Probability Basics

Sources of Uncertainty

The world is a very uncertain place...

- Uncertain **inputs**
  - Missing data
  - Noisy data
- Uncertain **knowledge**
  - Multiple causes lead to multiple effects
  - Incomplete enumeration of conditions or effects
  - Incomplete knowledge of causality in the domain
  - Stochastic effects
- Uncertain **outputs**
  - Abduction and induction are inherently uncertain
  - Incomplete deductive inference may be uncertain

Probabilities

- 30 years of AI research danced around the fact that the world was inherently uncertain

- Bayesian Inference:
  - Use probability theory and information about independence
  - Reason diagnostically (from evidence (effects) to conclusions (causes))...
  - ...or causally (from causes to effects)

- Probabilistic reasoning only gives probabilistic results
  - i.e., it summarizes uncertainty from various sources

Discrete Random Variables

- Let A denote a random variable
  - A represents an event that can take on certain values
  - Each value has an associated probability

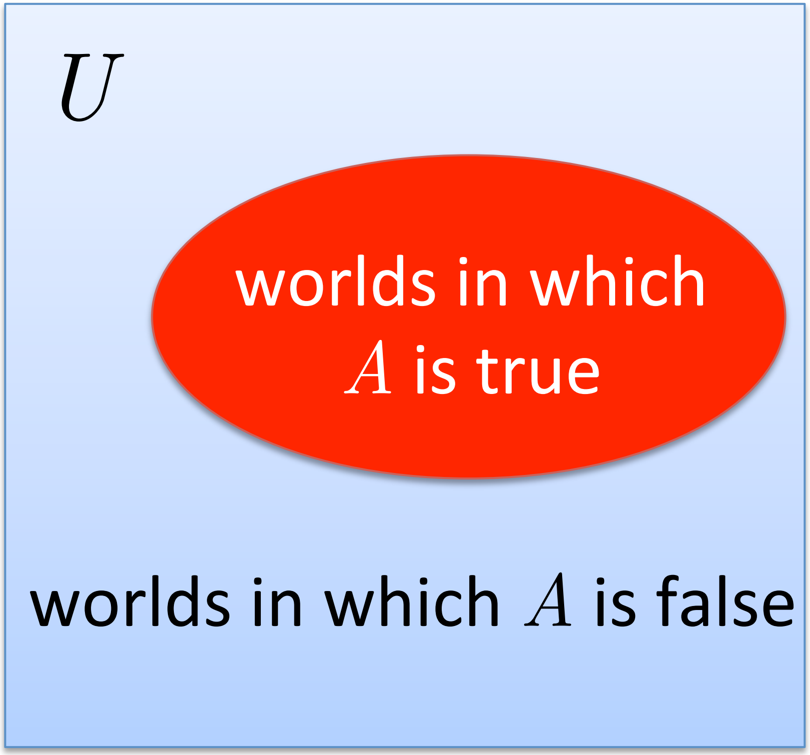
- Examples of binary random variables:
  - A = I have a headache
  - A = Sally will be the US president in 2020

- $\mathrm{P}(A)$ is "the fraction of possible worlds in which A is true"
  - We could spend hours on the philosophy of this, but we won't

Adapted from slide by Andrew Moore

Visualizing A

- Universe U is the event space of all possible worlds
  - Its area is 1
  - $\mathrm{P}(U)$ = 1



- $\mathrm{P}(A)$ = area of red oval

- Therefore:

$$P(A) + P(\neg A) = 1$$

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

Axioms of Probability

Kolmogorov showed that three simple axioms lead to the rules of probability theory

- de Finetti, Cox, and Carnap have also provided compelling arguments for these axioms

1. All probabilities are between 0 and 1:

$$0 \leq \mathrm{P}(A) \leq 1$$

2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0:

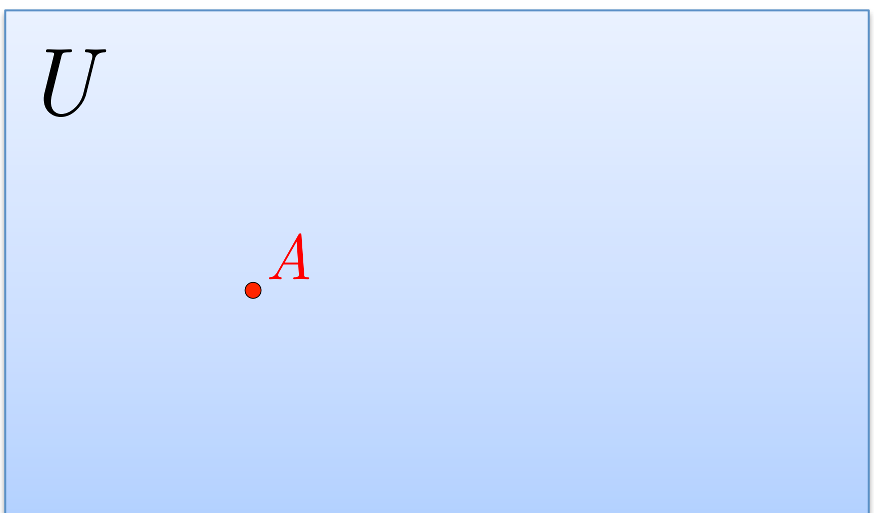
$$\mathrm{P}(\mathrm{true}) = 1 \; ; \quad \mathrm{P}(\mathrm{false}) = 0$$

3. The probability of a disjunction is given by:

$$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$$

Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

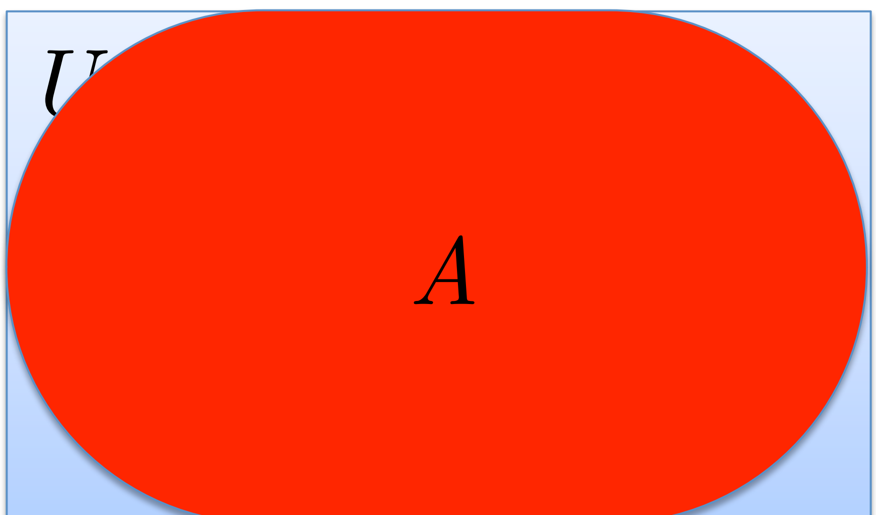


The area of A can't get any smaller than 0

A zero area would mean no world could ever have A true

Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

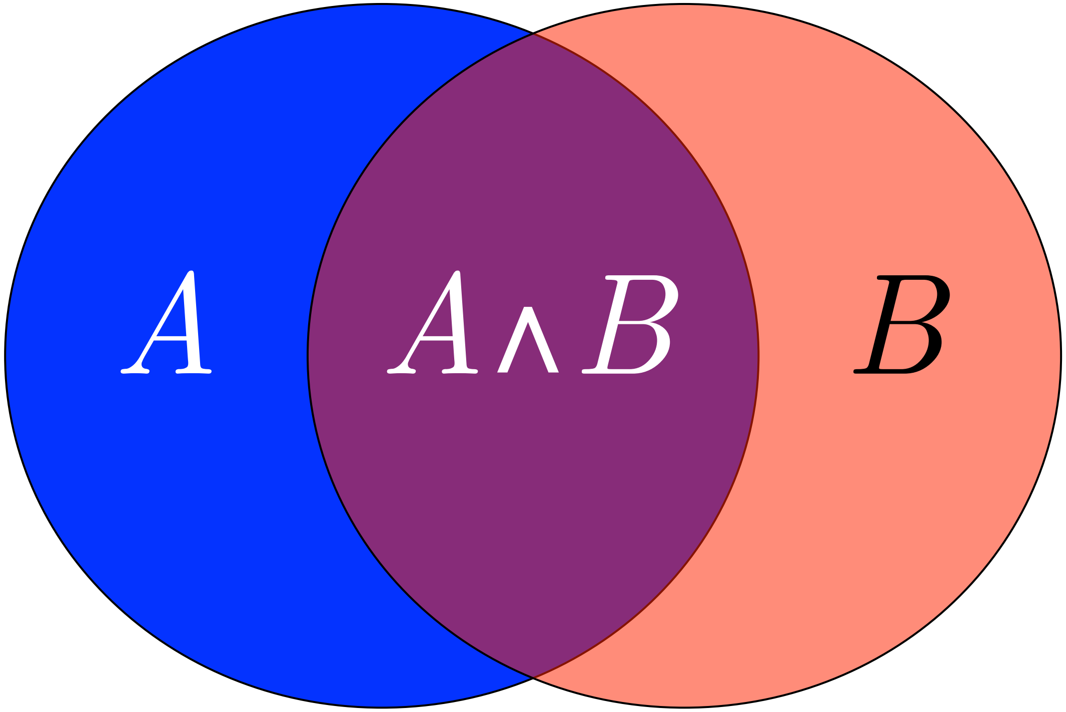


The area of A can't get any bigger than 1

An area of 1 would mean A is true in all possible worlds

Interpreting the Axioms

- $0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(\mathrm{true}) = 1$
- $\mathrm{P}(\mathrm{false}) = 0$
- $\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$



These Axioms are Not to be Trifled With

- There have been attempts to develop different methodologies for uncertainty:
  - Fuzzy Logic
  - Three-valued logic
  - Dempster-Shafer
  - Non-monotonic reasoning

- But the axioms of probability are the only system with this property:

  If you gamble using them you can't be unfairly exploited by an opponent using some other system [di Finetti, 1931]

An Important Theorem

$0 \leq \mathrm{P}(A) \leq 1$

$\mathrm{P}(\text{true}) = 1; \quad \mathrm{P}(\text{false}) = 0$

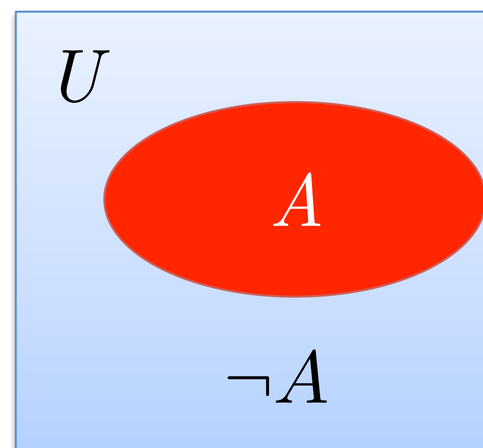
$\mathrm{P}(A \lor B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \land B)$

From these we can prove:

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

Proof: Let $B = \neg A$. Then, we have

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$
$$P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A)$$
$$P(\text{true}) = P(A) + P(\neg A) - P(\text{false})$$
$$1 = P(A) + P(\neg A) - 0$$
$$P(\neg A) = 1 - P(A) \qquad \square$$



Another Important Theorem

$0 \leq \mathrm{P}(A) \leq 1$

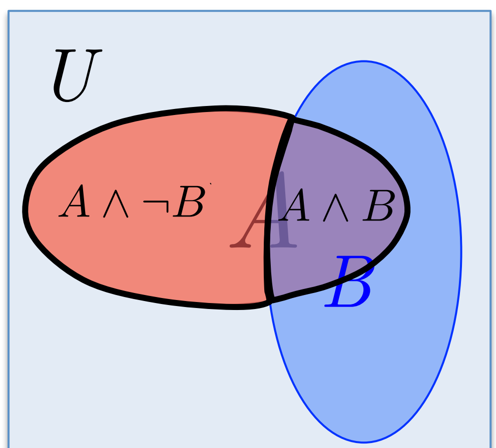
$\mathrm{P}(\mathrm{True}) = 1; \quad \mathrm{P}(\mathrm{False}) = 0$

$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$

From these we can prove:

$$P(A) = P(A \wedge B) + P(A \wedge \neg B)$$

How?



Multi-valued Random Variables

- Suppose A can take on more than 2 values
- A is a *random variable with arity k* if it can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \quad \text{if } i \neq j$$

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

$$1 = \sum_{i=1}^{k} P(A = v_i)$$

Multi-valued Random Variables

- We can also show that:

$$P(B) = P(B \wedge [A = v_1 \vee A = v_2 \vee \ldots \vee A = v_k])$$

$$P(B) = \sum_{i=1}^{k} P(B \wedge A = v_i)$$

- This is called **marginalization** over A

Prior and Joint Probabilities

- **Prior probability**: degree of belief without any other evidence
- **Joint probability**: matrix of combined probabilities of a set of variables

Russell & Norvig's Alarm Domain: (boolean RVs)

- A world has a specific instantiation of variables:

  $(\text{alarm} \wedge \text{burglary} \wedge \neg\text{earthquake})$

- The joint probability is given by:

$\mathrm{P}(\text{Alarm}, \text{Burglary})$ =

	alarm	¬alarm
burglary	0.09	0.01
¬burglary	0.1	0.8

Prior probability of burglary:

$\mathrm{P}(\text{Burglary}) = 0.1$

by marginalization over Alarm

The Joint Distribution

e.g., Boolean variables A, B, C

Recipe for making a joint distribution of d variables:

The Joint Distribution

Recipe for making a joint distribution of d variables:

1. Make a truth table listing all combinations of values of your variables (if there are d Boolean variables then the table will have 2^d rows).

e.g., Boolean variables A, B, C

A	**B**	**C**
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Recipe for making a joint distribution of d variables:

1. Make a truth table listing all combinations of values of your variables (if there are d Boolean variables then the table will have 2^d rows).

2. For each combination of values, say how probable it is.

e.g., Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

The Joint Distribution

Recipe for making a joint distribution of d variables:

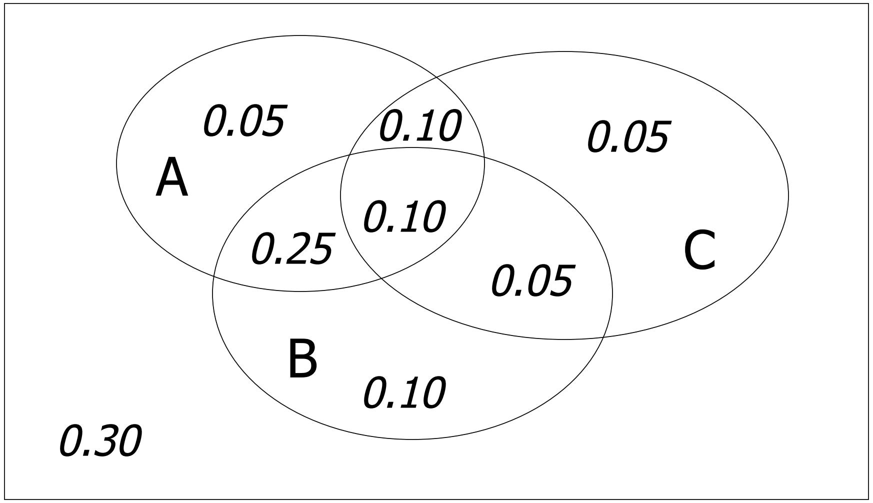
1. Make a truth table listing all combinations of values of your variables (if there are d Boolean variables then the table will have 2^d rows).

2. For each combination of values, say how probable it is.

3. If you subscribe to the axioms of probability, those numbers must sum to 1.

e.g., Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Inferring Prior Probabilities from the Joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	0.01	0.08	0.001	0.009
¬burglary	0.01	0.09	0.01	0.79

$$P(alarm) = \sum_{b,e} P(alarm \wedge \text{Burglary} = b \wedge \text{Earthquake} = e)$$

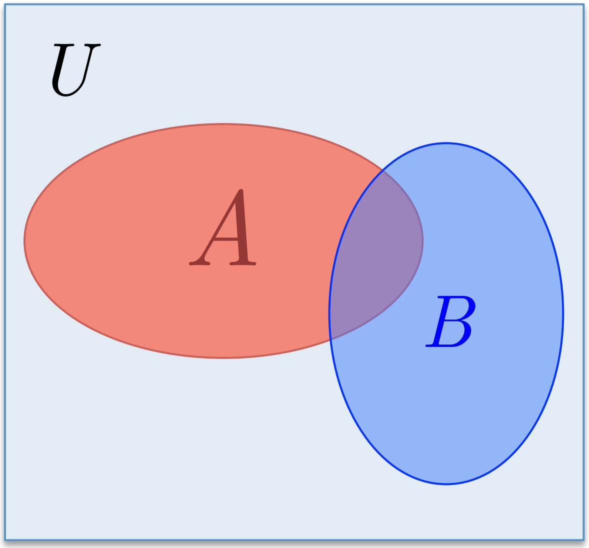
$$= 0.01 + 0.08 + 0.01 + 0.09 = 0.19$$

$$P(burglary) = \sum_{a,e} P(\text{Alarm} = a \wedge burglary \wedge \text{Earthquake} = e)$$

$$= 0.01 + 0.08 + 0.001 + 0.009 = 0.1$$

Conditional Probability

- $\mathrm{P}(A \mid B)$ = Fraction of worlds in which B is true that also have A true



What if we already know that B is true?

That knowledge changes the probability of A

- Because we know we're in a world where B is true

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

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

Example: Conditional Probabilities

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

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

P(Alarm, Burglary) =

	alarm	¬alarm
burglary	0.09	0.01
¬burglary	0.1	0.8

$$\begin{aligned} \text{P(burglary} \mid \text{alarm)} &= \text{P(burglary} \wedge \text{alarm)} / \text{P(alarm)} \\ &= 0.09 / 0.19 = 0.47 \end{aligned}$$

$$\begin{aligned} \text{P(alarm} \mid \text{burglary)} &= \text{P(burglary} \wedge \text{alarm)} / \text{P(burglary)} \\ &= 0.09 / 0.1 = 0.9 \end{aligned}$$

$$\begin{aligned} \text{P(burglary} \wedge \text{alarm)} &= \text{P(burglary} \mid \text{alarm) P(alarm)} \\ &= 0.47 * 0.19 = 0.09 \end{aligned}$$

Example: Inference from the Joint Without Explicitly Computing Priors

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	0.01	0.08	0.001	0.009
¬burglary	0.01	0.09	0.01	0.79

$$\begin{aligned}P(\text{Burglary} \mid \text{alarm}) &= \alpha\, P(\text{Burglary}, \text{alarm})\\ &= \alpha\, [P(\text{Burglary}, \text{alarm}, \text{earthquake}) + P(\text{Burglary}, \text{alarm}, \neg\text{earthquake})\\ &= \alpha\, [\, (0.01, 0.01) + (0.08, 0.09)\,]\\ &= \alpha\, [\, (0.09, 0.1)\,]\end{aligned}$$

Note: (d_1, d_2) represents a prob. distribution
Burglary = true   Burglary = false

Since $P(\text{burglary} \mid \text{alarm}) + P(\neg\text{burglary} \mid \text{alarm}) = 1$,
It must be that $\alpha = 1/(0.09+0.1) = 5.26$ (i.e., $P(\text{alarm}) = 1/\alpha = 0.19$)

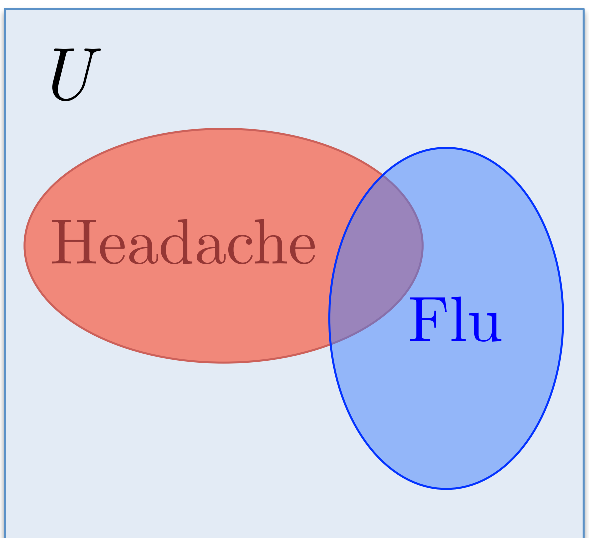
$$P(\text{burglary} \mid \text{alarm}) = 0.09 * 5.26 = 0.474$$

$$P(\neg\text{burglary} \mid \text{alarm}) = 0.1 * 5.26 = 0.526$$

Example: Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

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



P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

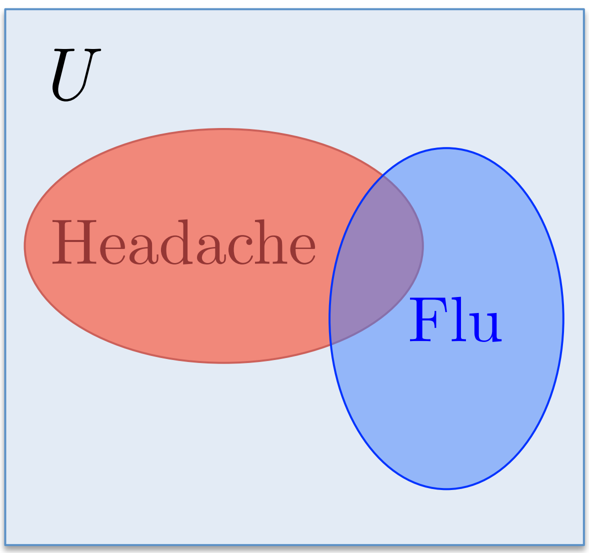
"Headaches are rare and flu is rarer, but if you're coming down with the flu there's a 50-50 chance you'll have a headache."

Based on slide by Andrew Moore

Example: Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

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



P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu."

Is this reasoning good?

Example: Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

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

P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

Want to solve for:
P(headache ∧ flu) = ?
P(flu | headache) = ?

P(headache ∧ flu) = P(headache | flu) x P(flu)
= 1/2 x 1/40 = 0.0125

P(flu | headache) = P(headache ∧ flu) / P(headache)
= 0.0125 / 0.1 = 0.125

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

- Exactly the process we just used
- The most important formula in probabilistic machine learning

(Super Easy) Derivation:

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

these are the same

Just set equal...

$$P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$$

and solve...

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Bayes' Rule

- Allows us to reason from **evidence** to **hypotheses**
- Another way of thinking about Bayes' rule:

$$P(\text{hypothesis} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{evidence})}$$

In the flu example:

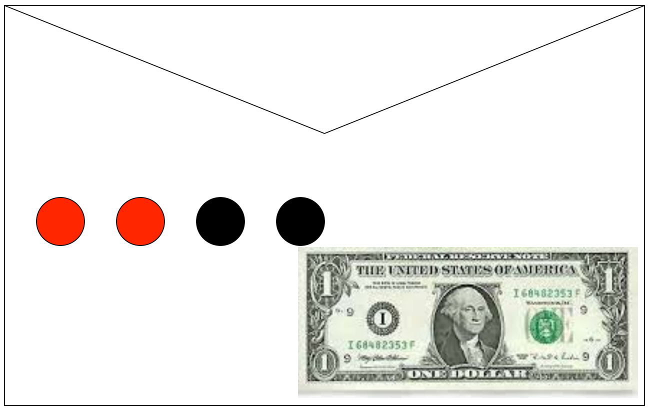
$\mathrm{P}(\text{headache}) = 1/10$ $\quad$ $\mathrm{P}(\text{flu}) = 1/40$

$\mathrm{P}(\text{headache} \mid \text{flu}) = 1/2$

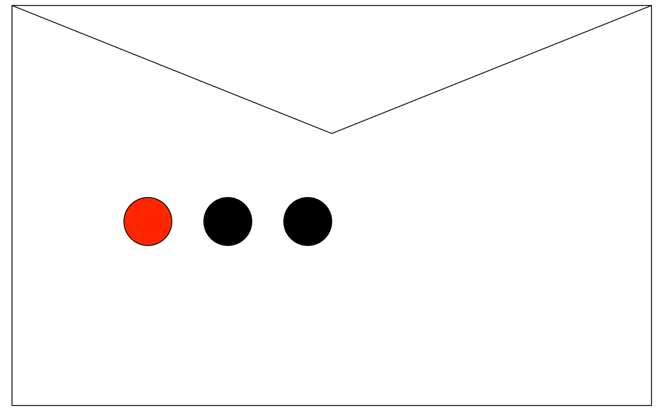
Given evidence of headache, what is $\mathrm{P}(\text{flu} \mid \text{headache})$?

Solve via Bayes rule!

Using Bayes Rule to Gamble



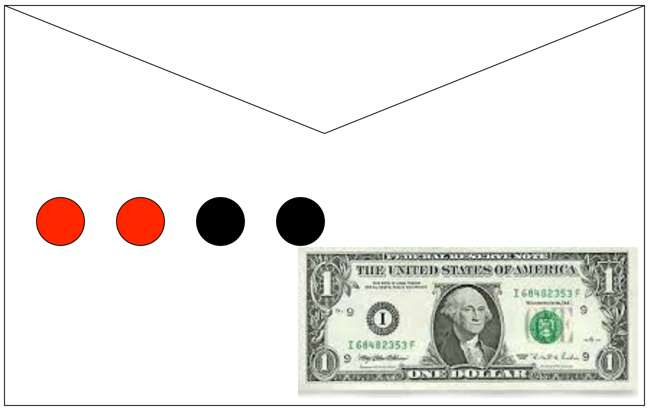
The "Win" envelope has a dollar and four beads in it



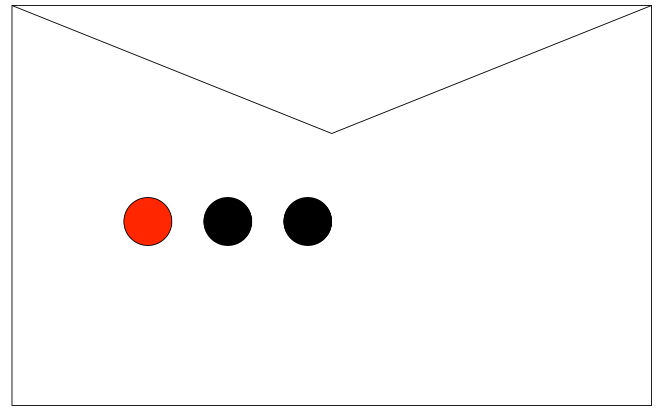
The "Lose" envelope has three beads and no money

Trivial question: Someone draws an envelope at random and offers to sell it to you.
How much should you pay?

Using Bayes Rule to Gamble



The "Win" envelope has a dollar and four beads in it



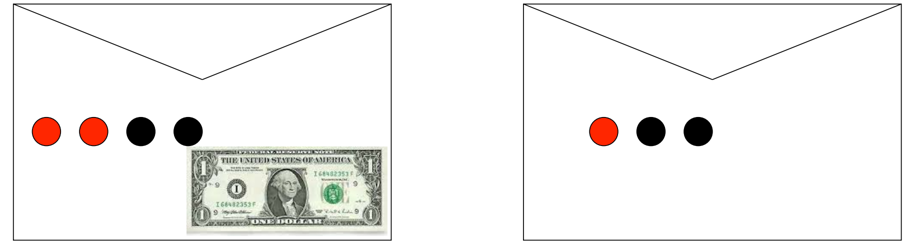
The "Lose" envelope has three beads and no money

Interesting question: Before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

Suppose it's red: How much should you pay?

Calculation…



Suppose it's black: How much should you pay?

$$P(b \mid win) = 1/2 \qquad P(b \mid lose) = 2/3$$

$$P(win) = 1/2$$

$$P(win \mid b) = \alpha \, P(b \mid win) \, P(win)$$
$$= \alpha \, 1/2 \times 1/2 = 0.25\alpha$$

$$P(lose \mid b) = \alpha \, P(b \mid lose) \, P(lose)$$
$$= \alpha \, 2/3 \times 1/2 = 0.3333\alpha$$

$$1 = P(win \mid b) + P(lose \mid b) = 0.25\alpha + 0.3333\alpha \Rightarrow \alpha = 1.714$$

$$P(win \mid b) = 0.4286 \qquad P(lose \mid b) = 0.5714$$

Based on example by Andrew Moore

Independence

- When two sets of propositions do not affect each others' probabilities, we call them **independent**
- Formal definition:

$$
\begin{aligned}
A \perp\!\!\!\perp B \quad & \leftrightarrow \quad P(A \wedge B) = P(A) \times P(B) \\
& \leftrightarrow \quad P(A \mid B) = P(A)
\end{aligned}
$$

For example, {moon-phase, light-level} might be independent of {burglary, alarm, earthquake}

- Then again, maybe not: Burglars might be more likely to burglarize houses when there's a new moon (and hence little light)
- But if we know the light level, the moon phase doesn't affect whether we are burglarized

Exercise: Independence

P(smart ∧ study ∧ prep)	smart		¬smart	
	study	**¬study**	**study**	**¬study**
prepared	0.432	0.16	0.084	0.008
¬prepared	0.048	0.16	0.036	0.072

Is *smart* independent of *study*?

Is *prepared* independent of *study*?

Exercise: Independence

P(smart ∧ study ∧ prep)	**smart**		**¬smart**	
	study	**¬study**	**study**	**¬study**
prepared	0.432	0.16	0.084	0.008
¬prepared	0.048	0.16	0.036	0.072

Is *smart* independent of *study*?

$$\mathrm{P}(\text{study} \wedge \text{smart}) = 0.432 + 0.048 = \boxed{0.48}$$

$$\mathrm{P}(\text{study}) = 0.432 + 0.048 + 0.084 + 0.036 = 0.6$$

$$\mathrm{P}(\text{smart}) = 0.432 + 0.048 + 0.16 + 0.16 = 0.8$$

$$\mathrm{P}(\text{study}) \times \mathrm{P}(\text{smart}) = 0.6 \times 0.8 = \boxed{0.48}$$

So yes!

Is *prepared* independent of *study*?

Conditional Independence

- Absolute independence of A and B:

$$A \perp\!\!\!\perp B \quad \leftrightarrow \quad P(A \wedge B) = P(A) \times P(B)$$
$$\leftrightarrow \quad P(A \mid B) = P(A)$$

Conditional independence of A and B given C

$$A \perp\!\!\!\perp B \mid C \quad \leftrightarrow \quad P(A \wedge B \mid C) = P(A \mid C) \times P(B \mid C)$$

  - e.g., Moon-Phase and Burglary are ***conditionally independent given*** Light-Level

- This lets us decompose the joint distribution:

$$P(A \wedge B \wedge C) = P(A \mid C) \times P(B \mid C) \times P(C)$$

  – Conditional independence is weaker than absolute independence, but still useful in decomposing the full joint

Take Home Exercise: Conditional independence

P(smart ∧ study ∧ prep)	smart		¬smart	
	study	**¬study**	**study**	**¬study**
prepared	0.432	0.16	0.084	0.008
¬prepared	0.048	0.16	0.036	0.072

Is *smart* conditionally independent of *prepared*, given *study*?

Is *study* conditionally independent of *prepared*, given *smart*?