CS 6347

Lecture 2

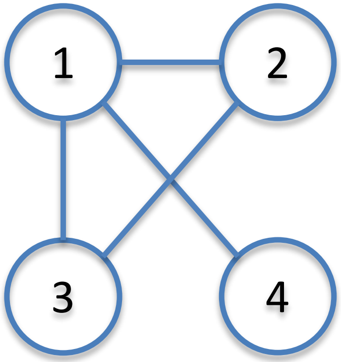
Bayesian Networks

Recap

- Last time:
  - Course logistics
  - Review of basic probability
- Today:
  - Independent set example
  - What makes one probability distribution “better” than another?
  - Bayesian networks

Graphs & Independent Sets

- A graph $G = (V, E)$ is defined by a set of vertices V and a set of edges $E \subseteq V \times V$ (i.e., edges correspond to pairs of vertices)

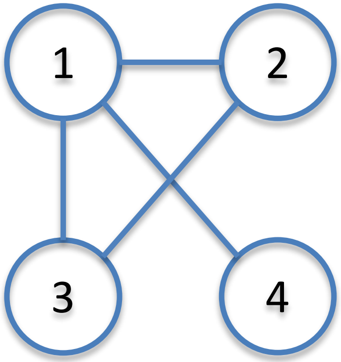


$$V = \{1,2,3,4\}$$

$$E = \{(1,2), (1,3), (2,3), (1,4)\}$$

Graphs & Independent Sets

- A set $S \subseteq V$ is an independent set if there does not exist an edge in E joining any pair of vertices in S

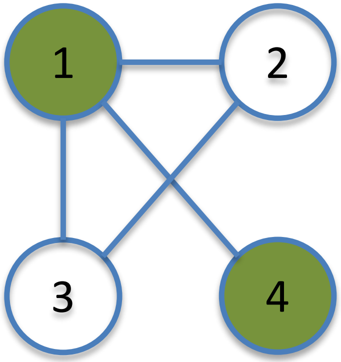


$$V = \{1,2,3,4\}$$

$$E = \{(1,2),(1,3),(2,3),(1,4)\}$$

Graphs & Independent Sets

- A set $S \subseteq V$ is an independent set if there does not exist an edge in E joining any pair of vertices in S



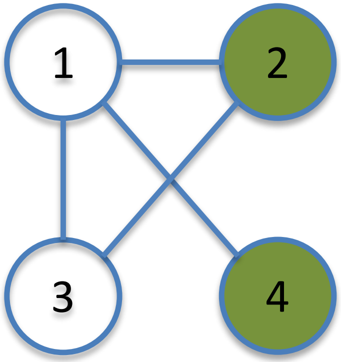
$$V = \{1,2,3,4\}$$

$$E = \{(1,2), (1,3), (2,3), (1,4)\}$$

{1,4} is not an independent set!

Graphs & Independent Sets

- A set $S \subseteq V$ is an independent set if there does not exist an edge in E joining any pair of vertices in S



$$V = \{1,2,3,4\}$$

$$E = \{(1,2),(1,3),(2,3),(1,4)\}$$

{2,4} is an independent set

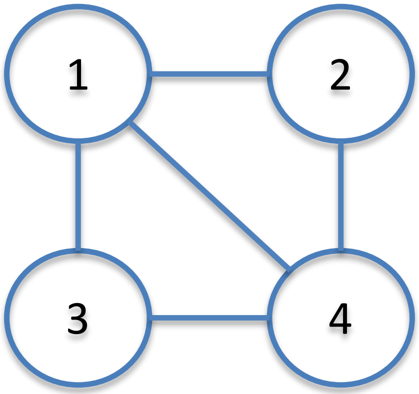
Example: Independent Sets

- Let Ω be the set of all vertex subsets in a graph $G = (V, E)$
- Let p be the uniform probability distribution over all independent sets in Ω
- Define for each $v \in V$ **and each subset of vertices** ω

$$X_v(\omega) = 1, \qquad \text{if } v \in \omega \text{ and}$$
$$X_v(\omega) = 0, \qquad \text{otherwise}$$

- $p(X_v = 1)$ is the fraction of all independent sets in G containing v
- $p(x_1, \ldots, x_n) \neq 0$ if and only if the x's define an independent set

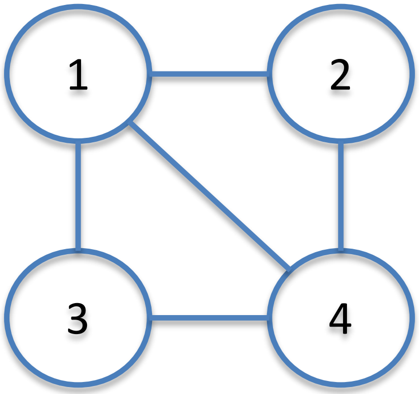
Example: Independent Sets



Consider the graph on the left, with the sample space and probabilities from the last slide

- $p(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1) = ?$
- $p(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) = ?$
- $p(X_2 = 1) = ?$

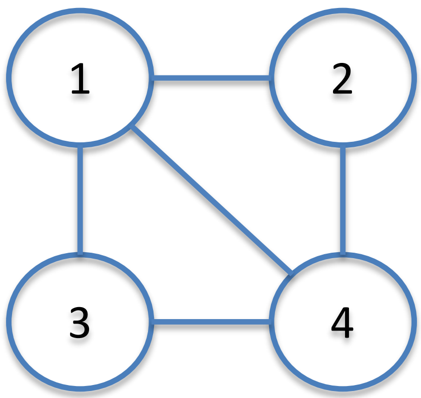
Example: Independent Sets



Consider the graph on the left, with the sample space and probabilities from the last slide

- $p(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1) = 0$
- $p(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) = ?$
- $p(X_2 = 1) = ?$

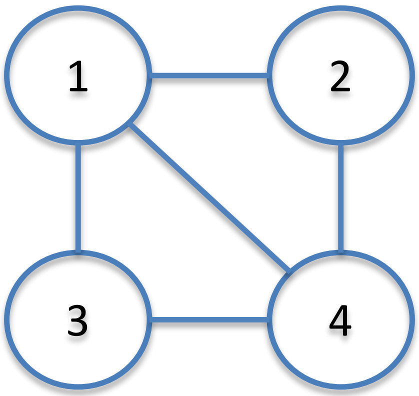
Example: Independent Sets



Consider the graph on the left, with the sample space and probabilities from the last slide

- $p(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1) = 0$
- $p(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) = 1/6$
- $p(X_2 = 1) = ?$

Example: Independent Sets



Consider the graph on the left, with the sample space and probabilities from the last slide

- $p(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1) = 0$
- $p(X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0) = 1/6$
- $p(X_2 = 1) = 1/3$

Example: Independent Sets

- How large of a table is needed to store an arbitrary distribution $p(X_V)$ over subsets of a given graph $G = (V, E)$?

Example: Independent Sets

- How large of a table is needed to store an arbitrary distribution $p(X_V)$ over subsets of a given graph $G = (V, E)$?

$$\mathbf{2^{|V|}\text{-}1}$$

Computational Issue #1

- How much storage space is required to represent a given joint probability distribution?
  - Can we do better than the worst case?
  - What properties of the joint distribution affect this number?

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 mutually independent random variables, then

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

- How much information is needed to store the joint distribution?

?

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 mutually independent random variables, then

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

- How much information is needed to store the joint distribution?

$\boldsymbol{n}$ numbers

- This model is boring: knowing the value of any one variable tells you nothing about the others

Structured Distributions

- Consider a general joint distribution $p(X_1, \ldots, X_n)$ over binary valued random variables

- If $X_1, \ldots, X_n$ are mutually, conditionally independent given a different random variable Y, then

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

  and

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

- These models turn out to be surprisingly powerful, despite looking nearly identical to the previous case!

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

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?

$$\mathbf{2n - 1}$$

- This distribution is chain-like

Computational Issue #2

- Given a joint probability distribution (as a table), how complicated is it to compute individual probabilities?
  - Computing $p(X_1 = x_1)$ from a joint probability distribution $p(X_1 = x_1, \ldots, X_n = x_n)$ is one type of statistical inference

Marginal Distributions

- Given a joint distribution $p(X_1, \ldots, X_n)$, the marginal distribution over the i^{th} random variable is given by

$$p_i(X_i = x_i) = \sum_{x_1} \sum_{x_2} \ldots \sum_{x_{i-1}} \sum_{x_i+1} \ldots \sum_{x_n} p(X_1 = x_1, \ldots, X_n = x_n)$$

- In general, marginal distributions are obtained by fixing some subset of the variables and summing out over the others
  - This can be an expensive operation!

Inference/Prediction

- Given fixed values of some subset, E, of the random variables, compute the conditional probability over the remaining variables, S

$$p(X_S | X_E = x_E) = \frac{p(X_S, X_E = x_E)}{p(X_E = x_E)}$$

- This involves computing the marginal distribution $p(X_E = x_E)$, so we refer to this as marginal inference

Inference/Prediction

- Given fixed values of some subset, E, of the random variables, compute the most likely assignment of the remaining variables, S

$$\underset{x_S}{\operatorname{argmax}}\, p(X_S = x_s | X_E = x_E)$$

- This is called maximum a posteriori (MAP) inference
- We don't need to do marginal inference to compute the MAP assignment, why not?

Computational Issues

- The amount of storage and the complexity of statistical inference are both affected by the independence structure of the joint probability distribution
  - More independence means easier computation and less storage
  - Want models that somehow make the underlying independence assumptions explicit, so we can take advantage of them (expensive to check all of the possible independence relationships)

Bayesian Networks

- A **Bayesian network** is a directed graphical model that represents independence relationships of a given probability distribution
  - Directed acyclic graph (DAG), $G = (V, E)$
    - Edges are still pairs of vertices, but the edges $(1,2)$ and $(2,1)$ are now distinct in this model
  - One node for each random variable
  - One conditional probability distribution per node
  - Directed edge represents a direct statistical dependence

Bayesian Networks

- A **Bayesian network** is a directed graphical model that represents 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

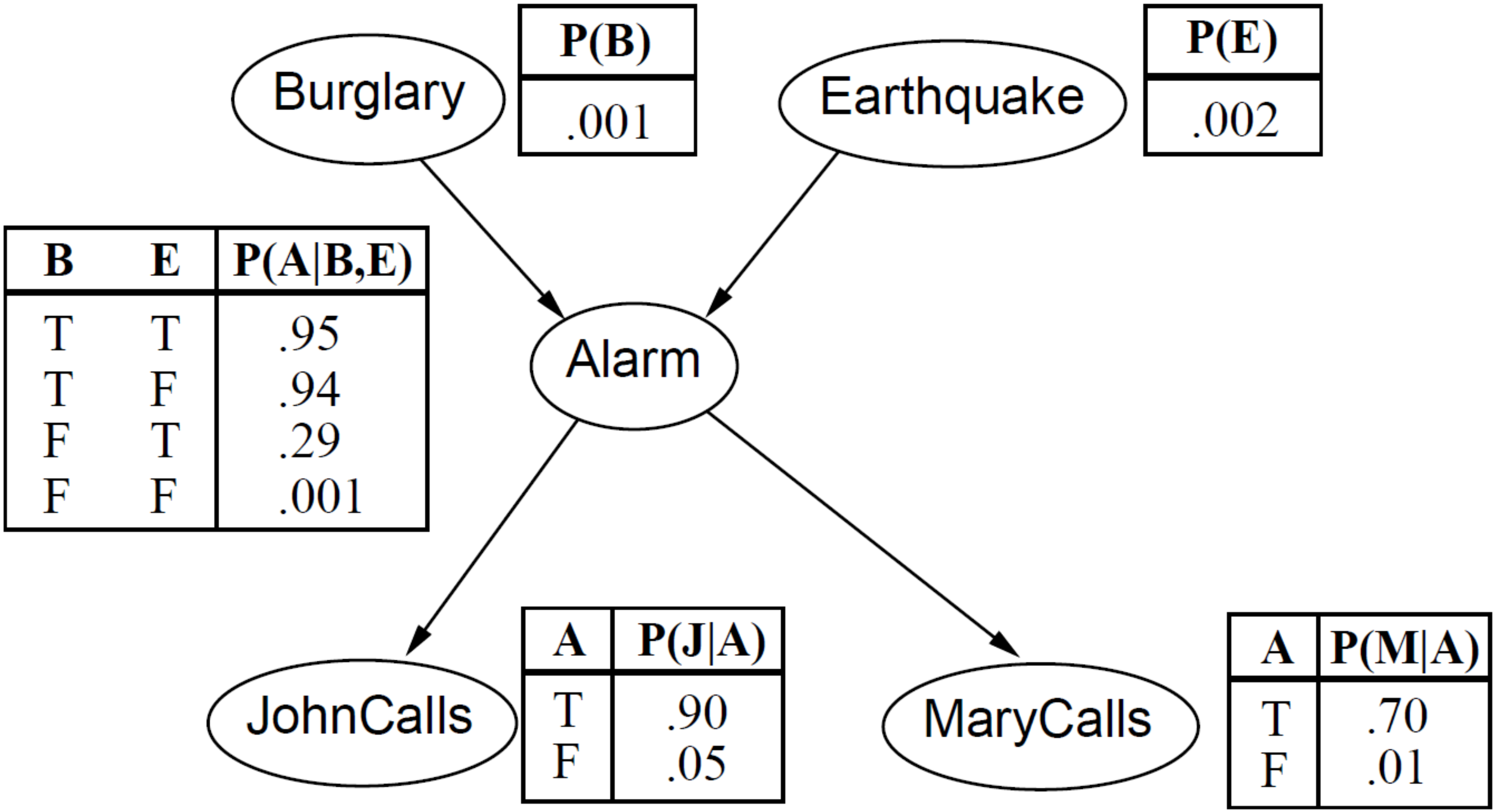
$$p(x_1, \ldots, x_n) = \prod_i p(x_i | x_{parents(i)})$$

Directed Chain

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

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_{n-1} \rightarrow X_n$$

An Example



from Artificial Intelligence: A Modern Approach