PROBABILITY: QUICK REVIEW

Two conceptions:

- certainty/likelihood of particular events
  - eg., Pr(particular coin flip = Heads);
  - e.g., Pr(ith word in a certain text is a noun)
- relative frequency of a certain kind of outcome in an infinitely large number of repetitions of some "repeatable experiment"
  - e.g., proportion of "heads" in ∞ no. of coin flips
  - e.g., proportion of nouns in an ∞ text corpus

- In estimating probabilities, we use the latter conception.
- In applying (estimated) probabilities, we use the first conception.

RANDOM VARIABLES

- A r.v. picks out some observable in a repeatable experiment
- Its value varies from repetition to repetition
  - e.g., "Flip" could be = H (for heads), = T (for tails) in repeated coin tosses
  - e.g., "CAT" could be N, V, P, A, Det, ... (i.e., POS) for a word found in some arbitrary location in a corpus

Pr(FLIP = H) = .497
says that the prob. of heads in flipping some coin is 49.7%

Pr(CAT = N) = .24
says that the prob. that an arbitrarily selected word in some large corpus (or in some type of corpus) is a noun is 24%.

Abbreviation:

Pr(H) = .497, Pr(N) = .24

r.v. is "understood"

Jointly distributed r.v.'s

- We can consider multiple r.v.'s at the same time, coming from one experiment or multiple experiments
- The values they assume may be <u>dependent</u> on one another (because there is some causal connection between the r.v.'s), or <u>independent</u>. (We'll formalize this.)

E.g., Let (X_1, X_2) be r.v.'s for the outcomes of 2 successive rolls of a die.

So, values are $(1,1), (1,2), \ldots, (6,6)$

For a fair die & causally independent, randomized rolls, $Pr(X_1=1, X_2=1) = Pr(X_1=1, X_2=2) = \ldots = Pr(X_1=6, X_2=6) = 1/36$ "joint probability"

E.g., Let (C_1, C_2, C_3) be r.v.'s for the POS of 3 successive words randomly selected from a large corpus.

Possible values: $(N,N,N), (N,N,V), \ldots, (Adv, Adv, Adj)$

<u>Interdependent</u>: $Pr(N,N,N) \neq (.24)^3$

$Pr(C_1=N, C_2=N, C_3=N)$

Probability Axioms

1. The probability of any outcome lies between 0 and 1

$$0 \leq Pr(X=x) \leq 1$$

where x is any possible value of r.v. X

2. The probabilities of all possible alternative outcomes add up to 1. E.g., $Pr(FLIP=H) + Pr(FLIP=T) = 1$

$$\sum_{x \in range(X)} Pr(X=x) = 1$$

Informally, 1. $0 \leq Pr(x) \leq 1$

2. $\sum_x Pr(x) = 1$

Similarly for jointly distributed r.v.'s:

1. $0 \leq Pr(\underline{x}) \leq 1$
2. $\sum_{\underline{x}} Pr(\underline{x}) = 1$

where $Pr(\underline{x})$ abbreviates $Pr(\underline{X} = \underline{x})$ and

$\underline{X} = (X_1, \ldots, X_n)$, i.e., n r.v.'s, and

$\underline{x} = (x_1, \ldots, x_n)$, i.e., an n-tuple of possible values of $X_1, \ldots, X_n$.

Conditional Probability

For jointly distributed r.v.'s, we may be interested in how probable a certain value is for one of them, given the value of another. Write as

$Pr(X=x \mid Y=y)$, or briefly, $Pr(x \mid y)$

(given)

E.g., Prob. that i^{th} word of a text is a noun, given that the preceding word is a determiner might be written $Pr(C_i = N \mid C_{i-1} = Det)$.

Def^n $$Pr(X=x \mid Y=y) = \frac{Pr(X=x, Y=y)}{Pr(Y=y)}$$ ("and") provided that $Pr(Y=y) > 0$

E.g., suppose $Pr(C_{i-1} = Det, C_i = N) = 0.2$

i.e., when we randomly select a pair of successive words from a corpus, 20% of the time they are a Det, N pair

Any 2 successive words

Venn diagram

& suppose $Pr(C_{i-1} = Det) = .25$

i.e., when we randomly select a word from a corpus, 25% of the time it is a Det

Then $$Pr(C_i = N \mid C_{i-1} = Det) = \frac{Pr(C_i = N, C_{i-1} = Det)}{Pr(C_{i-1} = Det)} = \frac{0.2}{0.25}$$

$$= .80$$

Chain rule

$Pr(a \& b) = Pr(a) Pr(b \mid a)$ from def^n of cond'l prob

Similarly

$Pr(a \& b \& c) = Pr(a) Pr(b \mid a) Pr(c \mid a \& b)$

Useful esp. if we can make certain independence assumptions. Here, suppose

a, c are conditionally independent, given b:

$Pr(c \mid a \& b) = Pr(c \mid b)$

Then

$Pr(a \& b \& c) = Pr(a) Pr(b \mid a) Pr(c \mid b)$

We can easily generalize to n events or r.v.'s.

E.g., Suppose $C_1, \ldots, C_n$ represent the parts of speech of a stretch of n successive words randomly selected from a text corpus.

$$Pr(c_1, \ldots, c_n) = Pr(c_1) Pr(c_2 \mid c_1) Pr(c_3 \mid c_1, c_2) \ldots$$
$$\ldots Pr(c_n \mid c_1, \ldots, c_{n-1})$$

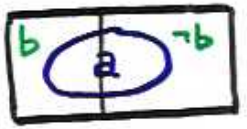
A "Markov assumption"

Suppose C_i is conditionally independent of all earlier parts of speech, given the immediately preceding one:

$$Pr(c_i \mid c_1, \ldots, c_{i-1}) = Pr(c_i \mid c_{i-1})$$

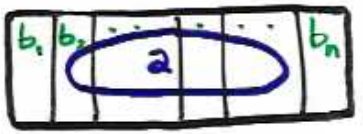
Then $$Pr(c_1, \ldots, c_n) = Pr(c_1) \prod_{i=2}^{n} Pr(c_i \mid c_{i-1})$$

"Marginal" (absolute) probabilities in terms of joint probabilities



$$Pr(a) = Pr(a \& b) + Pr(a \& \neg b)$$

More generally, suppose $b_1, \ldots, b_n$ are mutually exclusive (disjoint), jointly exhaustive events (total prob. 1). Then



$$Pr(a) = \sum_{i=1}^{n} Pr(a \& b_i) = \sum_{i=1}^{n} Pr(b_i) Pr(a|b_i)$$

"Marginalization"

Independence

Intuitively, two events A, B are independent if the fact that one occurred (or failed to occur) tells us nothing about the probability that the other occurred; i.e.,

$P(A|B) = P(A)$ "B provides no evidence about A"
$P(A|\neg B) = P(A)$
$P(B|A) = P(B)$
$P(B|\neg A) = P(B)$
$P(A \wedge B) = P(A)P(B)$
$P(A \wedge \neg B) = P(A)P(\neg B)$
etc.

Any of the variant formulations can be proved from any of the others

E.g., Flip a coin twice
Possible outcomes $\{\langle H,H\rangle, \langle H,T\rangle, \langle T,H\rangle, \langle T,T\rangle\}$
A (heads on 1st toss), B (heads on 2nd toss)

If all 4 prob's are 1/4, then A, B are independent.

Note: independent ⇏ NOT disjoint

Independent r.v.'s

E.g., For the above prob. space, let
$X(\langle H,H\rangle) = X(\langle H,T\rangle) = 1$, $X(\langle T,H\rangle) = X(\langle T,T\rangle) = 0$
$Y(\langle H,H\rangle) = Y(\langle T,H\rangle) = 1$, $Y(\langle H,T\rangle) = Y(\langle T,T\rangle) = 0$
i.e., "X = 1" if 1st toss is H, & 0 o.w.
"Y = 1" if 2nd — " —

Independent r.v.'s, cont'd

So again, if $P(X=1) = P(X=0) = P(Y=1) = P(Y=0)$,
X, Y are independent r.v.'s; & we have

$$P(X=x \wedge Y=y) = P(X=x)\,P(Y=y) \quad \text{for } x,y \in \{0,1\}$$

also $P(X=x \mid Y=y) = P(X=x)$ for $x,y \in \{0,1\}$
etc.

Often abbreviations are used, such as

$P(x)$ for $P(X=x)$, & similarly for Y.

So, for independent X, Y,

$P(x,y) = P(x)P(y)$, etc.

Another example

Space = CS grad students
X = verbal GRE score (up to 800)
Y = analytical GRE score (up to 800)
Z = 444 grade (up to 4.0)
These are probably nearly independent

Conditional independence

e.g., $P(Z=z \mid X=x \wedge Y=y) = P(Z=z \mid X=x)$

provides no further evidence about Z, beyond what X=x provides

Bayes' Rule

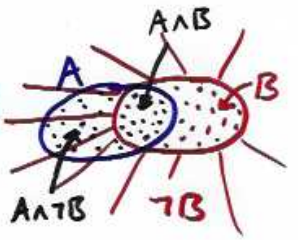
e.g. B = cold
A = sore throat

$$P(A \wedge B) = P(A)\,P(B \mid A) = P(B)\,P(A \mid B)$$

$\therefore P(B \mid A)$ prob. of cold, given sore throat

$$= \frac{P(B)\,P(A \mid B)}{P(A)} \quad (1)$$

$$= \frac{P(B)\,P(A \mid B)}{P(A \wedge B) + P(A \wedge \neg B)} \quad (2)$$



$P(A \wedge B)$ and $P(A \wedge \neg B)$: mutually exclusive (disjoint)

$$P(B \mid A) = \frac{P(B)\,P(A \mid B)}{P(B)\,P(A \mid B) + P(\neg B)\,P(A \mid \neg B)} \quad (3)$$

$P(\neg B) = 1 - P(B)$

E.g., if we know:
$P(B) = P(\text{cold}) = .01$ Prior prob of cold hypothesis
$P(A \mid B) = P(\text{sore throat} \mid \text{cold}) = .9$
Likelihood of evidence, given the cold hypothesis
$P(A \mid \neg B) = P(\text{sore throat} \mid \text{no cold}) = .005$
Likelihood of evidence, given falsity of cold hypothesis

Then we can compute P(hypoth. | new evidence):

$P(B|A) = P(\text{cold} \mid \text{sore throat})$

= posterior probability of cold hypothesis

$$= \frac{(.01)(.9)}{(.01)(.9) + (1-.01)(.005)} \quad \text{(prior = .01)}$$

$$= .645$$

NB: much bigger than prior, .01

Odds-Likelihood form of Bayes' Rule

Above we needed 3 numbers to compute $P(B|A)$, but it really only takes 2....

By analogy with (3):

$$P(\neg B|A) = \frac{P(\neg B)\,P(A|\neg B)}{P(\neg B)P(A|\neg B) + P(B)P(A|B)} \quad (3')$$

Divide (3) by (3') (Note: same denominator!):

$$\frac{P(B|A)}{P(\neg B|A)} = \frac{P(B)\,P(A|B)}{P(\neg B)\,P(A|\neg B)}$$

$$\boxed{O(B|A) = O(B)\,L(A|B)} \quad (4)$$

O "odds"
L "likelihood ratio"

- where the "odds" in favor of an event (or proposition) A is the ratio of its probability of occurrence to its probability of nonoccurrence

$$O(B) = \frac{P(B)}{P(\neg B)} = \frac{P(B)}{1-P(B)}$$

Similarly for conditional (or posterior) odds,

$$O(B|A) = \frac{P(B|A)}{P(\neg B|A)} = \frac{P(B|A)}{1-P(B|A)}$$

- and the likelihood ratio $L(A|B)$ is the ratio of the likelihood of the evidence, given the hypothesis, to the likelihood of the evidence, given the denial of the hypothesis

$$L(A|B) = \frac{P(A|B)}{P(A|\neg B)}$$

How to recover probabilities from (4):

From $O(B) = \frac{P(B)}{1-P(B)}$, it's easily shown that

$$P(B) = \frac{O(B)}{1+O(B)}$$

Similarly

$$P(B|A) = \frac{O(B|A)}{1+O(B|A)}$$

Combining Evidence

E.g. might have 2 pieces of evidence for a "cold" hypothesis: sore throat (E_1), runny nose (E_2); hypothesis H

Write down Bayes rule for $P(H|E_1)$ again:

$$P(H|E_1) = \frac{P(H)\,P(E_1|H)}{P(H)\,P(E_1|H) + P(\neg H)\,P(E_1|\neg H)} \quad (5)$$

Take this posterior prob. as the prior prob. for the next piece of evidence, E_2:

$$P(H|E_1 \wedge E_2) = \frac{P(H|E_1)\,P(E_2|H \wedge E_1)}{P(H|E_1)\,P(E_2|H \wedge E_1) + P(\neg H|E_1)\,P(E_2|\neg H \wedge E_1)}$$

Now assume that relative to a situation where the hypothesis H is given (known to be true), E_1 & E_2 are independent (conditional independence):

$$P(E_2|H \wedge E_1) = P(E_2|H)$$

Similarly if $\neg H$ is given (H is known to be false):

$$P(E_2|\neg H \wedge E_1) = P(E_2|\neg H)$$

Then (5) simplifies to:

$$P(H|E_1 \wedge E_2) = \frac{P(H|E_1)\,P(E_2|H)}{P(H|E_1)\,P(E_2|H) + P(\neg H|E_1)\,P(E_2|\neg H)} \quad (6)$$

So, we already know how to compute $P(H|E_1)$ using $P(H)$, $P(E_1|H)$, & $P(E_1|\neg H)$

So (6) shows us how to allow for both pieces of evidence, E_1 & E_2, with the additional data $P(E_2|H)$, $P(E_2|\neg H)$

E.g. Let's use $P(\text{cold}) = .01$, $P(\text{sore throat}|\text{cold}) = .9$, $P(\text{sore throat}|\neg\text{cold}) = .005$ as before, so we get

$$P(\text{cold}|\text{sore throat}) = .645$$

as before.

Now assume new evidence "sniffles", where

$$P(\text{sniffles}|\text{cold}) = .95$$
$$P(\text{sniffles}|\neg\text{cold}) = .1$$

From (6): $P(\text{cold}|\text{sore throat} \wedge \text{sniffles}) =$

$$\frac{.645(.95)}{.645(.95) + (1-.645)(.1)} = .945$$

note the increase!

We can get an odds-likelihood version as before: Write an expression for

$$P(\neg H | E_1 \wedge E_2)$$

analogous to (6). Call this (6'). Divide (6)/(6'):

$$\frac{P(H|E_1 \wedge E_2)}{P(\neg H|E_1 \wedge E_2)} = \frac{P(H|E_1)\,P(E_2|H)}{P(\neg H|E_1)\,P(E_2|\neg H)}$$

$$O(H|E_1 \wedge E_2) = O(H|E_1)\,L(E_2|H)$$

But $O(H|E_1) = O(H)\,L(E_1|H)$

So

$$\boxed{O(H|E_1 \wedge E_2) = O(H)\,L(E_1|H)\,L(E_2|H)} \quad (7)$$

To "update" the odds in favor of H as new evidence $E_1, E_2, \ldots$ comes in, we just multiply by the corresponding likelihood ratios $L(E_1|H), L(E_2|H), \ldots$

(but keep in mind the conditional independence assumption!)