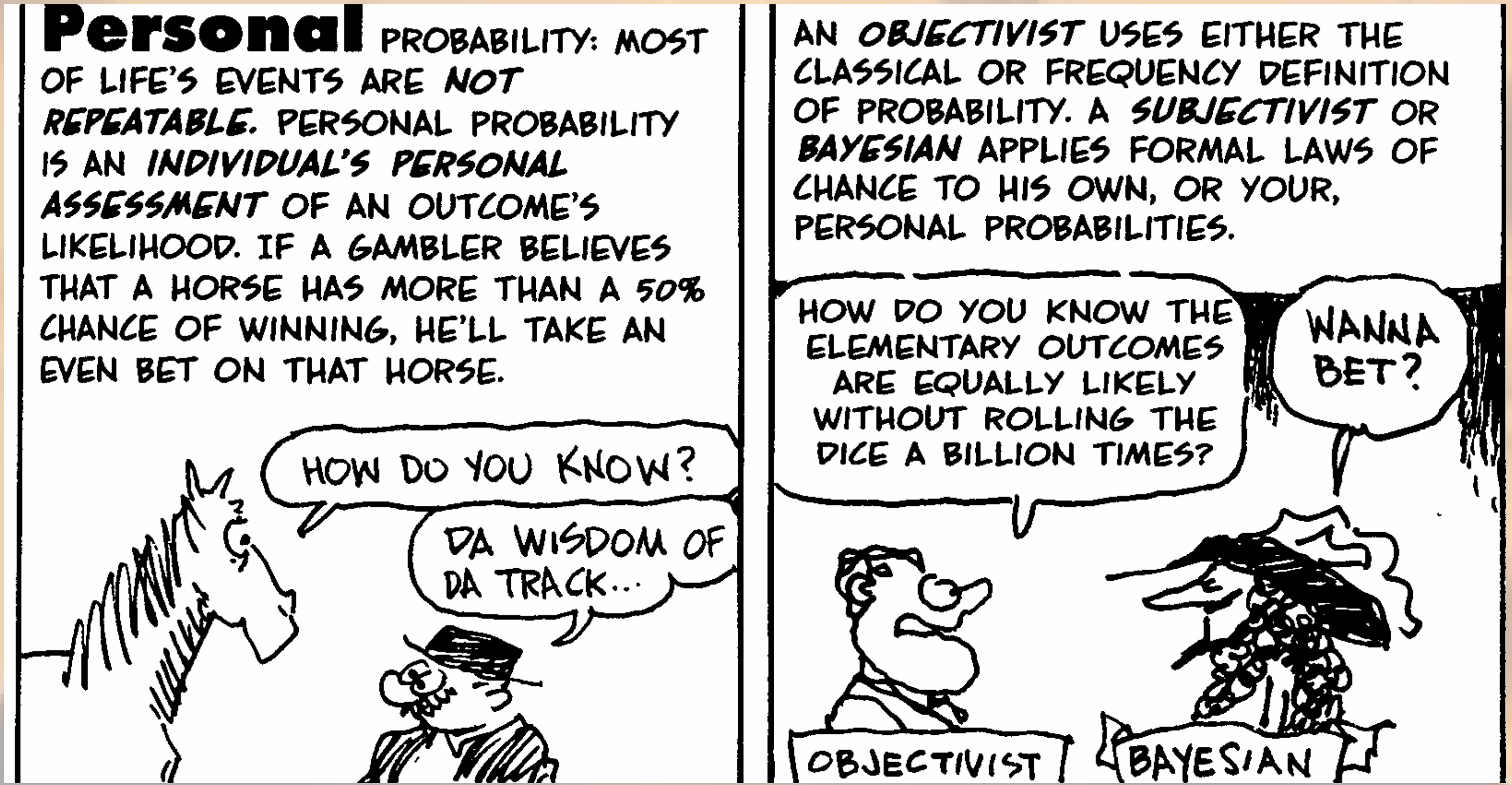
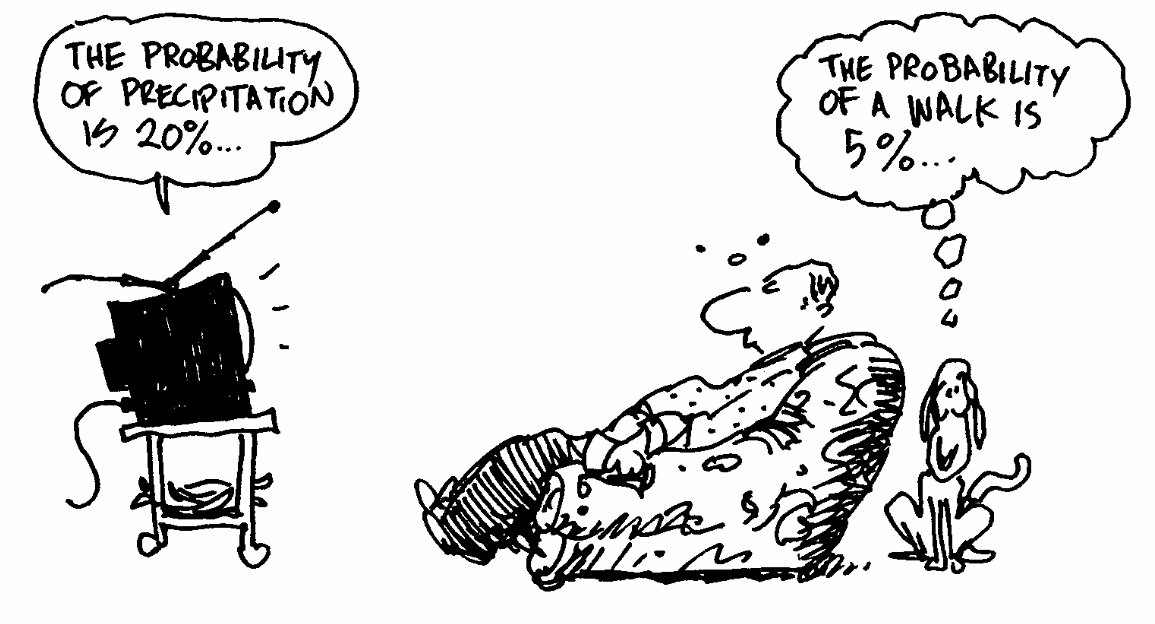
Probability as a measure of belief



Probabilities are to be interpreted



Dictionary definition:
probability = chance = likelihood = probability

?

Numerical measures of belief

- Belief in a proposition, *f*, can be measured in terms of a number between 0 (definitely false) and 1 (definitely true) — this is the probability of *f*
- *f* has a probability between 0 and 1, doesn't mean *f* is true to some degree, but means that you are ignorant of its truth value. Probability is a measure of your ignorance

Random Variables

- A random variable is a term in a language that can take one of a number of different values
- *dom(x)*, the domain of a variable, is the set of values *x* can take
- a tuple of random variables $<x_1,\ldots, x_n>$ is a complex random variable with domain $dom(x_1) \times \ldots \times dom(x_n)$
- a proposition is a Boolean formula made from assignments of values to variables

Possible world semantics

- A possible world specifies an assignment of one value to each random variable

$$w \models x = v$$

means variable *x* is assigned value *v* in world w

- logical connectives have their standard meaning

$$w \models \alpha \wedge \beta \text{ if } w \models \alpha \text{ and } w \models \beta$$

$$w \models \alpha \vee \beta \text{ if } w \models \alpha \text{ or } w \models \beta$$

$$w \models \neg\alpha \text{ if } w \not\models \alpha$$

Semantics of probability

- For a finite number of variables with finite domains:
  - define a nonnegative measure $\mu(w)$ to each world w so that the measures of the possible worlds sum to 1. The measure specifies how much you think the world w is like the real world.
  - The probability of a proposition f is defined by

$$P(f) = \sum_{w \models f} \mu(\omega)$$

Axioms of probability

- Axiom 1. $P(f) = P(g)$ if $f \leftrightarrow g$ is a tautology. That is, logically equivalent formulae have the same probability.
- Axiom 2. $0 \leq P(f)$ for any formula f.
- Axiom 3. $P(\tau) = 1$ if τ is a tautology.
- Axiom 4. $P(f \vee g) = P(f) + P(g)$ if $\neg(f \wedge g)$ is a tautology.

These axioms are sound and complete with respect to the semantics

Conditioning

- Specifies how to revise beliefs based on new information
- Building a probabilistic model starts by taking all background information into account. This gives the prior probability.
- All other information must be conditioned on.
- If evidence *e* is all the info obtained subsequently, the conditional probability $P(h|e)$ of h given *e* is the posterior probability of *h*

Semantics of conditional probability

- Evidence *e* rules out possible worlds incompatible with *e*.
- Evidence e induces a new measure, μ_e, over possible worlds

$$\mu_e = \begin{cases} \frac{1}{P(e)} \times \mu(\omega) \text{ if } \omega \models e \\ 0 \qquad\qquad\quad \text{ if } \omega \not\models e \end{cases}$$

- The conditional probability of formula *h* given evidence *e* is

$$P(h \mid e) = \sum_{\omega \models h} \mu_e(w) = \frac{P(h \wedge e)}{P(e)}$$

Properties of conditional probabilities

- Chain rule

$$P(f_1 \wedge f_2 \wedge \ldots \wedge f_n)$$
$$= P(f_1) \times P(f_2 \mid f_1) \times P(f_3 \mid f_1 \wedge f_2)$$
$$\times \ldots \times P(f_n \mid f_1 \wedge \ldots . \wedge f_{n-1})$$
$$= \prod_{i=1}^{n} P(f_i \mid f_1 \wedge \ldots . \wedge f_{i-1})$$

Law of total probability

- use a weighted average of conditional probabilities to calculate a probability

Propositions W,J

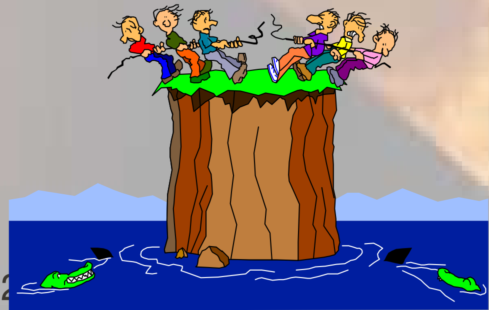
$W = (W \cap J) \cup (W \cap (\sim J))$ (disjointness)

$$\left.\begin{aligned} P(W) &= P(W \cap J) + P(W \cap (\sim J)) \\ P(W \cap J) &= P(W \mid J)P(J) \\ P(W \cap (\sim J)) &= P(W \mid \sim J)P(\sim J) \end{aligned}\right\} \Rightarrow$$

$$P(W) = P(W \mid J)P(J) + P(W \mid \sim J)P(\sim J)$$

Beating classifiers

- in LEE-97 Cup your favorite classifier program YFC is tested against decision tree classifier (DTC) and Naïve Bayes classifier (NBC).
- from experience probability of YFC performing better than DTC is 7/10 but against NBC only 2/10
- you assess the probability of YFC facing DTC to be 1/4 (and thus P(NBC) to be 3/4)
- How likely is it that your YFC is better than your next competitor?



Beating classifiers

Propositions J = against DTC, ~ J = against NBC,

B = beating

$$P(B \mid J) = 7/10, P(B \mid \sim J) = 2/10$$

$$P(J) = 1/4, P(\sim J) = 3/4$$

By the law of total probability

$$P(B) = \frac{1}{4} \cdot \frac{7}{10} + \frac{3}{4} \cdot \frac{2}{10} = \frac{13}{40} = .325$$

Bayes Rule

- let us reverse the previous situation: assuming same probability assignments, you tell me that YFC outperformed the competitor. Which classifier (DTC or NBC) did you compete with?

Bayes Rule

$$P(J \mid W) = \frac{P(W \cap J)}{P(W)}$$

from chain rule we have

$$P(J \mid W) = \frac{P(W \mid J)P(J)}{P(W)}$$

using the law of total probability we have

$$P(J \mid W) = \frac{P(W \mid J)P(J)}{P(W \mid J)P(J) + P(W \mid \sim J)P(\sim J)}$$

In the example …

Propositions J = against DTC, ~ J = against NBC,

B = beating

$P(B \mid J) = 7/10, P(B \mid \sim J) = 2/10$

$P(J) = 1/4, P(\sim J) = 3/4$

By Bayes Rule

$$P(J \mid B) = \frac{(7/10)(1/4)}{(7/10)(1/4) + (2/10)(3/4)} = \frac{7}{13}$$

Side note

- Note that P(B |J) + P(B | ~J) is not necessarily 1, but P(B | J) + P(~B | J) is!
- Why?
  - Intuitive argument
    - √ If you now that you will surely beat DTC (P(B | J) =1), this does NOT mean that you will surely loose against NBC, right?
  - Semantic argument:
    - √ if you condition on two different things, you obtain two different measures on the possible worlds ⇨ two different distributions
- Mnemonic rule of thumb:
  - Given P(A | B):
    - √ If you sum over the *left* hand side (A), the result is 1
    - √ If you sum over the *right* hand side (B), the result can be anything
    - √ If you sum over the values of B weighted by P(B), the result is P(A)

Marginalization

A	B	C	P(A,B,C)	P(A=1)?	P(A=1,B=1)?	P(B=1 | A=1)?
1	1	1	0,10	X	X	
1	0	0	0,20	X		
1	1	0	0,05	X	X	
1	0	1	0,15	X		
0	1	1	0,30			
0	0	0	0,05			
0	1	0	0,10			
0	0	1	0,05			
			1,0	0,50	0,15	0,15/0,50 = 0,30

Probabilistic inference = marginalization

- H: something you do not know and want to know
- U: everything you do not know and do not need to know ("nuisance")
- I: background knowledge
- D: observed data

$$P(H|D,I) = \sum_{U} P(H|D,I,U)P(U|D,I)$$

Another version of BR

$$P(J \mid W) = \frac{P(W \mid J)P(J)}{P(W \mid J)P(J) + P(W \mid \sim J)P(\sim J)}$$

divide by the numerator

$$P(J \mid W) = \frac{1}{1 + \dfrac{P(W \mid \sim J)}{P(W \mid J)} \dfrac{P(\sim J)}{P(J)}}$$

If we express $P(\sim J \mid W)$ with this expression,

we get

$$\frac{P(\sim J \mid W)}{P(J \mid W)} = \frac{P(W \mid \sim J)}{P(W \mid J)} \frac{P(\sim J)}{P(J)}$$

Bayes Rule in Terms of Odds

Bayes factor in favor of J
(**likelihood**)

$$\frac{P(J \mid W)}{P(\sim J \mid W)} = \frac{P(W \mid J)}{P(W \mid \sim J)} \frac{P(J)}{P(\sim J)}$$

Posterior odds in favor of J

Prior odds in favor of J

Example: Which die?

- two dice: 4-sider and 20-sider
- each side equally likely (for each die)
- F = “pick 4-sider”, T=~F=“pick 20-sider”
- for you P(F)=P(~F)=1/2
- I roll the die picked. The result is 3. Which die did I pick?

Example: Which die?

likelihoods are $P(3 | F) = 1/4$ and $P(3 | T) = 1/20$

Bayes's rule says

$$P(F | 3) = \frac{P(3 | F)P(F)}{P(3 | F)P(F) + P(3 | T)P(T)}$$

$$= \frac{(1/4)(1/2)}{(1/4)(1/2) + (1/20)(1/2)} = \frac{5}{6}$$

The Bayes factor in favor of F

$$\frac{P(3 | F)}{P(3 | T)} = \frac{1/4}{1/20} = 5$$