Oct 1~~4~~ ~~Multivariate Probability~~

16 Bayesian Networks

Suppose my car won't start. Why not? How would I diagnose the problem?

Announcements and To-Dos

Announcements:

1. Working on Practicum official write-up:
   1.1 Implement and compare Simulated Annealing and Djikstra's algorithm on "The traveling salesman" problem.
   1.2 Short-ish paper on ethics and AI: think about a topic where AI might have important impacts on the world, get some sources, and write about the costs and benefits of certain AI tuning and implementation choices! Consider e.g. smartphones, GPS, game-playing, cars, USPS routes, national security, whatever interests you!

Last time we learned:

1. Some multivariate probability!

Probability Roundup

We have added a *multivariate* emphasis to our past understanding of probability. This means a few crucial distinctions are in play:

1. **Definition:** The *joint* distribution over a set of random variables $X_1, X_2 \dots X_n$ specifies a real number for each outcome $P(X_1 = x_1, X_2 = x_2, \dots X_n = x_n)$, often just written as $P(x_1, x_2, \dots x_n)$.

   In short, it tracks **and** probabilities for the random variables.

   Definition: The *marginal* distribution for X is $P(X = x)$ ignoring other variables.

   It tabulates the probability for X summing over all *other* random variables (group by x, then sum probabilities.)

   $P(X = \text{True}) = p$
   $P(X = \text{False}) = 1 - p$

   Definition: The *conditional distribution* of X given Y is $P(X|Y)$.

   $\begin{cases} P(X=T \mid Y=T) = q \\ P(X=F \mid Y=T) = 1-q \end{cases}$
   $\begin{cases} P(X=T \mid Y=F) = q_2 \\ P(X=F \mid Y=F) = 1-q_2 \end{cases}$

   It tabulates the probability for X summing over levels of y: (group by y, then list the x probabilities - may not need to sum, but do need to normalize!)

Independence

1. **Definition:** X and Y are *independent* if for all events x and y, $P(x,y) = P(x)P(y)$. We write $X \perp\!\!\!\perp Y$.

   We interpret this as: X and Y are in no way related. Computationally, the joint distribution *factoring* into $P(x)$ times $P(y)$ is often *very* convenient.

2. **Definition:** X and Y are *conditionally independent* if for all events x and y, and z, $P(x,y|z) = P(x|z)P(y|z)$. We write $X \perp\!\!\!\perp Y|Z$.

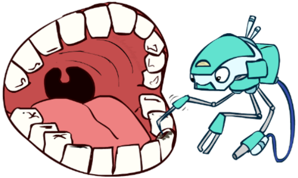
   We interpret this as: X and Y may be related, but they're only related *because* of Z. Once we know Z, then X and Y become unrelated outcomes.

   In probability we're often hesitant to think of this as Z "causing" two results X and Y, but that's one way to think about what this could mean.

Conditional Independence

Imagine now we have three variables:

1. Cavity (**C**)
2. Toothache (**T**)
3. Whether the dentist *finds* a cavity with a probe (**P**)



Conditional Independence

Imagine now we have three variables:

1. Cavity (**C**)
2. Toothache (**T**)
3. Whether the dentist *finds* a cavity with a probe (**P**)

If we have a cavity, the probability a probe catches it is not affected by whether we have a toothache, or:

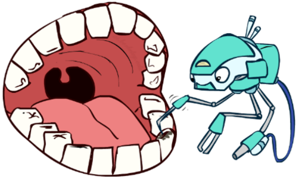
$$P(\underline{+p}|\boxed{+c,}\cancel{+t}) = P(+p|+c)$$

This also holds if we don't have a cavity:

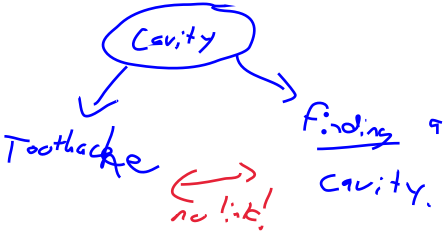
$$P(\underline{+p}|\boxed{\neg c,}\cancel{+t}) = P(+p|\neg c)$$

Or: probe is *conditionally* independent of toothache when given the presence of a cavity... because the cavity causes each!

$$P(p|t,c) = P(p|c)$$



Bayesian Networks



Definition: A *Bayesian Network* is:

arrows for edges

- A directed acyclic graph (DAG).

  no loops

- Shows a "flow" of cause and effects
- The point of a Bayes net is to **represent full joint probability distributions** *and*
- encode an interrelated set of **conditional independence** and related **probability statements**

  between "T" & "P"

Example: The cavity/toothache/probe example from the prior slide would be written as:

Bayesian Networks

Example: Represent the full joint distribution for $P(traffic, umbrella, rain)$.

1. "Trivial" decomposition:

$$P(T,U,R) = P(U \mid TR)\, \underbrace{P(TR)}_{P(T \mid R) \cdot P(R)}$$

2. Conditional Independence:

3. Visually:

Bayesian Networks

Example: Represent the full joint distribution for $P(traffic, umbrella, rain)$.

1. "Trivial" decomposition:

$$P(T, R, U) = P(U|TR)P(TR) = P(U|TR)P(T|R)P(R)$$

   ...but this is kind of useless: both pedestrians using umbrellas and drivers driving slowly are responding to the weather, so $P(U|TR)$ feels awkward.

2. Conditional Independence:

3. Visually:

Bayesian Networks

Example: Represent the full joint distribution for $P(traffic, umbrella, rain)$.

1. "Trivial" decomposition:

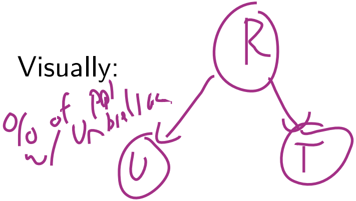
$$P(T, R, U) = P(U|TR)P(TR) = P(U|TR)P(T|R)P(R)$$

...but this is kind of useless: both pedestrians using umbrellas and drivers driving slowly are responding to the weather, so $P(U|TR)$ feels awkward.

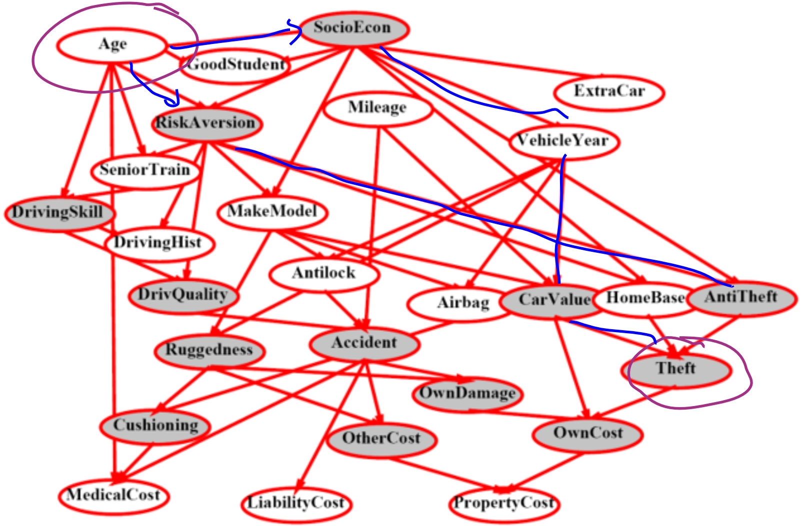
2. Conditional Independence: If we assume that U and T are *conditionally independent* given R, we can simplify more

$$P(T, R, U) = P(U|TR)P(TR) = \underline{P(U|\cancel{T}R)}P(T|R)P(R) = \underline{P(U|R)}P(T|R)P(R)$$

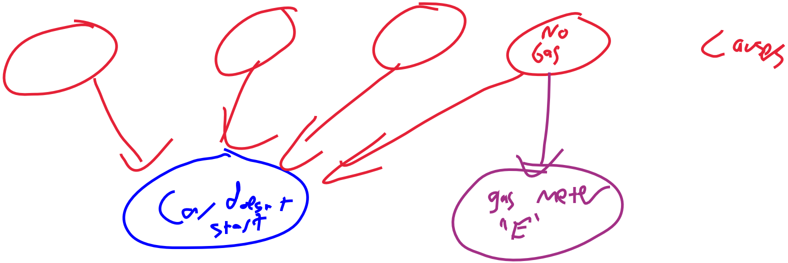
3. Visually:



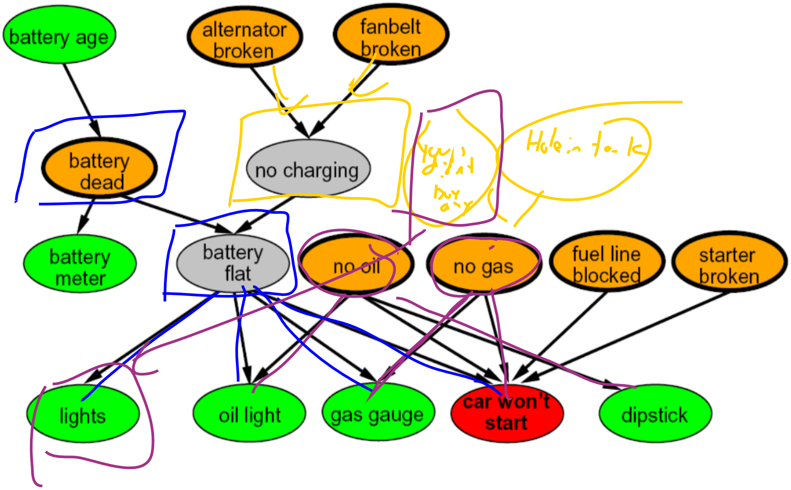
Bayesian Networks Examples: Insurance



Bayesian Networks: Your Car Won't Start



Bayesian Networks: Your Car Won't Start

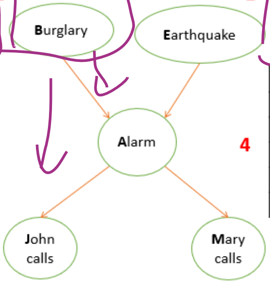


P(you didn't buy gas AND lights not on)

Example: Two steps of alarm:

P(B)	
true	false
0.001	0.999

1



1

P(E)	
true	false
0.002	0.998

4

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

marginal

$P(A \mid B, E) = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \end{cases}$

2 levels   2 levels

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

2

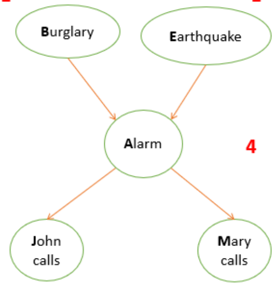
A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

Two big ideas: what should $\underbrace{P(B|A,\cancel{J})}_{P(B|A)}$ be? What should $\underbrace{P(J|A,\cancel{M})}_{P(J|A)}$ be?

Example: Two steps of alarm:

P(B)	
true	false
0.001	0.999

1



1

P(E)	
true	false
0.002	0.998

4

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

2

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

Two big ideas: what should $P(B|A, J)$ be? What should $P(J|A, M)$ be?

We want various pairs of events like (J and B) or (J and M) to be *conditionally independent*, given A. Once the alarm has been triggered *regardless of what triggered it* we have a probability that John calls, and unrelated probability that Mary calls

Bayes Nets: Independence

As with any probability, the joint distribution of random variables $X_1, X_2, \dots X_n$ can factor:

And

$P(U \mid T R) \cdot P(T R)$

$$
\begin{aligned}
& P(X_1 = x_1, X_2 = x_2, \dots X_n = x_n) \\
&= P(x_n, x_{n-1} \dots x_1) \\
&= P(x_n | x_{n-1} \dots x_1) P(x_{n-1}, x_{n-2} \dots x_1) \\
&= P(x_n | x_{n-1} \dots x_1) P(x_{n-1} | x_{n-2} \dots x_1) P(x_{n-2}, \dots x_1) \\
&= \vdots \\
&= P(x_n | x_{n-1} \dots x_1) P(x_{n-1} | x_{n-2} \dots x_1) P(x_{n-2} \dots x_1) \dots P(x_2 | x_1) P(x_1) \\
&= \prod_{i=1}^{n} P(x_i | \underbrace{x_{i-1}, x_{i-2}, \dots x_1}_{\text{all smaller indices}})
\end{aligned}
$$

... but in a Bayes network, most of the "conditioned" events are conditionally independent!

Bayes Nets: Independence

X_1 parent to X_2 X_3:

$P(X_3 | X_2 \boxed{X_1})$ $P(X_2 | \boxed{X_1})$

So in a Bayes nets, we can simplify. Parents would not only be neighbors on the graph, but also now the events we think of "causing" their successors or children.

If we are careful to write the nodes in an order such that parents precede the children then, then the parents of node X_i are in the set of prior nodes $\{X_1, X_2, X_3, \ldots X_{i-1}\}$, and:

B E A

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$= \prod_{i=1}^{n} P(x_i | x_{i-1}, x_{i-2}, \ldots x_1) \qquad \boxed{= \prod_{i=1}^{n} P(x_i | \text{parents}(X_i))}$$

A M J

The last statement is the crux of the Bayesian network: **each node ~~if~~ is conditionally independent of its other predecessors, given its parents**

Bayes Nets: Traffic

Consider a simple network of rain and traffic.

$P(R)$

+r	1/4
¬r	3/4

$P(T|R)$

+r→	+t	3/4
	¬t	1/4
¬r→	+t	1/2
	¬t	1/2

What is $P(+r, \neg t)$?

$P(A \mid B) \cdot P(B)$

$P(R,T) = P(T \mid R) \cdot P(R)$

Bayes Nets: Traffic

Consider a simple network of rain and traffic.

$P(R)$

+r	1/4
¬r	3/4

$P(T|R)$

+r →	+t	3/4
	¬t	1/4
¬r →	+t	1/2
	¬t	1/2

What is $P(+r, \neg t)$?

$$P(R, T) = P(T|R)P(R)$$

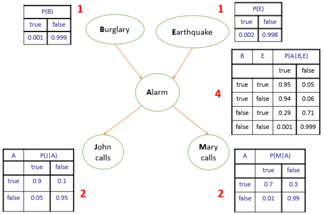
$$P(\underbrace{R = +r}_{\text{yes}}, \underbrace{T = \neg t}_{\text{no}}) = P(T = \neg t | R = +r)P(R = +r)$$

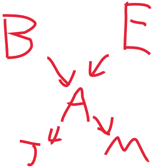
$$\left(\tfrac{1}{4}\right)\left(\tfrac{1}{4}\right)$$

$$\boxed{= 1/16}$$

Bayes Nets: the Alarm

Recall: Bayes nets encode joint distributions as the product of local conditionals:
$P(X_1 = x_1, X_2 = x_2, \dots X_n = x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i))$





What is the entire joint distribution? How would we perform calculations on it? We want

$$P(B, E, A, J, M)$$

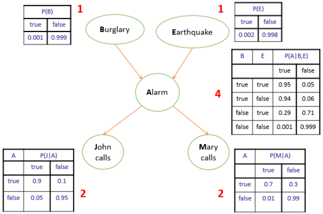
and in that order is nice, since everything can be conditioned on events *prior to it*.

Bayes Nets: the Alarm

$P(B, E, A, J, m)$ 32 entries

Recall: Bayes nets encode joint distributions as the product of local conditionals:

$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i))$



P(B)	
true	false
0.001	0.999

P(E)	
true	false
0.002	0.998

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

What is the entire joint distribution?

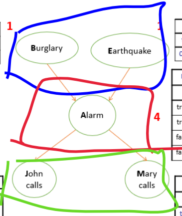
1. Events B and E are independent. So $P(B, E) = P(B)P(E)$. They're also class-conditionally independent *given* an alarm, but we don't even need that!
2. Event A depends on B and E. We know $P(A|B)$, $P(A|E)$, $P(A|BE)$.
3. Once we know A, we know the class-independent probabilities of $P(J|A), P(M|A)$. Due to their independence, $P(\underbrace{JM}_{\text{and}}|A) = P(J|A)P(M|A)$.

Bayes Nets: the Alarm

Everything factors!

P(B)	
true	false
0.001	0.999

P(E)	
true	false
0.002	0.998



B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

What is the entire joint distribution?

$$P(B, E, A, J, M)$$
$$= P(BE)P(AJM|BE) \qquad \text{Defn. Conditional}$$
$$= P(B)P(E)P(AJM|BE) \qquad \text{B \& E ind.}$$
$$= P(B)P(E)P(A|BE)P(JM|A\cancel{BE}) \qquad \text{Defn. Conditional}$$
$$= P(B)P(E)P(A|BE)P(J|A)P(M|A) \qquad \text{Class-cond. indep!}$$

$P(JM|A)$

In effect we've divided the graph into tiers: a top tier of B, E, a middle tier of A, and the last tier of J, M. The overall probability is just the product of the statements on each tier conditioned on the tier above it!

Bayes Nets: the Alarm

$$P(B, E, A, J, M) =$$

$$P(B)P(E)P(A|BE)P(J|A)P(M|A)$$

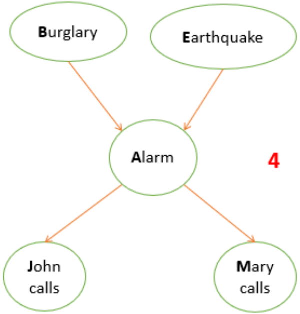
Example: What is
$P(B = True|J = True, M = True)$?

P(B)	
true	false
0.001	0.999

1

1

P(E)	
true	false
0.002	0.998



B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

4

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

2

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

Bayes Nets: the Alarm

$$P(B, E, A, J, M) =$$

$$P(B)P(E)P(A|BE)P(J|A)P(M|A)$$

Example: What is
$P(B = True|J = True, M = True)$?
By hand Solution:

$$P(B|JM) = \frac{P(BJM)}{P(JM)}$$

$P(BJM)$ happens 4 ways to add up: over both values of E and both values of A. $P(JM)$ happens 8 ways to add up: over both values of E, both values of A, and both values of B.

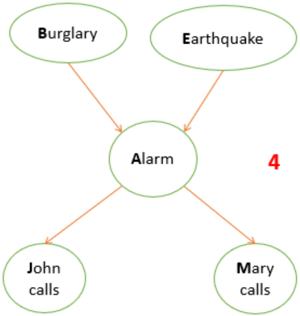
That's 8 outcomes to consider! But they're all quick and easy! A specific outcome is on the tables. E.g.
$P(B = T, E = F, A = T, J = T, M = T)$
is one we need, and is
$P(B = T)P(E = F)P(A = T|B = TE = F)P(J = T|A = T)P(M = T|A = T)$ or
$(.001)(.998)(.94)(.9)(.7)$

P(B)	
true	false
0.001	0.999

1

P(E)	
true	false
0.002	0.998

1



B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

4

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

2

Bayes Nets: the Alarm

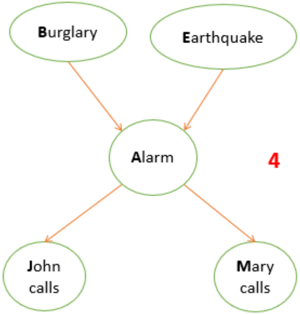
$$P(B, E, A, J, M) =$$

$$P(B)P(E)P(A|BE)P(J|A)P(M|A)$$

Example: What is
$P(B = True|J = True, M = True)$?

P(B)	
true	false
0.001	0.999

1



1

P(E)	
true	false
0.002	0.998

4

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

2

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

That's 8 outcomes to consider! But they're all quick and easy! A specific outcome is on the tables. E.g.
$P(B = T, E = F, A = T, J = T, M = T)$
is one we need, and is
$P(B = T)P(E = F)P(A = T|B = TE = F)P(J = T|A = T)P(M = T|A = T)$ or
$(.001)(.998)(.94)(.9)(.7)$

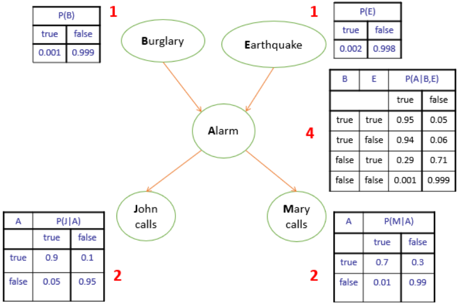
Bayes Nets: Space

Are we really saving space and time, here?

The full joint distribution on $n = 5$ nodes would take $2^5 = 32$ probabilities. We're specifying only 10.

Suppose the Bayes net in general has $k = 2$ parents per node. Then for $n = 5$ nodes we're only specifying $25 \cdot^{2} = 20$ probabilities at worst!

But what if $n = 30$ and $k = 5$? The Bayes net would require $n \cdot 2^k = 960$ probabilities. The full joint distribution holds $2^{30} \approx 1e9$ entries.



Moving Forward

- Next Week:
  1. Inference and Sampling on Bayes' networks
  2. Now: some distributions and intuitions on Bayesian thinking!