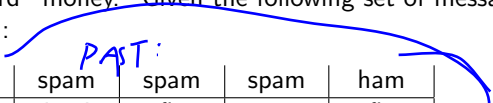


Oct 14 Multivariate Probability

Example: Consider building an e-mail spam filter, where we classify e-mails as SPAM or HAM (valid) based on our prior knowledge of what's contained in SPAM versus HAM exmails. We get a new e-mail that's just the word "money." Given the following set of messages, compute $P(\text{money}|\text{SPAM})$ and $P(\text{money}|\text{HAM})$:



TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

PAST:

NEW:
money

Announcements and To-Dos

Next up: "Practicum"

1) Traveling salesman
via simulated
annealing

Announcements:

1. AB pruning another example (with the α and β tracked):

<https://youtu.be/xBXHtz4Gbdo>

Last time we learned:

1. Finished up with EVs.

2) Short 3-5 page
paper on
AI Ethics

Probability Roundup

1. Probabilities must sum to 1 when considering all possible outcomes.
2. $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$.
3. $P(A \text{ given } B) = P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$, and represents our thoughts about A after gaining the knowledge that event B definitely happened.
4. The *multiplication rule*: $P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$
5. If events A and B are *independent*:
 $P(A \text{ and } B) = P(A)P(B)$; $P(A|B) = P(A)$; $P(B|A) = P(B)$.
factor; *lack of effect*
6. The Law of Total Probability

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \cdots + P(A|E_k)P(E_k)$$

7. Bayes Theorem: $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$

Naive Bayes

Example: Consider building an e-mail spam filter. We receive an e-mail with the words buy, pills and deal. Is this a SPAM e-mail, or valid (HAM)?

Definition: *Class conditional independence* is the assumption that the features of \mathbf{x} are conditionally independent given y . This means that $P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$. We often make this naive assumption.

Handwritten notes:
 → words: buy pill deal
 → HAM SPAM
 Ca factor

Under class conditional independence, we would be assuming that

$$P(x = [\text{buy}, \text{pills}, \text{deal}] | y = \text{SPAM}) = P(\text{buy} | \text{SPAM}) P(\text{pills} | \text{SPAM}) P(\text{deal} | y = \text{SPAM})$$

The Naive Bayes Classifier is just Bayes' theorem:

$$P(\text{spam} | \text{words}) = P(\text{words} | \text{spam}) \cdot P(\text{spam}) / P(\text{words})$$

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

where we have y as the event of "it's spam" and x as the words in the e-mail.

Naive Bayes

We use some special vocabulary when we use Bayes theorem to describe the four pieces.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Definition: $P(x)$ and $P(y)$ are marginal distributions of x and y . Because we're trying to compute a final result on $P(y)$, the right-hand side $P(y)$ is called our prior distribution: it's what we know about y prior to observing any data x .

Definition: $P(x|y)$ is called the *likelihood*. Given a class y , how would we observe x ? It typically comes from an assumed pdf of x .

For our classifier, it's the likelihood of a specific e-mail or set of words *given* that the e-mail is SPAM (or HAM). For a classification problem, it's thusly sometimes called the class-conditional probability.

Definition: $P(y|x)$ and $P(x)$ is the *posterior distribution*. It holds the (classification) probability that data y belongs to class c , given observation of its features x .

SPAM and HAM

Example: Suppose we get a new e-mail that's just the word "money." Given the following set of messages, compute $P(\text{money}|\text{SPAM})$ and $P(\text{money}|\text{HAM})$:

TYPE:	ham	spam	spam	spam	ham	$\leftarrow P(\text{SPAM}), P(\text{HAM})$
	work	nigeria	fly	money	fly	
	buy	money	buy	buy	home	
	money	viagra	nigeria	fly	nigeria	

Properties of HAM: $P(x | \text{HAM})$

$P(x | \text{SPAM})$

SPAM and HAM

Example: Suppose we get a new e-mail that's just the word "money." Given the following set of messages, compute $P(\text{money}|\text{SPAM})$ and $P(\text{money}|\text{HAM})$:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Solution: For this small set of e-mails, these are just discrete values from counting the numbers of occurrences of each event. $P(\text{money}|\text{SPAM}) = \frac{|(\text{both})|}{|(\text{SPAM})|} = \frac{2}{3}$

$$P(\text{money}|\text{HAM}) = \frac{|(\text{both})|}{|(\text{HAM})|} = \frac{1}{2}$$

SPAM and HAM

Example, Cont'd:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

We want $P(SPAM|money)$; $P(HAM|money)$. We can get there with Bayes' Theorem:

$$P(SPAM|w) = \frac{\overbrace{P(w|SPAM)P(SPAM)}^{\text{likelihood}}}{P(w)} \quad \text{prior knowledge of SPAM/HAM}$$

and we can fully write out the denominator using LTP:

$$P(SPAM|w) = \frac{P(w|SPAM)P(SPAM)}{\underbrace{P(w|SPAM)P(SPAM)}_{\text{AND SPAM}} + \underbrace{P(w|HAM)P(HAM)}_{\text{word AND HAM}}}$$

SPAM and HAM

Example, Cont'd: Compute $P(SPAM|money)$ and $P(HAM|money)$:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

SPAM and HAM

Example, Cont'd: Compute $P(SPAM|money)$ and $P(HAM|money)$:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Solution:

$$P(SPAM|money) = \frac{P(money|SPAM)P(SPAM)}{P(money|SPAM)P(SPAM) + P(money|HAM)P(HAM)}$$

$$P(SPAM|money) = \frac{\frac{2}{3} \frac{3}{5}}{\frac{2}{3} \frac{3}{5} + \frac{1}{2} \frac{2}{5}} = \frac{2}{3}$$

$\frac{2}{3} \text{ r } 5$
 $3 / 5$

Sanity check: 2 of the 3 times "money" appears is in a SPAM message!

Classifiers

So imagine we get an e-mail with "money" in it. According to our system, how do we *classify* it?

$\approx 2/3$

$\approx 1/3$

Since $P(SPAM|money) > P(HAM|money)$ (or more than half the time it's SPAM), it's probably reasonable to classify the e-mail as SPAM.

But we should always be worried about making mistakes, and estimating the probability that our classifier gets things wrong! Marking a SPAM message as HAM is annoying but not a big deal. Missing a crucial HAM e-mail because it got filtered out might be disastrous... is $P(SPAM|money) > .5$ the right threshold for this type of problem?

Bigger Better BSFs

Our initial math can filter messages based on the appearances of certain words, but we might want to do this with all of the word simultaneously. We can do that! (with some extra assumptions...)

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Suppose we receive an email with just the words $\{\text{buy}, \text{nigeria}\}$. Now we want to compute things like $P(\text{SPAM}|\{\text{buy}, \text{nigeria}\})$ and $P(\text{HAM}|\{\text{buy}, \text{nigeria}\})$. In turn, we'll have to compute the pieces like $P(\{\text{buy}, \text{nigeria}\}|\text{HAM})$.

Data Science!

For pieces like $P(\{buy, nigeria\} | HAM)$ now we have a *set* of words: we could tally up e-mails that contain *both* words, but this might be very tedious to compute if we receive new emails.

Scary Example: How many distinct *sets* of words are in the sentence "This is not a spam e-mail about money from nigeria" ?

Maybe we don't want to count all possible subsets of words from e-mails, so we might do all the math where we assume that word are *independent*. Is this a good assumption?

Data Science!

For pieces like $P(\{buy, nigeria\}|HAM)$ now we have a *set* of words: we could tally up e-mails that contain *both* words, but this might be very tedious to compute if we receive new emails.

Scary Example: How many distinct *sets* of words are in the sentence "This is not a spam e-mail about money from nigeria"?

Solution: This is the *Power set*! It's asking about all of the ways we can chop up the e-mail into subsets. For even that 10 word sentence that $2^{10} = 1024$ sets. Yikes!

Maybe we don't want to count all possible subsets of words from e-mails, so we might do all the math where we assume that word are independent. Is this a good assumption?

Data Science!

To compute $P(\{buy, nigeria\}|HAM)$ where we assume the words are *independent*, we have an **AND** statement before the conditional... we can just use multiplication!

$$P(\{buy, nigeria\}|HAM) = P(buy|HAM)P(nigeria|HAM)$$

Then we can reuse all of our work for individual words, and multiply the values for all of our individual words together to get the resulting probabilities for an e-mail. Note that the denominators are the same:

$$P(HAM|email) = \frac{P(email|HAM)P(HAM)}{P(email)};$$

$$P(SPAM|email) = \frac{P(email|SPAM)P(SPAM)}{P(email)}$$

The process of making simplifications like independence to save from a computationally disastrous "exact" answer like all possible subsets is a major concern in Data Science.

Bigger Better BSFs

A final BSF: predict if {buy, nigeria} is SPAM.

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Bigger Better BSFs

A final BSF: predict if {buy, nigeria} is SPAM.

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Solution: $P(SPAM|\{buy, nigeria\}) = \frac{P(buy|SPAM)P(nigeria|SPAM)P(SPAM)}{P(\{buy, nigeria\})}$. From before, we have the *numerator* values as $\frac{2}{3} \frac{2}{3} \frac{3}{5} = 0.267$.

$P(HAM|\{buy, nigeria\}) = \frac{P(buy|HAM)P(nigeria|HAM)P(HAM)}{P(\{buy, nigeria\})}$. From before, we have the *numerator* values as $\frac{1}{2} \frac{1}{2} \frac{2}{5} = 0.1$.

Great idea: let's just compare those and ignore the denominators, since they're the same anyways! Since $0.267 > 0.1$, we predict the e-mail is SPAM.

Multivariate models

Today we move onto probability models that tend to assign and track probabilities of *lots* of variables at once. We could track, e.g.:

1. Precipitation, temperature, and barometric pressure
2. Whether we have an illness, whether we feel well, whether the doctor can diagnose us, and whether we perform worse in school
3. Whether our car starts, and whether each and every component of the car is in working order

Probability Models

Probabilistic reasoning gives us a framework for managing uncertain beliefs and knowledge.

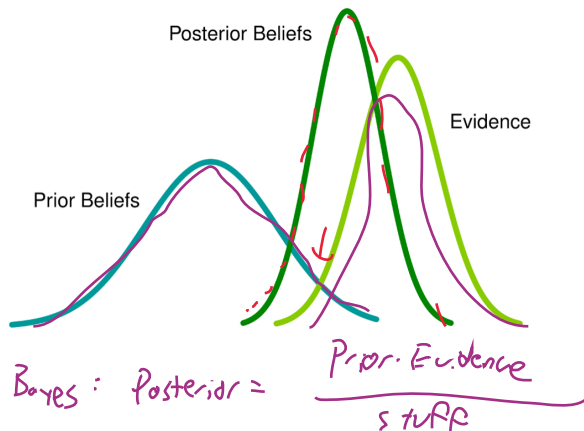
In general, we decompose our variables into classes, and then figure out a way to describe their relationships.

1. Observed variables (evidence): Agent knows certain things about the state of the world (e.g. sensor readings or symptoms)
2. Unobserved variables: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
3. Model: Agent knows something about how the known variables relate to the unknown variables

Bayesian Models

The whole Bayesian statistical framework can be summarized as follows:

1. What is your process of interest?
2. Get data (evidence) for your process of interest
3. Build a model $\nu(\theta)$ of this process
The model depends on uncertain parameters $\theta \rightarrow \theta$ holds: probs, means, variances
4. Formalize your a priori knowledge or assumptions of the model parameters θ in prior distributions, $P(\theta)$
5. Update your prior knowledge using the match between your model and the data



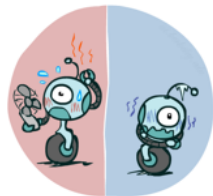
Probability Models

The first tool is the single-variable probability distribution, which assigns a probability to each outcome.

► We require that over all outcomes x : $\sum P(x) = 1$

► $P(x) \geq 0$ for any one outcome.

• Temperature:



$P(T)$

T	P
hot	0.5
cold	0.5

↓
sum to 1!

▪ Weather:



$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

↓
sum to 1!

Joint Probabilities

$\{ \#1, \#2, \dots, \#n \}$
domains

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 $P(x_1, x_2, \dots, x_n)$.

► We still require that over all outcomes,

$$\sum_{\{x_1, x_2, \dots, x_n\}} P(x_1, x_2, \dots, x_n) = 1$$

► Again, $P(x_1, x_2, \dots, x_n) \geq 0$ for any one outcome.

If we have n variables ^{columns} and each has domain sizes or outcomes d , how large is this?

d^n rows (distinct probab. distributions/ outcomes) ^{levels}

For all but the smallest distributions, it's impractical to write out.

temp weather

Variables		P
T	W	
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

2 rows

return sun
medium rain
hot cone +
cold cone +
medium cone +
Sum to 1!

The Marginal Distribution

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

It's the subtable that eliminates other columns (groups by X) and combines/collapses rows by adding:

Variables		P
T	W	
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\begin{aligned}
 \underline{P(t)} &= \sum_w P(t, w) \\
 p(\text{Hot}) &\Rightarrow \underbrace{(.4)}_{w=\text{sun}} + \underbrace{(.1)}_{w=\text{rain}} \\
 P(w) &= \sum_t P(t, w)
 \end{aligned}$$

T	P
hot	0.5
cold	0.5

W	P
sun	0.6
rain	0.4

Example: Joint/Marginal

Consider the joint distribution for X, Y , at left:

X	Y	P
+X	+y	0.2
+X	-y	0.3
^{not} -X	+y	0.4
-X	-y	0.1

$x \neq y$
 $x \neq y$

What are:

1. $P(+x, \neg y)?$

$\neg y \text{ is not } y! = .3$

2. $P(+x)?$

$P(+x \text{ AND } +y) + P(+x \text{ AND } \neg y)$

3. $P(\neg y \text{ OR } +x)?$ $= .3 + .2 = .5$

$= .6$

Example: Joint/Marginal

Consider the joint distribution for X , Y , at left:

X	Y	P
$+X$	$+y$	0.2
$+X$	$-y$	0.3
$-X$	$+y$	0.4
$-X$	$-y$	0.1

What are:

1. $P(+x, \neg y)$?

Sol: .3

2. $P(+x)$?

Sol: $.3 + .2 = .5$

3. $P(\neg y \text{ OR } +x)$?

Sol: $.4 + .2 = .6$

Example: Joint/Conditional

Consider the joint distribution for X, Y , at left.

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

We again compute by summing over rows, but here we group by Y and then sum the X values... but normalize so it adds up to one!

Variables		P
T	W	
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

\Rightarrow

$$P(W | T = \text{hot}) =$$

$$P(\text{sun} | \text{hot}) = \frac{.4}{.5}$$

$$P(W | T = \text{cold})$$

$$P(\text{both}) / P(T = \text{hot})$$

div by .5

What are:

1. $P(\text{sun} | \text{hot}) = .8$

2. $P(\neg \text{sun} | \neg \text{hot})?$

$$P(\text{both}) \quad .3 / .5$$

$P(\text{cold})$

Example: Joint/Conditional

Consider the joint distribution for X, Y , at left.

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

We again compute by summing over rows, but here we group by Y and then sum the X values... but normalize so it adds up to one!

Variables		P
T	W	
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

\Rightarrow

$P(W T = \text{hot})$	
W	P
sun	0.8
rain	0.2

$P(W T = \text{cold})$	
W	P
sun	0.4
rain	0.6

What are:

1. $P(\text{sun}|\text{hot})?$

Sol: .8

2. $P(\neg \text{sun}|\neg \text{hot})?$

Sol: .6

Both tables *combined* are the full marginal distribution of $P(W|T)$.

Emphasis on Independence

Imagine we know the marginals:

$P(W T = \text{cold})$	
W	P
sun	0.4
rain	0.6

$P(W T = \text{cold})$	
W	P
hot	0.5
cold	0.5

These events could be *independent*:

Variables		P
T	W	
hot	sun	0.2
hot	rain	0.3
cold	sun	0.2
cold	rain	0.3

Handwritten notes: Blue brackets on the left of the first two columns. A red arrow points from the circled 0.2 to the text "Cold half the time".

...or not, as we had before:

Variables		P
T	W	
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Handwritten notes: Blue brackets on the left of the first two columns. A red bracket on the right of the last two rows.

Emphasis on Independence

To track multiple variables we need to carefully define what we mean by independence.

1. **Definition:** X and Y are *independent* if for all events x and y , $\underline{P(x, y) = P(x)P(y)}$. 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 $(X \perp\!\!\!\perp Y)|Z$.

Words in spam filter

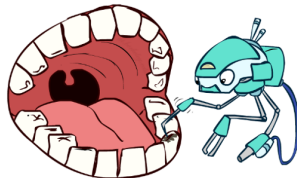
We interpret this as: X and Y may be related, but they're only related *because of Z* .
Once we know Z , then X and Y become unrelated outcomes.

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:

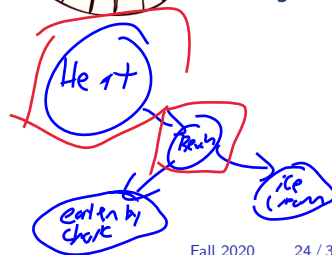
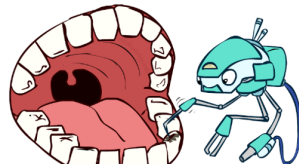
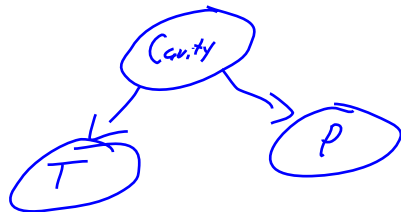
$$P(+p | +c, +t) = P(+p | +c)$$

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

$$P(+p | \neg c, +t) = P(+p | \neg c)$$

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:

- ▶ A directed acyclic graph (DAG).
- ▶ 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**

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

Bayesian Networks

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

1. “Trivial” decomposition:
2. Conditional Independence:
3. Visually:

Bayesian Networks

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

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 $P(\text{traffic}, \text{umbrella}, \text{rain})$.

1. “Trivial” decomposition:

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

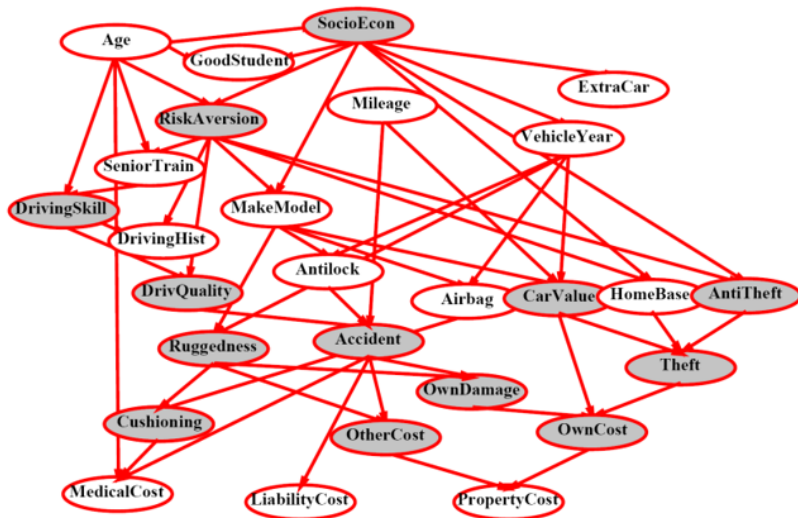
...but this is kind of useless: both pedestrians using umbrellas and drivers driving slowly are responding to the weather, so $P(U|TR)$ feels awkward.

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

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

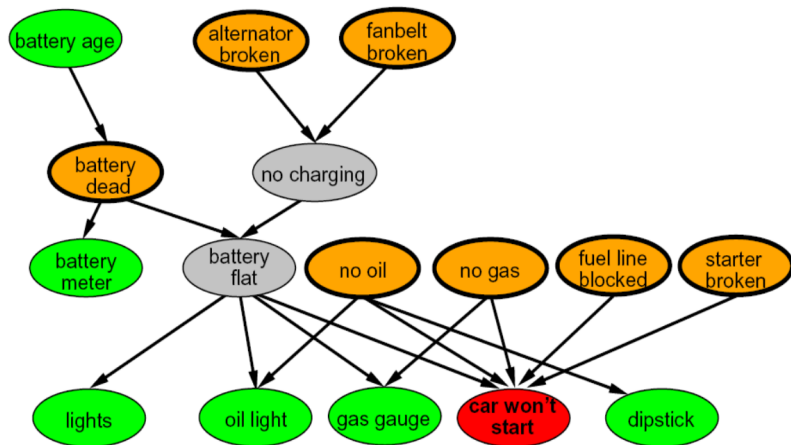
3. Visually:

Bayesian Networks Examples: Insurance

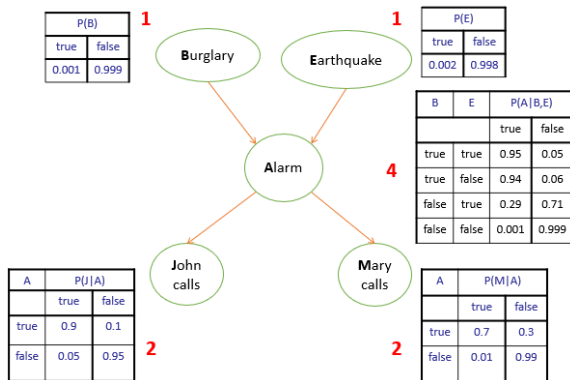


Bayesian Networks: Your Car Won't Start

Bayesian Networks: Your Car Won't Start



Example: Two steps of alarm:



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

Moving Forward

- ▶ This Week:
 - 1. nb day Friday
- ▶ Next time: Math on Bayes' networks