CS486/686: Introduction to Artificial Intelligence
Lecture 7a - Probability and Bayesian Networks

Jesse Hoey & Victor Zhong

School of Computer Science, University of Waterloo

February 3, 2025

Readings: Poole & Mackworth Chap. 9-9.6

Uncertainty

Why is uncertainty important?

- Agents (and humans) don't know **everything**,
  but need to **make decisions** anyways!
- Decisions are made in the **absence of information**,
  or in the presence of **noisy** information (sensor readings)

The best an agent can do:

Know how uncertain it is, and act accordingly

Probability: Frequentist vs. Bayesian

Frequentist view:

probability of heads = # of heads / # of flips

probability of heads **this time** = probability of heads (history)

Uncertainty is **ontological**: pertaining to the world

Bayesian view:

probability of heads **this time** = agent's **belief** about flip

belief of agent A : based on **previous experience** of agent A

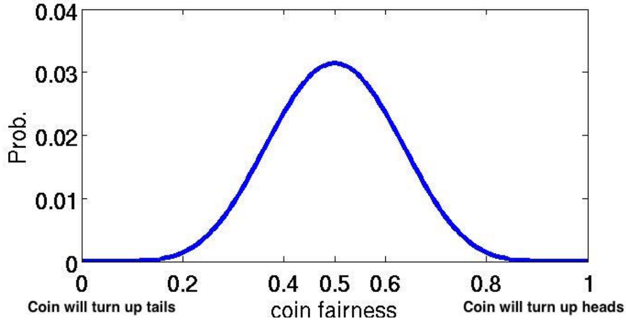
Uncertainty is **epistemological**: pertaining to knowledge

Probability: Bayesian

Bayesian probability

all else being equal (Prior)

before any flips

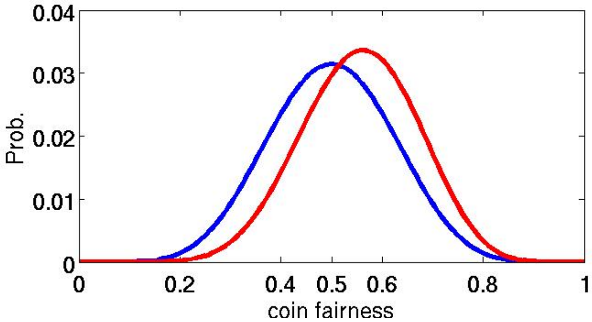


Probability: Bayesian

Bayesian probability

all else being equal (Prior)

after 2 flips **heads, heads** (Posterior)

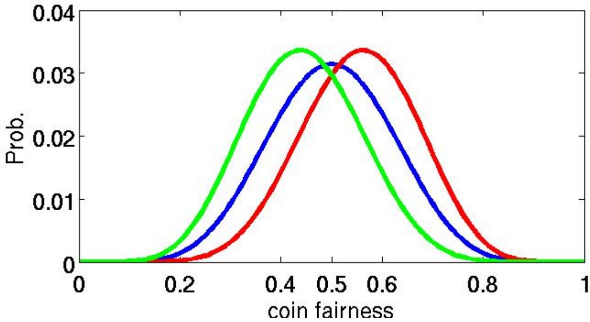


Probability: Bayesian

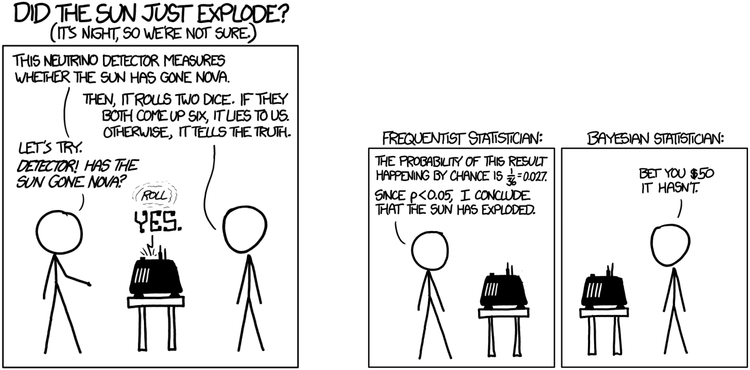
Bayesian probability

all else being equal (Prior)

after 2 flips **tails,tails** (Posterior)



Probability: Bayesian



Probability

Should you **wear your seatbelt**? Estimate $P(injury)$ given you do/don't wear it

Probability

Should you **wear your seatbelt**? Estimate $P(injury)$ given you do/don't wear it

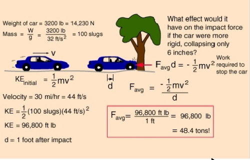
Frequentist:

test	day	result	P(injury)
-	Sunday (prior to start)	-	?
1	Monday	OK	0.0
2	Tuesday	OK	0.0
3	Tuesday	Crash	0.33333
4	Thursday	OK	0.25
5	Friday	OK	0.2
...	...	...	...
N		Crash	Number of injuries / N

Probability

Should you **wear your seatbelt**? Estimate $P(injury)$ given you do/don't wear it

Bayesian:



UK government:
"Seatbelts save 2200 lives/year"

car–accidents.com:
"in 63% of fatalities, no seatbelts were worn"

Probability Measure

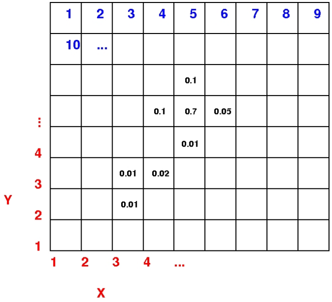
if X is a **random variable** (feature, attribute),

it can take on values x, where $x \in Domain(X)$ (or $Dom(X)$)

Assume x is discrete,
$\mathbf{P}(\mathbf{x})$ is the probability that $X = x$

Joint probability $\mathbf{P}(\mathbf{x}, \mathbf{y})$ is the
probability that $X = x$ and $Y = y$ at the same time

Joint probability distribution:



Where is the robot?
Features: X,Y

Axioms of Probability

Axioms are things we have to assume about probability:

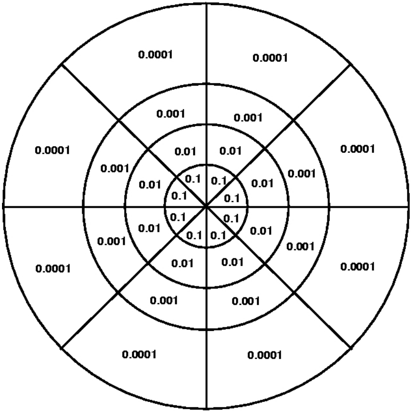
- $P(X) \geq 0$
- $\sum_x P(X = x) = 1.0$
- $P(a \vee b) = P(a) + P(b)$ if a and b are contradictory - can't both be true at the same time e.g. $P(win \vee lose) = P(win) + P(lose) = 1.0$

Some notes:

- probability between 0-1 is purely **convention**
- $P(a) = 0$ means you think a is **definitely false**
- $P(a) = 1$ means you think a is **definitely true**
- $0 < P(a) < 1$ means you have **belief** about the truth of a. It does **not** mean that a is true to some degree, just that you are ignorant of its truth value.
- Probability = measure of **ignorance**

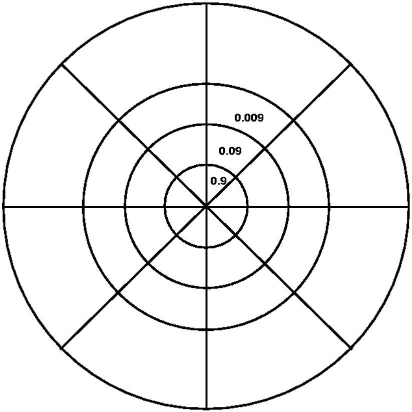
Independence

- describe a system with n features: $2^n - 1$ probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, P(hit a sector)
- $P(sector) = P(r, \theta)$ where $r = 1, \ldots, 4$ and $\theta = 1, \ldots, 8$.
- 32 sectors in total - need to give 31 numbers



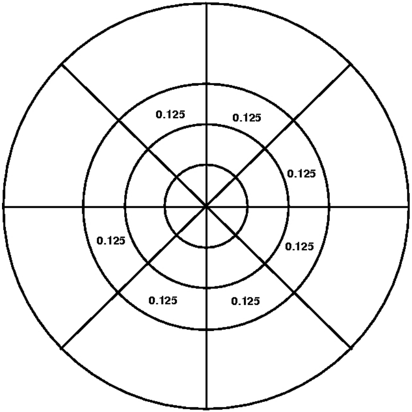
Independence

- describe a system with n features: $2^n - 1$ probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, P(hit a sector)
- assume **radial independence**: $P(r, \theta) = P(r)P(\theta)$
- only need 7+3=10 numbers



Independence

- describe a system with n features: $2^n - 1$ probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, P(hit a sector)
- assume **radial independence**: $P(r, \theta) = P(r)P(\theta)$
- only need 7+3=10 numbers



Conditional Probability

if X and Y are random variables, then

$P(x|y)$ is the probability that $X = x$ **given** that $Y = y$.

e.g.

$P(\mathit{flies}|\mathit{is\_bird})$ is different than $P(\mathit{flies})$

Incorporate independence:

$P(\mathit{flies}|\mathit{is\_bird}, \mathit{has\_feathers}) = P(\mathit{flies}|\mathit{is\_bird})$ if *flies* and *has_feathers* are independent given *is_bird*

Product rule (Chain rule):

$P(\mathit{flies}, \mathit{is\_bird}) = P(\mathit{flies}|\mathit{is\_bird})P(\mathit{is\_bird})$

$P(\mathit{flies}, \mathit{is\_bird}) = P(\mathit{is\_bird}|\mathit{flies})P(\mathit{flies})$

leads to : Bayes' rule $\quad P(\mathit{is\_bird}|\mathit{flies}) = \frac{P(\mathit{flies}|\mathit{is\_bird})P(\mathit{is\_bird})}{P(\mathit{flies})}$

Sum Rule

We know (an Axiom):

$$\sum_x P(X = x) = 1.0 \quad \text{and therefore that} \quad \sum_x P(X = x|Y) = 1.0$$

This means that **(Sum Rule)**

$$\sum_x P(X = x, Y) = P(Y)$$

proof:

$$\begin{aligned}\sum_x P(X = x, Y) &= \sum_x P(X = x|Y)P(Y) \\ &= P(Y)\sum_x P(X = x|Y) \\ &= P(Y)\end{aligned}$$

We call $P(Y)$ the **marginal** distribution over Y

Conditional Indenpendence

- X and Y are **independent** iff

$$P(X) = P(X|Y)$$

$$P(Y) = P(Y|X)$$

$$P(X, Y) = P(X)P(Y)$$

so learning Y doesn't influence beliefs about X

- X and Y are **conditionally independent** given Z iff

$$P(X|Z) = P(X|Y, Z)$$

$$P(Y|Z) = P(Y|X, Z)$$

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

so learning Y **doesn't influence beliefs** about X **if you already know** Z...does **not** mean X and Y are independent

Expected Values

Expected value of a function on X, $V(X)$:

$\mathbb{E}(V) = \sum_{x \in Dom(X)} P(x)V(x)$

where $P(x)$ is the probability that $X = x$

This is useful in **decision making**, where $V(X)$ is the utility of situation X

Bayesian decision making is then

$\mathbb{E}(V(\texttt{decision})) = \sum_{outcome} P(outcome|decision)V(outcome)$

Value of Independence

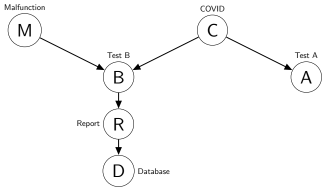
- complete independence reduces both **representation** and **inference** from $O(2^n)$ to $O(n)$
- Unfortunately, complete mutual independence is **rare**
- Fortunately, most domains do exhibit a fair amount of **conditional independence**
- **Bayesian Networks** or **Belief Networks** (BNs) encode this information

Belief Networks

Bayesian network or **belief network**

- Directed Acyclic graph
- Encodes independencies in a graphical format
- Edges give $P(X_i | parents(X_i))$

COVID diagnosis example:



- Two tests A and B
- Test A is quick and cheap, but has a high false positive rate
- Test A results are read directly
- Test B uses a machine that sometimes malfunctions, but has a lower FP rate
- Test B results are not read directly,
- a Report is written (by a human who makes mistakes)
- the Report is entered into a database (by another human who makes mistakes)

Correlation and Causality

- Directed links in Bayes' net $\approx$ **causal**
- However, **not always** the case:
  chocolate $\rightarrow$ Nobel or
  Nobel $\rightarrow$ chocolate?
- In a Bayes net, it doesn't matter!
- But, some structures will be **easier to specify**

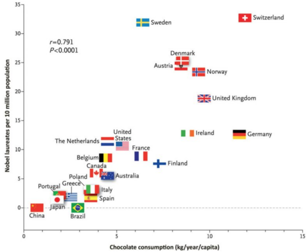


Figure 1. Taken from Messerli (2012). Correlation between countries' annual per capita chocolate consumption and the number of Nobel laureates per 10 million population. Copyright "The New England Journal of Medicine".

In this example, its probably
chocolate $\leftarrow$ *"Switzerland − ness"* $\rightarrow$ *Nobel*

Bayesian Networks - Example

If Jesse's alarm doesn't go off (A), Jesse probably won't get coffee (C); if Jesse doesn't get coffee, he's likely grumpy (G). If he is grumpy then it's possible that the lecture won't go smoothly L. If the lecture does not go smoothly then the students will likely be sad S.

A=Jesse's alarm doesn't go off
C=Jesse doesn't get coffee
G=Jesse is grumpy
L=lecture doesn't go smoothly
S=students are sad

all variables **binary (true/false)**

Conditional Independence

- If you learned any of A, C, G, or L, would your assessment of $P(S)$ **change?**
  - If any of these are seen to be true, you would increase $P(s)$and decrease $P(\bar{s})$
  - So S is **not independent** of A, C, G, L
- If you knew the value of L, would learning the value of A, C, or G influence $P(S)$?
  - Influence that these factors have on S is mediated by their influence on L
  - Students aren't sad because Jesse was grumpy, they are sad because of the lecture
  - Therefore, S is **conditionally independent** of A, C, and G (given L)

Conditional Independence

$$A \rightarrow C \rightarrow G \rightarrow L \rightarrow S$$

- We say: S is **independent** of A, C, and G, **given** L
- (this is **conditional independence**)
- Similarly, we can say
  - S is **independent** of A and C, given G
  - G is **independent** of A, given C
  - ...
- This means that:
  - $P(S|L, G, C, A) = P(S|L)$
  - $P(L|G, C, A) = P(L|G)$
  - $P(G|C, A) = P(G|C)$
  - $P(C|A)$ and $P(A)$ don't "simplify"

Conditional Independence

Chain rule (**product rule**):

$$P(S, L, G, C, A) = P(S|L, G, C, A)P(L|G, C, A)P(G|C, A)P(C|A)P(A)$$

Independence:

$$P(S, L, G, C, A) = P(S|L)P(L|G)P(G|C)P(C|A)P(A)$$

So we can specify the full **joint probability**
using the five local **conditional probabilities**:

$$P(S|L), P(L|G), P(G|C), P(C|A), P(A)$$

Bayesian Networks

A **Bayesian Network** (Belief Network, Probabilistic Network) or BN over variables $\{X_1, X_2, \ldots, X_N\}$ consists of:

- a **DAG** whose nodes are the variables
- a set of **Conditional Probability tables** (CPTs) giving $P(X_i|Parents(X_i))$ for each X_i

Example probability tables for the Coffee Bayes Net:

$\mathbf{P}(\mathbf{A} = \mathbf{true}) = 0.3$

$\mathbf{P}(\mathbf{C} = \mathbf{true}|\mathbf{A}) =$

A	$P(C = true\|A)$
t	0.8
f	0.15

$\mathbf{P}(\mathbf{G} = \mathbf{true}|\mathbf{C}) =$

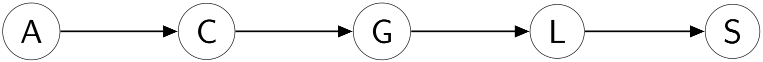
C	$P(G = true\|C)$
t	1.0
f	0.2

$\mathbf{P}(\mathbf{L} = \mathbf{true}|\mathbf{G}) =$

G	$P(L = true\|G)$
t	0.7
f	0.2

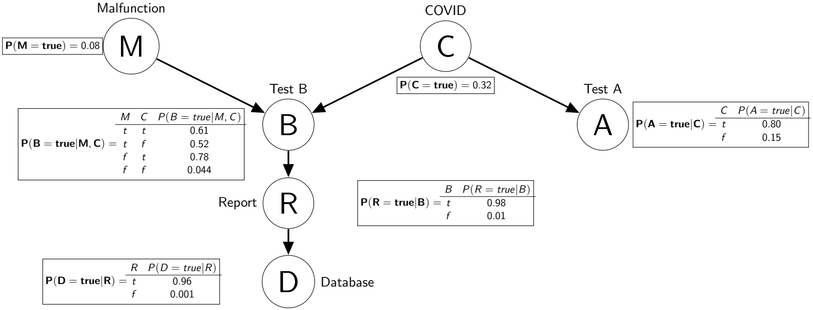
$\mathbf{P}(\mathbf{S} = \mathbf{true}|\mathbf{L}) =$

L	$P(S = true\|L)$
t	0.9
f	0.3



Another Example Quantification

COVID diagnosis:



$\mathbf{P}(\mathbf{B} = \mathbf{true} | \mathbf{M}, \mathbf{C}) =$

M	C	$P(B = true \| M, C)$
t	t	0.61
t	f	0.52
f	t	0.78
f	f	0.044

$\mathbf{P}(\mathbf{A} = \mathbf{true} | \mathbf{C}) =$

C	$P(A = true \| C)$
t	0.80
f	0.15

$\mathbf{P}(\mathbf{R} = \mathbf{true} | \mathbf{B}) =$

B	$P(R = true \| B)$
t	0.98
f	0.01

$\mathbf{P}(\mathbf{D} = \mathbf{true} | \mathbf{R}) =$

R	$P(D = true \| R)$
t	0.96
f	0.001

Semantics of a Bayes' Net

The structure of the BN means that:

every X_i is **conditionally independent** of all its **nondescendants** given its parents:

$$P(X_i|S, Parents(X_i)) = P(X_i|Parents(X_i))$$

for any subset $S \subseteq NonDescendants(X_i)$

The BN defines a **factorization** of the **joint probability** distribution. The joint distribution is formed by multiplying the conditional probability tables together.

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i|parents(X_i))$$

Next

Uncertainty (cont.): BN Construction and Inference