6.034
Probabilistic Reasoning

Peter Szolovits

ai6034.mit.edu

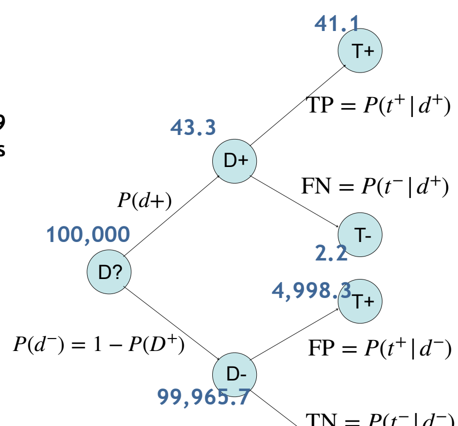
November 8, 2019

Probabilistic Reasoning

- 1970's, Dr. William Schwartz posed this problem at a medical convention:
  - MIT biological researchers create a new, simple test for cancer
  - 5% error rate
  - Give test to a random MIT student, with, alas, a positive result
  - What is the probability that student actually has cancer?
- Common answers:
  - 95%, 50%
- What is the real answer?

How certain are we after a test?

- How common is cancer?
  - "cancers as those occurring between the ages of 20 and 39 years … 43.3 new cancer cases per 100 000 people per year"
  - Thus, prevalence ≈ 0.0433%, or 0.000433
- "Accuracy" refers to either false positives or false negatives; assume
  - $P(t^+ | d^+) = 0.95$, $P(t^- | d^+) = 0.05$, $P(t^- | d^-) = 0.95$, $P(t^+ | d^-) = 0.05$,



$$\frac{41.1}{41.1 + 4998.3} = .00082$$

https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(17)30677-0/fulltext

---

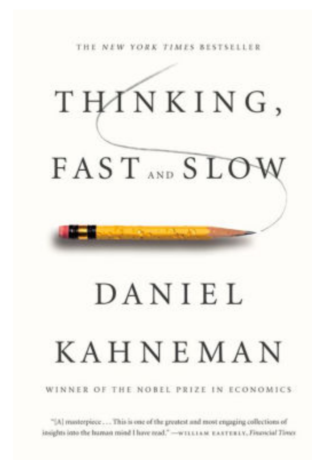
- Doctors (and people, in general) are lousy at probabilistic reasoning

Judgment under Uncertainty: Heuristics and Biases

Biases in judgments reveal some heuristics of thinking under uncertainty.

Amos Tversky and Daniel Kahneman

Science, 1974



2011

Lessons

- Tversky & Kahneman, 1974, "Judgment under Uncertainty: Heuristics and Biases"
  - Representativeness
    - "Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail."
    - What is his profession?
farmer, salesman, airline pilot, librarian, or physician
  - Availability
  - Anchoring

https://science.sciencemag.org/content/sci/185/4157/1124.full.pdf

Lessons

- Tversky & Kahneman, 1974, "Judgment under Uncertainty: Heuristics and Biases"
  - Representativeness
  - Availability — what comes to mind?
    - instances of large classes are usually recalled better and faster than instances of less frequent classes
    - "Different lists were presented to different groups of subjects. In some of the lists the men were relatively more famous than the women, and in other the women were relatively more famous than the men. In each of the lists, the subjects erroneously judged that the class (sex) that had the more famous personalities was the more numerous.
  - Anchoring

Lessons

- Tversky & Kahneman, 1974, "Judgment under Uncertainty: Heuristics and Biases"
  - Representativeness
  - Availability
  - Anchoring
    - "subjects were asked to estimate … the percentage of African countries in the United Nations.
    - "… a number between 0 and 100 was determined by spinning a wheel of fortune in the subject's presence. The subjects were instructed to indicate first whether that number was higher or lower than the value of the quantity, and then to estimate the quantity.
    - "… these arbitrary numbers had a marked effect on estimates. … the median estimates … were 25 and 45 for groups that received 10 and 65, respectively, as starting points. Payoffs for accuracy did not reduce the anchoring effect."

Interpreting a Test

Relationship between true state of the world and a diagnostic test

	Test Positive	*Test Negative*	
Disease Present	True Positive	False Negative	TP+FN
Disease Absent	False Positive	True Negative	FP+TN
	TP+FP	FN+TN	

	Test Positive	*Test Negative*	
Disease Present	True Positive	False Negative	TP+FN
Disease Absent	False Positive	True Negative	FP+TN
	TP+FP	FN+TN	

Definitions

Sensitivity (true positive rate): TP/(TP+FN)

False negative rate: 1-Sensitivity = FN/(TP+FN)

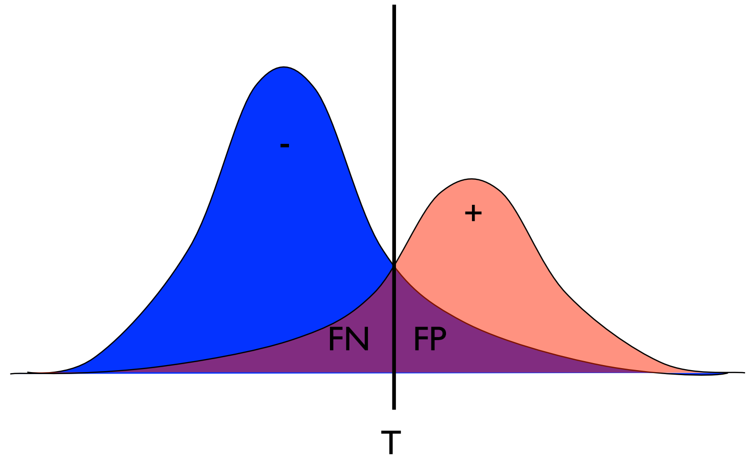
Specificity (true negative rate): TN/(FP+TN)

False positive rate: 1-Specificity = FP/(FP+TN)

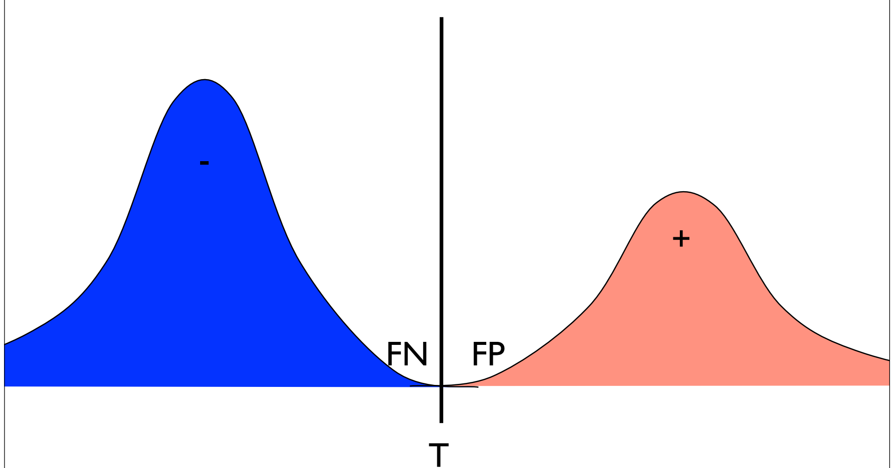
Positive Predictive Value: TP/(TP+FP)

Negative Predictive Value: TN/(FN+TN)

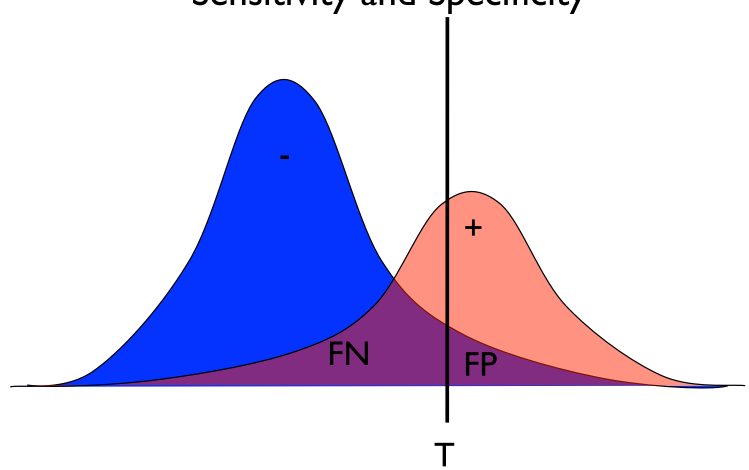
Thresholds for Continuous Tests



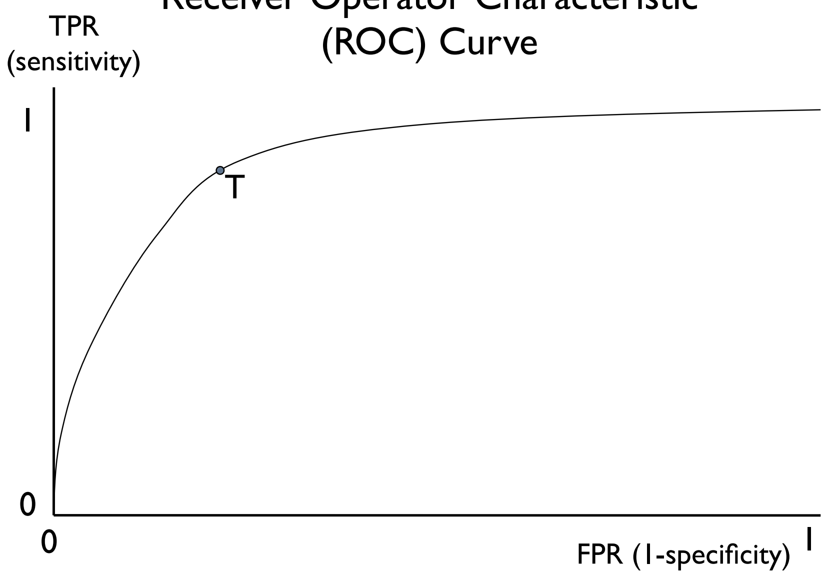
Wonderful Test



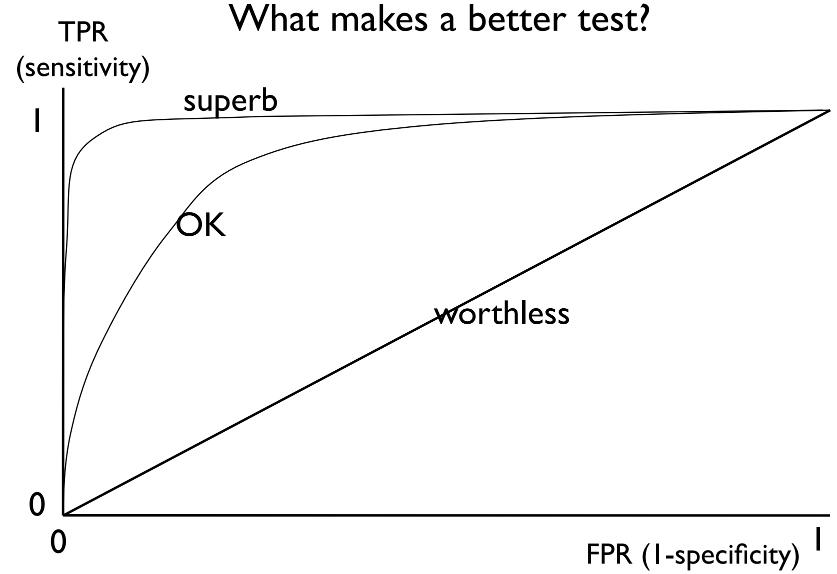
Test Thresholds Change Trade-off between Sensitivity and Specificity



Receiver Operator Characteristic (ROC) Curve



What makes a better test?



Thinking about Multiple Variables

Dog barks	Burglar	Raccoon	Tally	P	Selected
false	false	false	405	0.405	☐
false	false	true	225	0.225	☐
false	true	false	0	0.000	☑
false	true	true	0	0.000	☑
true	false	false	45	0.045	☐
true	false	true	225	0.225	☐
true	true	false	40	0.040	☑
true	true	true	60	0.060	☑
○T ○F ●?	○T ○F ●?	○T ○F ●?	1000	1.000	0.100

Discrete Random Variables

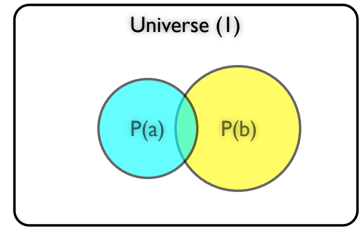
- E is a (Boolean) random variable if it denotes an uncertain event
  - You will receive a grade of "A" in this class
  - Global warming will cause Florida to be under water by 2100
  - Your fever is caused by malaria
- We can extend this to discrete variables with more than two possible values
- Random variables can also be continuous
  - Your first child will be 6'3" in adulthood
- Sources of uncertainty
  - Lack of knowledge ("Is population of Bhutan > 1M?")
  - Imperfection of models (climate change and Florida)
  - Physics (radioactive decay)

Meaning(s) of Probability

- P(E): fraction of "possible worlds" in which E is true
  - The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible. —Laplace
- Interpretations of probability (*semester course in philosophy…*)
  - Physical/Objective — related to random events
    - Frequentist
      - measure long runs of experimental trials; e.g., coin flipping
    - Propensity/Inductive
      - Quantum Mechanics; predicted by models, supported by evidence
  - Subjective/Belief

A Review of Probabilities

- Axioms of Probability
  1. $0 \leq P(a) \leq 1$
  2. $P(\text{TRUE}) = 1$
     $P(\text{FALSE}) = 0$
  3. $P(a) + P(b) - P(a, b) = P(a \vee b)$



Venn Diagram

- Useful Theorems
  - $P(\neg a) = 1 - P(a)$
  - $P(a) = P(a, b) + P(a, \neg b)$
  - $P(a) + P(b) - P(a \vee b) = P(a, b)$

Multi-Valued Random Variables

- A is a random variable with <u>arity</u> k if it can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$

$P(A = v_i, A = v_j) = 0$ if $i \neq j$

$P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k) = 1$

$$P(A = v_1 \vee A = v_2 \vee \ldots \vee A = v_i) = \sum_{j=1}^{i} P(A = v_j)$$

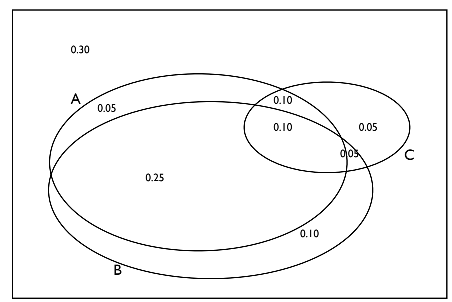
thus, $\sum_{j=1}^{k} P(A = v_j) = 1$

$$P(B, A = v_1 \vee A = v_2 \vee \ldots \vee A = v_i) = \sum_{j=1}^{i} P(B, A = v_j)$$

thus, $P(B) = \sum_{j=1}^{k} P(B, A = v_j)$

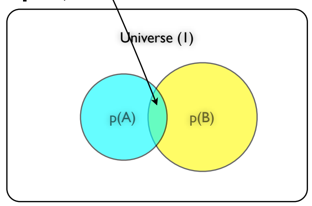
Joint Distribution with 3 Variables

A	B	C	*Prob*
0	0	0	0.3
0	0	1	0.05
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.25
1	1	1	0.1



Conditional Probability

- $P(a \mid b)$ = fraction of possible worlds in which b is true in which a is also true
- $P(a \mid b) = P(a, b)/p(b)$
- thus, $P(a, b) = P(b)P(a \mid b)$



Chain Rule

- What about $P(a, b, c)$?
  - $y = b, c$
  - $P(a, b, c) = P(a, y) = P(a \mid y)P(y)$
  - $= P(a \mid b, c)P(b, c)$
  - $= P(a \mid b, c)P(b \mid c)P(c)$
- Chain rule: nothing depends on anything to its left
- $P(x_1, \ldots, x_n) = \prod_{i=n}^{1} P(x_i \mid x_{i-1}, \ldots x_1)$

Independence (and Conditional Independence)

- Definition of ***independence***
  - $P(a \mid b) \equiv P(a)$ iff a is independent of b
- Implies
  - $P(a, b) = P(a)P(b)$
- Definition of ***conditional independence***
  - $P(a \mid b, z) \equiv P(a \mid z)$ iff a is independent of b given z
- Implies
  - $P(a, b \mid z) = P(a \mid z)P(b \mid z)$

"Naive" Bayes

Bayes' Rule

- $P(d_j \mid s_i) = \dfrac{P(d_j)P(s_i \mid d_j)}{P(s_i)}$
- where $P(s_i) = \sum_{k=i}^{n} P(d_k)P(s_i \mid d_k)$
- By conditional independence,

$$P(d_j \mid s_1, \ldots, s_k) = \frac{P(d_j)P(s_1, \ldots, s_k \mid d_j)}{P(s_1, \ldots, s_k)}$$

$$= \frac{P(d_j)}{P(s_1, \ldots, s_k)} \prod_l P(s_l \mid d_j)$$

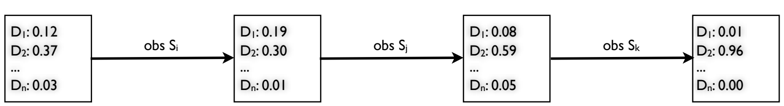
- Note: $P(s_1, \ldots, s_k)$ is the same for all d_j, just to normalize P

Diagnostic Reasoning

- Assume patient has exactly one disease
  - modeled as a multi-valued random variable
  - *Example: Acute renal (kidney) failure*; i.e., you stop peeing
    - Possible diagnoses: infection, blockage, low blood flow, …
- Can have many signs/symptoms
  - *Example: pain*, *blood pressure*, *blood in urine*, *strep infection*, *sodium concentration*, …
- Assume that the symptoms are all conditionally independent given the disease.

Diagnostic Reasoning with Naive Bayes

- Exploit assumption of conditional independence among symptoms
  $P(s_1, s_2, \ldots, s_k \mid d_j) = P(s_1 \mid d_j)P(s_2 \mid d_j)\ldots P(s_k \mid d_j)$
- Sequence of observations of symptoms, s_i, each revise the distribution via Bayes' Rule
- $P^j(d_i \mid s_1, \ldots, s_j) = \dfrac{P^{j-1}(d_i)P(s_j \mid d_i)}{P^{j-1}(s_j)} = \dfrac{P^{j-1}(d_i)P(s_j \mid d_i)}{\sum_{i=0}^{n} P^{j-1}(d_i)P(s_j \mid d_i)}$



Data Represented in (Conditional) Probability Tables

Priors

	Prior Probability
D_1	0.25
D_2	0.4
…	…
D_m	0.05

Conditionals

S_1 = Cough	P(Cough|D_i)
D_1	0.001
D_2	0.9
…	
D_m	0.4

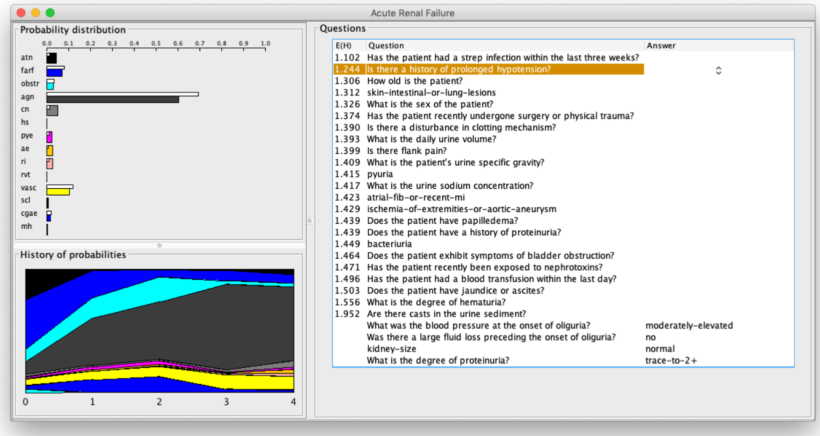
S_2 = Fever	P(F=none|D_i)	P(F=mild|D_i)	P(F=severe|D_i)
D_1	0.05	0.8	0.15
D_2	0.9	0.07	0.03
…			
D_m	0.4	0.4	0.2

How to choose which observation to make next? Entropy Redux

- Compute the expected entropy of $P(d_i)$ after requesting each possible observation
  $q = \arg\min_j E(H(P(d \mid s_j)))$
  - For each observation, s_j, we can get n_j possible answers
    - For each answer, we can compute the revised (by Bayes rule) posterior probability distribution
    - For that distribution, we compute its entropy
    - The expected entropy weights these entropies by the probability that we would get that answer if we asked that question, namely

$$P(s_j = v_k) = \sum_{i=1}^{n} P(d_i)P(s_j = v_k \mid d_i)$$

Acute Renal Failure Program



Gorry, G. A., Kassirer, J. P., Essig, A., & Schwartz, W. B. (1973). Decision analysis as the basis for computer-aided management of acute renal failure. The American Journal of Medicine, 55(3), 473-484. Modern GUI by P. Szolovits.

Odds-Likelihood

- In gambling, "3-to-1" odds means 75% chance of success
  $O = P/(1-P) = P/\neg P$
  - P = 0.5 means O=1
- Likelihood ratio $L(s \mid d) = P(s \mid d)/P(s \mid \neg d)$
- Odds-likelihood form of Bayes rule
  $O(d \mid s_1, \ldots, s_n) = O(d)L(s_1 \mid d)\ldots L(s_n \mid d)$
- Log transform to weighted sum
  - basis for many clinical scoring systems

$$\log[O(d \mid s_1, \ldots, s_n)[= \log[O(d)L(s_1 \mid d)\ldots L(s_n \mid d)]$$
$$= \log[O(d)] + \log[O(s_1 \mid d)] + \ldots + \log[O(s_n \mid d)]$$
$$= W(d) + W(s_1 \mid d) + \ldots + W(s_n \mid d)$$