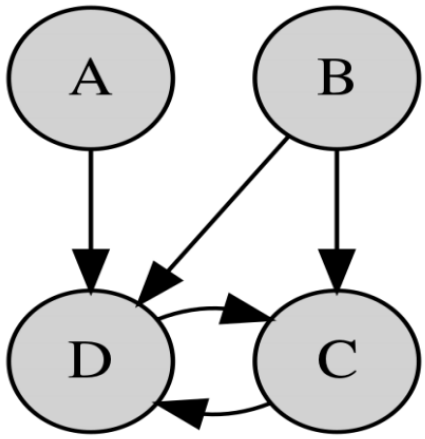
Math189R The Math of Big Data
Probabilistic Graphical Models

1 Probabilistic Graphical Models

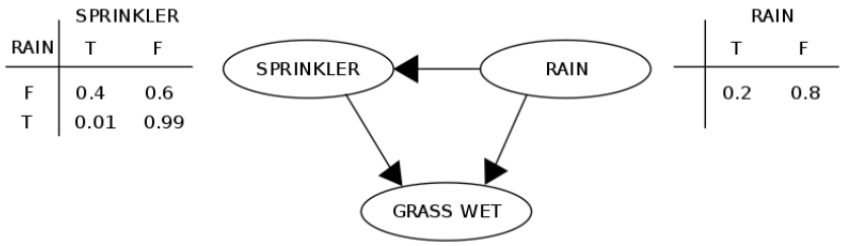
- A probabilistic graphical model (PGM) is a probabilistic model for which a graph expresses the conditional dependence structure between random variables.
- They are commonly used in probability theory, statistics—particularly Bayesian statistics—and machine learning.
- An example of a graphical model. Each arrow indicates a dependency. In this example: D depends on A, B, and C; and C depends on B and D; whereas A and B are each independent.



- Generally, probabilistic graphical models use a graph-based representation as the foundation for encoding a distribution over a multi-dimensional space and a graph that is a compact or factorized representation of a set of independences that hold in the specific distribution.
- Two branches of graphical representations of distributions are commonly used, namely, Bayesian networks and Markov random fields. Both families encompass the properties of factorization and independences, but they differ in the set of independences they can encode and the factorization of the distribution that they induce.

2 Bayesian Networks

- Example of a simple Bayesian network: Two events can cause grass to be wet - an active sprinkler or rain. Rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler usually is not active). This situation can be modeled with a Bayesian network (shown to the right). Each variable has two possible values, T (for true) and F (for false).



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- The joint probability distribution is given by

$$\Pr(G, S, R) = \Pr(G|S, R)\Pr(S|R)\Pr(R)$$

- Key: Once you have written out this probability function according the graphic model, you can answer basically "any" probability questions!
- Recall the chain rule for random variables. For two random variables, this is

$$\Pr(X, Y) = \Pr(X|Y)\Pr(Y)$$

  Extended to n random variables, this becomes:

$$\Pr(X_n, \ldots, X_1) = \Pr(X_n|X_{n-1}, \ldots, X_1)\Pr(X_{n-1}|X_{n-2}, \ldots, X_1)\ldots\Pr(X_1)$$

  We can write this as

$$\Pr(\bigcap_{k=1}^{n} X_i) = \prod_{k=1}^{n} \Pr(X_k|\bigcap_{j=1}^{k-1} X_j)$$

- For instance, if we want to answer a question like "What is the probability that it is raining, given the grass is wet?", we can use the conditional probability formula as follows:

$$\Pr(R = T|G = T) = \frac{\Pr(G = T, R = T)}{\Pr(G = T)} = \frac{\sum_{S\in\{T,F\}} \Pr(G = T, S, R = T)}{\sum_{S,R\in\{T,F\}} \Pr(G = T, S, R)}$$

  Then the numerical results (subscripted by the associated variable values) are

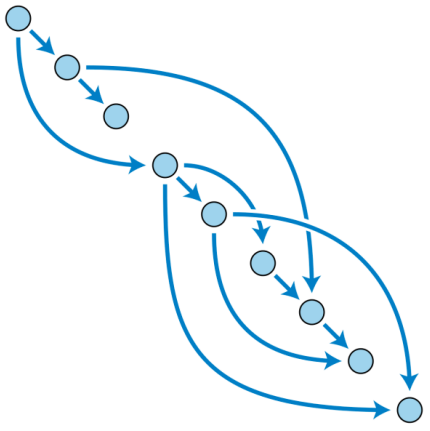
$$\Pr(R = T|G = T) = \frac{0.00198_{TTT} + 0.1584_{TFT}}{0.00198_{TTT} + 0.288_{TTF} + 0.1584_{TFT} + 0.0_{TFF}} = \frac{891}{2491} = 35.77\%$$

3 Review of Independence

- Let A and B be random variables taking values $a \in \mathcal{A}$ and $b \in \mathcal{B}$. Then A and B are independent if
$$p(a,b) = p(a)p(b)$$
for all a and b. For x_i and x_j, we use the notation $x_i \perp x_j$.
- Another way of conceptualizing this is to observe that if $p(a,b) = p(a)p(b)$, then
$$p(a|b) = \frac{p(a,b)}{p(b)} = \frac{p(a)p(b)}{p(b)} = p(a)$$
so the knowledge of B does not affect A.
- Useful fact: $a \perp b$ iff $p(a,b) = f(a)g(b)$ for some functions f and g.
- A is conditionally independent of B given C if
$$p(a,b|c) = p(a|c)p(b|c)$$
Alternatively, we can write
$$p(a|b,c) = p(a|c)$$
- In Markov chains, the Markov assumption is
$$p(x_j|x_{j-1}, x_{j-2}, \ldots, x_1) = p(x_j|x_{j-1})$$

4 Directed Acyclic Graphs (DAGs)

- A DAG is a finite directed graph with no directed cycles. That is, it consists of finitely many vertices and edges (also called arcs), with each edge directed from one vertex to another, such that there is no way to start at any vertex v and follow a consistently-directed sequence of edges that eventually loops back to v again. Equivalently, a DAG is a directed graph that has a topological ordering, a sequence of the vertices such that every edge is directed from earlier to later in the sequence.



- DAG models use a factorization of the joint distribution

$$p(x_1, x_2, \ldots, x_d) = \prod_{j=1}^{d} p(x_j | x_{pa(j)})$$

  where $pa(j)$ are the "parent" nodes of node j. This assumes the Markov property, which states that in general,

$$p(x_j | x_{1:j-1}) = p(x_j | x_{pa(j)})$$

- Rather than factoring by variables, it is also possible to factor the distribution into fully-connected "blocks" like

$$p(x) = \prod_{b} p(x_b | x_{pa(b)})$$

- All conditional independences implied by a DAG can be read from the graph. In particular, A and B are conditionally independent given C if "D-separation blocks all undirected paths in the graph from any variable in A to any variable in B.
- The rules of d-separation are intuitive in a simple model of gene inheritance:
  - Each person has a singular number, called a "gene".
  - If you have no parents, your gene is a random number.
  - If you have parents, your gene is the sum of your parents plus noise.
  - Genes of people are independent if knowing one says nothing about the other.
  - Your gene is dependent on your parents.
  - Your gene is independent of your friends.
  - Genes of people can be conditionally independent given a third person - for example, knowing your grandparent's gene tells you something about yours, but isn't useful if you know your parent's gene.