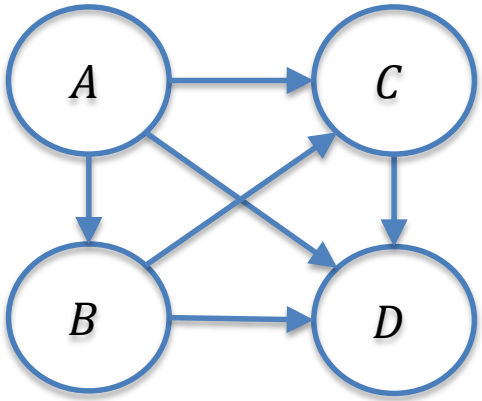
Equivalent Models?





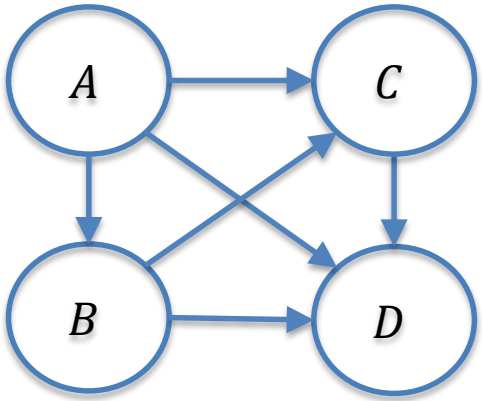
Do these models represent the same independence relations?

I-Maps



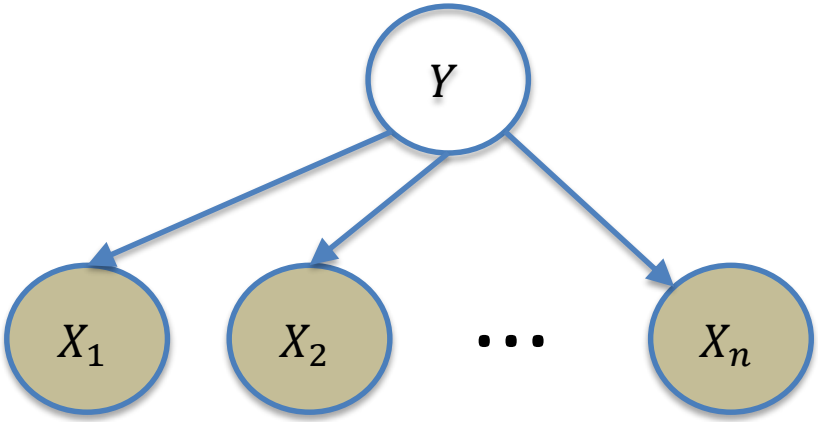
What independence relations does this model imply?

I-Maps



$I(G) = \emptyset$, this is an I-map for any joint distribution on four variables!

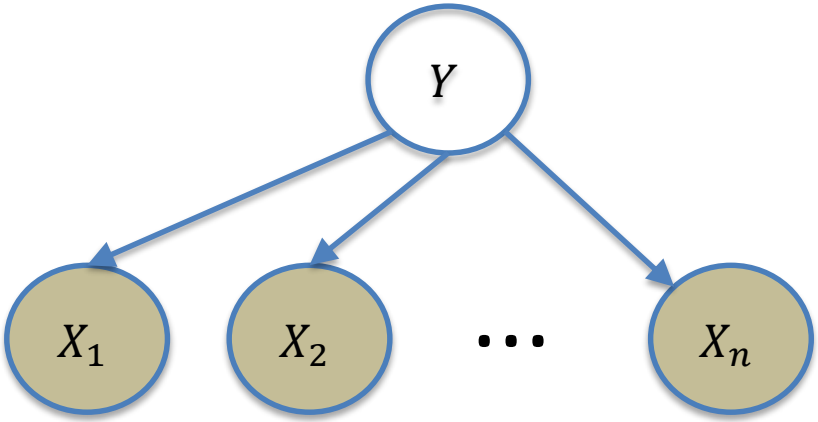
Naïve Bayes



$$p(y, x_1, \ldots, x_n) = p(y)p(x_1|y) \ldots p(x_n|y)$$

- In practice, we often have variables that we observe directly and those that can only be observed indirectly

Naïve Bayes



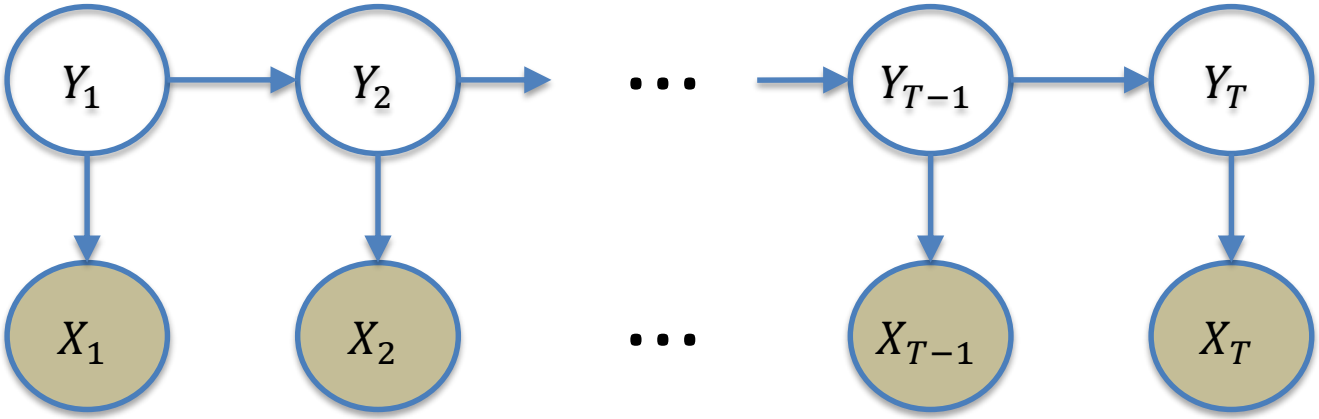
$$p(y, x_1, \ldots, x_n) = p(y)p(x_1|y) \ldots p(x_n|y)$$

- This model assumes that $X_1, \ldots, X_n$ are independent given Y, sometimes called naïve Bayes

Example: Naïve Bayes

- Let Y be a binary random variable indicating whether or not an email is a piece of spam
- For each word in the dictionary, create a binary random variable X_i indicating whether or not word i appears in the email
- For simplicity, we will assume that $X_1, \ldots, X_n$ are independent given Y
- How do we compute the probability that an email is spam?

Hidden Markov Models



$$p(x_1, \ldots, x_T, y_1, \ldots, y_T) = p(y_1)p(x_1|y_1)\prod_t p(y_t|y_{t-1})p(x_t|y_t)$$

- Used in coding, speech recognition, etc.
- Independence assertions?