4/23/19 Lecture 7

Law of total probability:

$$P(A) = \sum_{j=1}^{k} P(B_j) P(A \mid B_j)$$

These probabilities add up to 1

These don't

Expressing the probability of A as a weighted average of the conditional probabilities of A given the partition sets weighted by the probabilities of those partition sets.

Extension of the LTP: Assuming all conditional probabilities are defined in what follows, if C is in $\mathbb{C}$ then $P(A \mid C) = \sum_{j=1}^{k} P(B_j \mid C) P(A \mid B_j \cap C)$.

Def: Events A, B are independent iff $P(A \cap B) = P(A) \cdot P(B)$, which (as long as $P(A) > 0$, $P(B) > 0$) is equivalent to the Bayesian view that $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$

Consequences of the def. of independence:

1) If A and B are independent, then so are A and B^c, A^c and B, and A^c and B^c

2) Extension of the def. to more than 2 events:

Def: Given events $A_1, \ldots, A_k$, they are mutually independent if for every subset $A_{i_1}, \ldots A_{i_j}$ of $(A_1, \ldots, A_k)$ $(j = 2, \ldots, k)$,

$$P(A_{i_1} \cap \ldots \cap A_{i_j}) = P(A_{i_1}) \ldots P(A_{i_j})$$

Bayesian Interpretation of Independence: A, B independent iff information about A doesn't change the chances associated with B, and vice versa

Def: Another useful extension of independence:
Events $\{A_1, \ldots, A_k\}$ are conditionally independent given event B if for every subset $\{A_{i_1}, \ldots, A_{i_j}\}$ of $\{A_1, \ldots A_k\}$ $(j = 2, \ldots, k)$,

$$P(A_{i_1} \cap \ldots \cap A_{i_j} \mid B) = \prod_{\ell=1}^{j} P(A_{i_\ell} \mid B)$$

(product)

Statistical Ex:
Suppose that there is a machine that can take an ordinary coin and produce IID tosses of the coin with $P(H) = \theta$ with H being any one toss, and θ can be set to any value in $[0,1]$ with a dial on the machine's control panel

Someone sets the dial to a θ value that's unknown to you and starts producing coin tosses $Y_1, Y_2, \ldots$

Suppose the first 10 tosses come out 1011100111
H T H H H T T H H H
(7 H, 3 T) "bits" (binary digits)

Q: Is there information in these first 10 tosses that helps you predict the next outcome Y_{11}?

A: Yes, definitely. It looks like θ is around $\frac{7}{10}$ so you would predict $Y_{11} = H$. Thus Y_{11} depends on Y_1 through Y_{10} probabilistically.

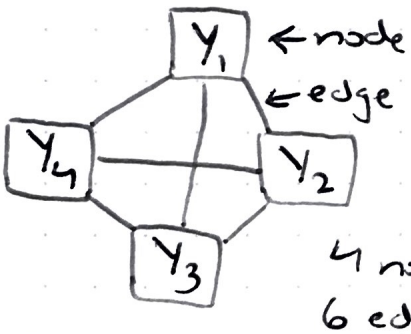
Now suppose instead that you watched the person with the machine set the dial to $\theta = 0.81$ so that θ is now known to you. The next 10 tosses come out HHHHTHHTHHHH (8H, 2T)

Q: Is there information in these 10 tosses that helps you predict the next toss?

A: No. You know that $\theta = 0.81$ so there's no information in any of the Y_i that helps you to predict any of the other Y_j. Given θ, the Y_i are independent.

Thus, the Y_i are <u>unconditionally dependent</u> but <u>conditionally dependent given θ</u>.

Ex with 4 Y values: Graph Theory



← node
← edge

4 nodes
6 edges

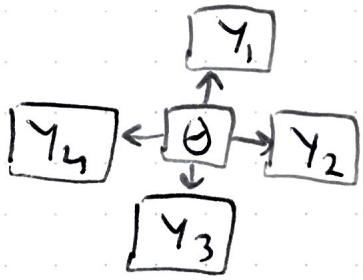
Edges denotes any dependence amongst the nodes

Y_i dependent

n nodes ← # of data points

$\binom{n}{2}$ edges $= \frac{n(n-1)}{2} = O(n^2)$

This is of order n^2
"Big O" of n^2



5 nodes
6 edges

$(n+1)$ nodes
n edges $= O(n)$

$X - Y$	$X \to Y$
X & Y are dependent	X causes Y

As soon as we know what θ is, the Y_i become conditionally dependent given θ.

The past and future are conditionally independent given the truth (θ)

Bayes' Theorem for events based on the possibility of a finite partition:

Suppose that the events $B_1, \ldots, B_k$ partition the sample space in such a way that $P(B_j) > 0$ for all $j = 1, \ldots, k$. If A is an event with $P(A) > 0$, then for each $i = 1, \ldots, k$ $P(B_i \mid A) = \frac{P(B_i) P(A \mid B_i)}{P(A)}$

and by the LTP, this is:

$$P(B_i \mid A) = \frac{P(B_j) \cdot P(A \mid B_i)}{\sum_{j=1}^{k} P(B_j) P(A \mid B_j)}$$

How this theorem is used in Bayesian statistics:

The B_i represent unknown states of the world: They're all possible — $P(B_i) > 0$ — and only one of them is true, but you don't know which one.

A represents data: information that will help you identify the most probable B_i

(A priori) — Before the dataset A arrives, you have background information about the plausibility of the B_i that you can represent with prior probabilities $P(B_i)$

(A posteriori) — After the dataset A arrives, you can use Bayes Theorem to update your prior probabilities to posterior probabilities $P(B_i \mid A)$

The probabilities $P(A \mid B_i)$ represent how likely the dataset A would be if B_i were the actual unknown state aka likelihood information

The denominator $P(A)$ does not depend on the B_i, and can therefore be regarded as a normalizing constant put into Bayes Theorem to make all the $P(B_i \mid A)$ add up to 1

Thus $\frac{P(B_i)\,P(A \mid B_i)}{P(A)}$ is interpreted as

$$\left(\begin{matrix}\text{Posterior}\\ \text{information}\end{matrix}\right) = \frac{\left(\begin{matrix}\text{prior}\\ \text{information}\end{matrix}\right) \cdot \left(\begin{matrix}\text{likelihood (data)}\\ \text{information}\end{matrix}\right)}{\left(\begin{matrix}\text{normalizing}\\ \text{constant}\end{matrix}\right)}$$

Back to credit card case study:

We know $P(B) = 0.01$ $P(S- \mid G) = 0.97$ $P(S+ \mid B) = 0.98$

We want $P(B \mid S+)$ (truth (unknown) ↑, system says (data) ↑)

Only 2 possible states: (B, G)

Method 1: 2×2 table

Method 2: Bayes' Theorem in odds form

$$\frac{P(B \mid S+)}{P(\text{not } B \mid S+)} = \left[\frac{P(B)}{P(\text{not } B)}\right] \cdot \left[\frac{P(S+ \mid B)}{P(S+ \mid \text{not } B)}\right]$$

We know all the probabilities on the right side

$\text{not } B = G$

$$\left(\begin{matrix}\text{posterior odds}\\ \text{in favor of } B\end{matrix}\right) = \left(\begin{matrix}\text{prior odds}\\ \text{in favor of } B\end{matrix}\right) \cdot \left(\begin{matrix}\text{Bayes factor}\\ \text{in favor of } B\end{matrix}\right)$$

$P(B) = 0.01$
$P(\text{not } B) = P(G) = 0.99$
$P(S+ \mid B) = 0.98$
$P(S+ \mid \text{Not } B) = 1 - P(S- \mid \text{Not } B) = 1 - P(S- \mid G) = 1 - 0.97 = 0.03$

So $\frac{P(B \mid S+)}{P(\text{not } B \mid S+)} = \left(\frac{0.01}{0.99}\right)\left(\frac{0.98}{0.03}\right)$

$= \begin{pmatrix} 99 \text{ to } 1 \text{ prior} \\ \text{odds against } B \end{pmatrix} \begin{pmatrix} 98 \text{ to } 3 \text{ odds} \\ \text{in favor of } B \end{pmatrix}$

$= \frac{98}{(99)(3)} = \frac{98}{297}$ ← posterior odds in favor of B

o = odds
p = probability

$o = \frac{p}{1-p} \rightarrow p = \frac{o}{1+o}$

$p = \frac{\frac{98}{297}}{1+\frac{98}{297}} = \frac{98}{98+297} = \frac{98}{395} = 0.25$ as before

Method 3: Bayes' Theorem in probability form

$P(B|S+) = \frac{P(B)\,P(S+|B)}{P(S+)} = \frac{(0.01)(0.98)}{?}$

↑ prediction

Truth: {B, G} (unknown)

Data: {S+, S−}

$P(B|S+)$ Search for truth on basis of data: statistical inference

↑ truth (unknown) ↑ data (observable)

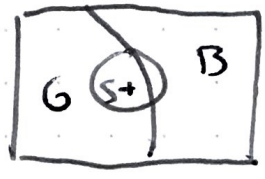
Unknown (true) state of world may well be unobservable

We don't know S+ by itself so we need B or G

When Y is hard, get help by extending the conversation. Find some other aspect of world X upon which Y depends and predict Y in 2 stages: $\begin{Bmatrix} X \\ (Y|X) \end{Bmatrix}$

P(S+) = data (Y)

Other info (X) = $\begin{Bmatrix} G \\ B \end{Bmatrix}$



Partition the world

$$P(S+) = P[(S+ \text{ and } G) \text{ or } P(S+ \text{ and } B)]$$

$$= P(S+ \text{ and } G) + P(S+ \text{ and } B)$$ There's no overlap

$$= P(G)P(S+|G) + P(B)P(S+|B)$$

$$= (0.99)(1 - P(S-|G)) + (0.01)(0.98)$$

$$= (0.99)(1-0.97) + (0.01)(0.98)$$

$$= (0.99)(0.03) + (0.01)(0.98)$$

$$= 0.0297 + 0.0098 = 0.0395$$

$$P(B|S+) = \frac{P(B)P(S+|B)}{P(S+)} = \frac{(0.01)(0.98)}{0.0395} = \frac{98}{395} = 0.25$$

Ch. 3 Random variables and their distributions

Ex: Tay-Sachs Disease

T = T-S baby

N = not T-S baby

S →		Y ← # of T-S babies
NNNNN	0	
TNNNN		
NTNNN		
NNTNN		
NNNTN		
NNNNT	1	
TTNNN		
TNTNN		
TNNTN		
TNNNT		
NTTNN		
NTNTN		
NTNNT		
NNTTN		
NNTNT		
NNNTT	2	
⋮	⋮	
TTTTT	5	

Def: Given a sample space S for an experiment $\mathcal{E}$, a (real-valued) random variable (rv) is a function from the non-empty collection $\mathbb{C}$ of subsets of S to the real number line $\mathbb{R}$

In the T-S case study, the elements s of S look like NNNTN and the rv Y counts how many Ts they contain.

For instance, $Y(TNNTN) = 2$ and $Y(NNNTT) = 2$ also. Y ignores the order of the children.

To simplify: $P(Y=y) = P(\{s : Y(s) = y\})$

$P(Y=y)$: T/F proposition; $\{s : Y(s) = y\}$: set

For ex) $P(Y=1) = P(\{s \in S : Y(s) = 1\})$

$= P(\{TNNNN, NTNNN, NNTNN, NNNTN, NNNNT\})$

In general the values a random variable takes on could be just about anything, but in this course all of our rvs will be real-valued.

In the T-S case study the rv Y can only take on the values $0, 1, \ldots, 5$

$P(Y=y)$

↑ random variable (process) ↑ a possible value of Y (outcome)

y	$P(Y=y)$
0	0.237
1	0.396
2	0.264
3	0.088
4	0.015
5	0.001

You can see that a rv Y is completely specified by 2 things: the values it can take on and the probabilities for those values.

Def: The probability distribution of a random variable Y is the collection of all probabilities of the form $P(Y \in A)$ for all sets A of real #s in the non-weird collection $\mathcal{C}_{\mathbb{R}}$ of subsets of the real number line $\mathbb{R}$

The rv Y in the T-S case study has a finite set of possible values — this time of same, but not all, rvs

Def: A random variable Y has a discrete distribution, or equivalently Y is a discrete rv if the set of possible distinct values of Y is finite or at most countably infinite;

rvs for which the set of all possible values is uncountable are called continuous random variables

Ex: The rv $X = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases}$ (with Y = # T-S babies)

is discrete, taking on only the values {yes, no} → {0, 1}
such rvs are called dichotomous or binary

Ex: Imagine a scale for weighing things that has a dial you can set to specify how many significant figures (sig figs) of precision you want.

Buy a "1 pound" package of butter and weigh it.

Possible Weights
16
16.0
15.99
15.9930
15.9928
⋮

If there's no conceptual limit to the number of sigfigs you could get, a rv Y = (the actual true weight of the package) should be modeled as continuous, having positive values on the real number line $\mathbb{R}$ $(0, \infty)$

Reality Check: Infinite precision is impossible in practice; every measurement you ever make is in actuality discrete, but it's useful to regard rvs that are conceptually continuous (i.e, no limit in principle to the precision of measurement) as continuous