Bayesian Networks

Today

- Why do we need Bayesian networks (Bayes nets)?
- “Factoring” the joint probability
- Conditional independence
- What is a Bayes net?
  - Mostly, we discuss discrete r.v.’s
- What can you do with a Bayes net?

Exponential growth of joint probability tables

- The size of a table representing the joint probabilities of discrete r.v.'s is exponential in the number of r.v.'s.

recur	not recur
0.24	0.76

	recur	not recur
big cells	0.16	0.33
little cells	0.08	0.43

	recur		not recur	
	rough cells	smooth cells	rough cells	smooth cells
big cells	0.06	0.10	0.17	0.16
little cells	0.02	0.06	0.18	0.25

- Large joint probability tables are:
  - Awkward or impossible to store and maintain, if there are many r.v.'s.
  - Difficult to reason about or visualize.
  - Computing marginal or conditional probabilities requires intractably large summations.
  - Difficult to learn. (Because there are so many free parameters!)
- To represent an arbitrary joint probability distribution, an exponentially-large table is *necessary*.
- But. . .

- The world is not arbitrarily complex!
  - “Effects” have a limited number of “causes”.
  - R.v.’s have limited relationships with other r.v.’s.
- Bayes nets are a technique for representing and reasoning about “big” joint probability distributions in a compact way. They rely on two things:
  1. Writing the joint probability as a produce of conditional probabilities.
  2. Simplifying based on conditional independence.

Rewriting the joint as a product of conditionals

- By the definition of conditional probability, any joint probability can be rewritten as

$$
\begin{aligned}
& P(X_1, X_2, \ldots, X_m) \\
= \quad & P(X_1 | X_2, \ldots, X_m) P(X_2, \ldots, X_m) \\
= \quad & P(X_1 | X_2, \ldots, X_m) P(X_2 | X_3, \ldots, X_m) P(X_3, \ldots, X_m) \\
= \quad & P(X_1 | X_2, \ldots, X_m) P(X_2 | X_3, \ldots, X_m) \cdots P(X_{m-1} | X_m) P(X_m)
\end{aligned}
$$

- We can "factor" the joint probability into a product of conditional probabilities in different ways. Another one is:

$$
\begin{aligned}
& P(X_1, X_2, \ldots, X_m) \\
= \quad & P(X_m | X_1, \ldots, X_{m-1}) P(X_1, \ldots, X_{m-1}) \\
= \quad & P(X_m | X_1, \ldots, X_{m-1}) P(X_{m-1} | X_1, \ldots, X_{m-2}) \cdots P(X_2 | X_1) P(X_1)
\end{aligned}
$$

Example

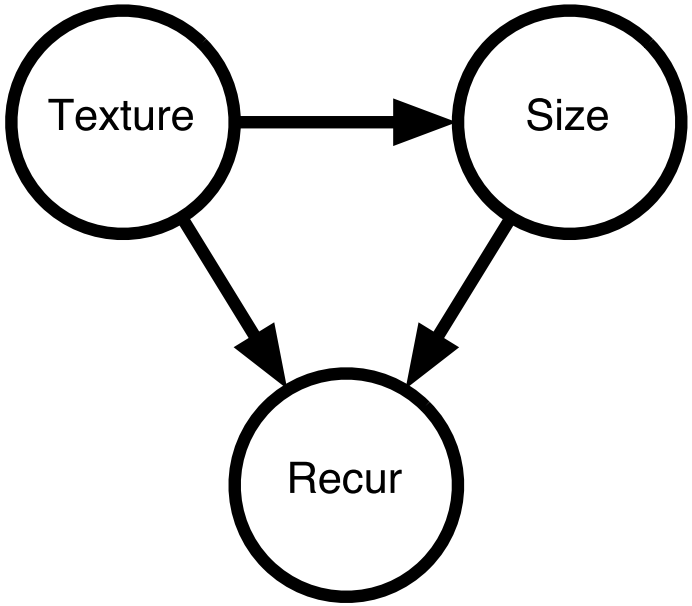
- $Recur =$ whether or not the cancer recurs.
- $Size =$ whether tumor cells are big or little.
- $Texture =$ whether tumor cells are rough or smooth.

$$
\begin{aligned}
& P(Recur, Size, Texture) \\
= \quad & P(Recur|Size, Texture)P(Size|Texture)P(Texture) \\
= \quad & P(Recur|Size, Texture)P(Texture|Size)P(Size) \\
= \quad & P(Size|Recur, Texture)P(Recur|Texture)P(Texture) \\
= \quad & P(Size|Recur, Texture)P(Texture|Recur)P(Recur) \\
= \quad & P(Texture|Recur, Size)P(Recur|Size)P(Size) \\
= \quad & P(Texture|Recur, Size)P(Size|Recur)P(Size)
\end{aligned}
$$

- With m variables, there are $m!$ ways of doing this.

Viewing it graphically

- A factorization can be depicted graphically.
  - Nodes correspond to r.v.'s.
  - Arcs to a r.v. come from the r.v.'s upon which it is conditioned.
- $P(Recur|Size, Texture)P(Size|Texture)P(Texture)$:



Space savings? None yet...

- If we imagine the terms, P, are represented by tables, there is no advantage to rewriting.

$$\underbrace{P(Recur, Size, Texture)}_{2^3=8 \text{ cells}}$$

$$= \underbrace{P(Recur|Size, Texture)}_{2^3=8 \text{ cells}} \underbrace{P(Size|Texture)}_{2^2=4 \text{ cells}} \underbrace{P(Texture)}_{2 \text{ cells}}$$

Conditional independence — a generalization of independence

- Let X, Y, and Z each represent one or more r.v.'s
- X is *conditionally independent* of Y *given* Z if

$$P(X|Y,Z) = P(X|Z)$$

  That is, if we know Z, knowing Y too doesn't help us predict X any better.

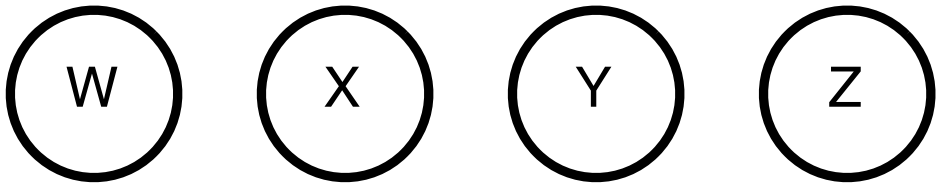
- Examples: Is there conditional independence or not?
  - $X =$ die roll, $Y =$ roll is even, $Z =$ roll is odd.
  - $X =$ die roll, $Y =$ roll is even, $Z =$ roll is prime.
  - $X =$ coin flip, $Y =$ another coin flip, $Z = X$ and Y match.
- Independence of X and Y is conditional independence with $Z = \emptyset$.

Taking advantage of conditional independencies

- If there are conditional independencies between r.v.'s, and if we factor the joint correctly, we can represent the joint more compactly.
- Example, if W, X, Y, and Z are independent binary r.v.s:

$$\underbrace{P(W,X,Y,Z)}_{16 \text{ cells}} = P(W|X,Y,Z)P(X|Y,Z)P(Y|Z)P(Z)$$

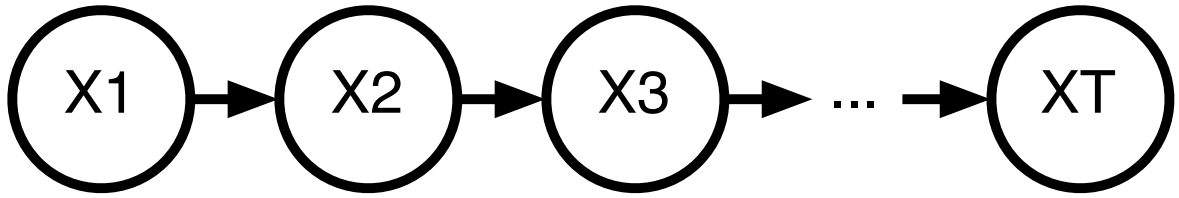
$$= \underbrace{P(W)}_{2 \text{ cells}}\underbrace{P(X)}_{2 \text{ cells}}\underbrace{P(Y)}_{2 \text{ cells}}\underbrace{P(Z)}_{2 \text{ cells}}$$



- For m independent binary random variables, the space requirement goes from 2^m to $2m$.

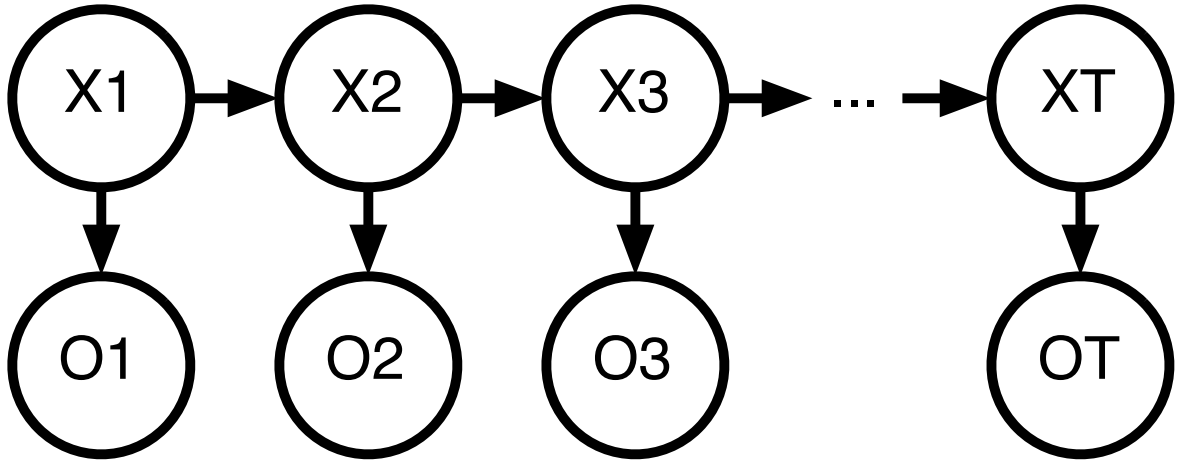
Example: Markov chains

- Let X_t denote the state of a dynamical system at time t.
- The Markov assumption: X_{t+1} is conditionally independent of $X_0, \ldots, X_{t-1}$ given X_t.
- $P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1) \cdots P(X_T|X_{T-1})$



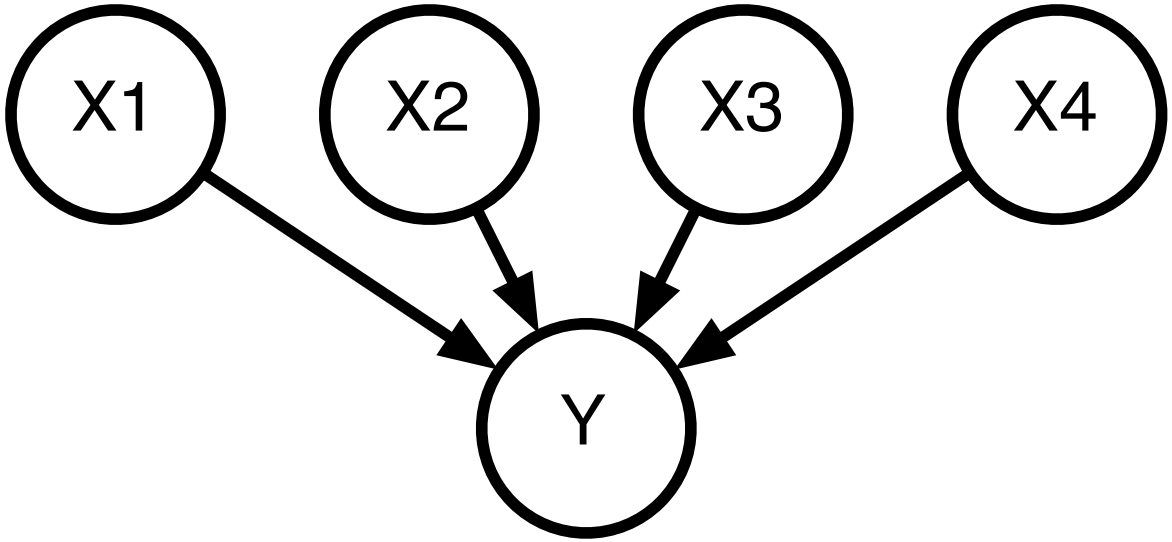
Example: Hidden markov model

- Let X_t denote the unobserved state of a dynamical system at time t.
- Let O_t denote an observation made at time t.
- $P(X_1, O_1, \ldots, X_T, O_T) =$
  $P(X_1)P(O_1|X_1)P(X_2|X_1)P(O_2|X_2)\cdots P(X_T|X_{T-1})P(O_T|X_T)$



Example: Prediction problems

- Let $X_1, \ldots, X_m$ represent input features and Y the r.v. to be predicted.
- We might assume the X_i are independent, but Y depends on them.

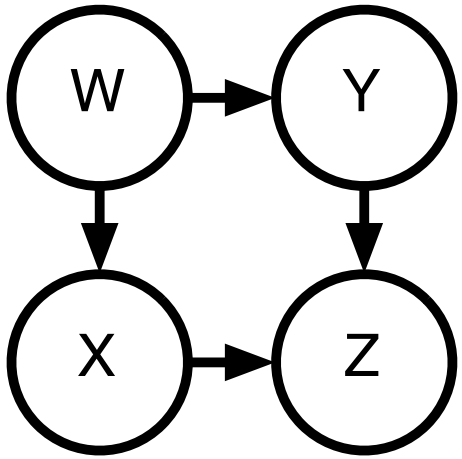


What is a Bayes net?

- It is a representation for a joint probability in terms of conditional probabilities.

$$P(W, X, Y, Z) = P(W)P(Y|W)P(X|W)P(Z|Y, X)$$

- The corresponding graphical structure is a directed, acyclic graph.



What is a Bayes net (2)?

- For discrete r.v.'s, conditional probabilities can be represented as tables (CPTs) or in more compact forms, such as trees.
- Continuous r.v.'s can be included too, with conditional probabilities represented, e.g., parametrically.

What do we do with Bayes nets?

Mainly, we compute conditional probabilities after observing some data.

- Diagnosis: What's the probability of cancer, given symptoms x, y, z?
- Prediction/causal reasoning: What's the probability of recurrence, given measurements x, y, z?
- What's the probability of a UFO sighting being a true alien spaceship?

Advantages of Bayes nets

- Represent the joint compactly.
- Marginal and conditional probabilities can be computed more efficiently than by naive summing-out over the joint. (at least if all r.v.'s are discrete)
- Provides a visual representation for the relationships between variables.
- Generalizes the prediction problem, allowing other forms of reasoning.