Stochastic methods in AI

Geraint A. Wiggins
Professor of Computational Creativity
Department of Computer Science
Vrije Universiteit Brussel

Probability, Likelihood and Truth

Queen Mary
University of London

- In PC and FOPC, we were able to
  - write down things that were true
  - write down rules of inference that were known to hold
  - make explicit truths which were previously implicit
    - by applying rules of inference
- We showed how a proof could be constructed, to show that a given truth (a goal, or conclusion) followed from a set of facts
- In all of this, however, everything is true or false
  - there's no way of saying "perhaps" or "maybe"
- We can refine all this by viewing truth values as extreme probabilities
  - 0 replaces false, 1 replaces true, but now we have all the values in between
  - (But note that there are limits to this analogy - it's not completely strict)

Probability, Likelihood and Truth

- Conjunction is exactly like multiplication
  - $F \wedge F = 0 \times 0 = 0 = F$
  - $F \wedge T = 0 \times 1 = 0 = F$
  - $T \wedge T = 1 \times 1 = 1 = T$
- Disjunction is somewhat like addition, but with a maximum value of 1
  - $F \vee F = \min(0 + 0, 1) = 0 = F$
  - $F \vee T = \min(0 + 1, 1) = 1 = T$
  - $T \vee T = \min(1 + 1, 1) = 1 = T$
  - however, NB, this is not a complete, proper definition – we'll give one later
- This idea corresponds precisely with the idea of probability
  - Something which is certain (T) has probability 1
  - Something which is certainly not (F) has probability 0

Using probability for logic

- We write P(A) to mean "the probability of A"
    - NB, ***not*** Predicate P applied to Object A!
  - But what does P(A) mean?
    - P(I will draw ace of hearts)
    - P(the coin will come up heads)
    - P(it will snow tomorrow)
    - P(the sun will rise tomorrow)
    - P(the problem is in the third cylinder)
    - P(the patient has measles)

Frequentist interpretation

Queen Mary University of London

- We can interpret frequencies of events as their probability
  - Draw a card from a normal deck
    - h = 13 hearts, s = 13 spades, d = 13 diamonds, c = 13 clubs
    - Total number of cards, n = 13 x 4 = 52
  - The probability that the proposition A = "the card is a heart" is true corresponds to the relative frequency with which we expect to draw a heart
    - note that, here, each card is equally likely to be drawn
  - P(A) = h / n = 13 / 52 = 0.25
- Definition: the *probability* of an event A is the number of possible occurrences where A holds divided by the all the possible occurrences
  - thus, if A always holds, P(A) = n / n = 1 = T in PC/FOPC, conversely for F

Simple exercises

Queen Mary
University of London

- Compute
  - P (I will draw ace of hearts)
  - P (I will draw a spade)
  - P (I will draw a heart or a spade)
  - P (I will draw a heart and a spade)

Definitions

Queen Mary University of London

- An *elementary* or *atomic event* is an occurrence that cannot be made up of other events
  - cf. Proposition/Sentence in PC, Literal/Atomic Sentence in FOPC
- An *event* is a set of atomic events
- The set of all possible outcomes of an event E is the *sample space* or *universe* for that event
- The probability of an event E in a sample space S is the ratio of the number of elements in E to the total number of possible outcomes of the sample space S of E
  - P(E) = |E| / |S|

Subjective interpretations

Queen Mary University of London

- There are many situations in which there is no objective frequency interpretation
  - Just before hang-gliding from the top of Ben Nevis, I say "there is probability 0.2 that I am going to have a broken neck"
  - You are working hard on your AI practical and you believe that the probability that you will get 90% is 0.9
- The probability that proposition A is true corresponds to the degree of *subjective belief*
  - It's better to use probabilities based on evidence!
  - but subjective belief can get us a very long way, sometimes

Axioms of probability

Queen Mary University of London

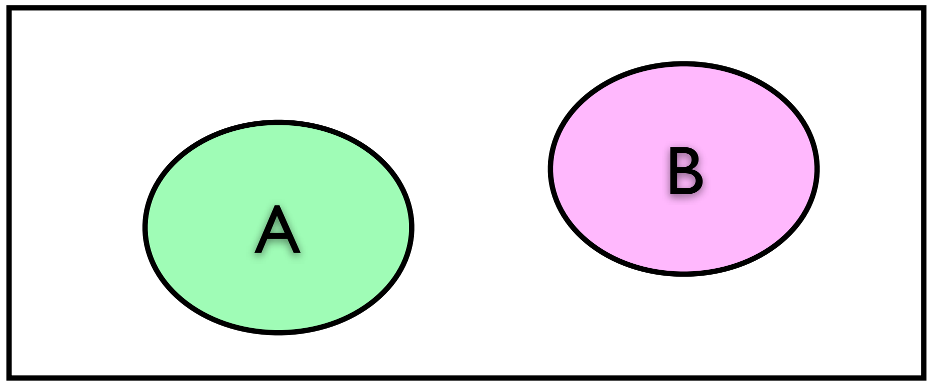
- There is a debate about which interpretation to adopt
    - frequentist
    - probabilistic
  - both are useful
  - but there is agreement about the underlying maths
- Three basic requirements of probability values:
  - $0 \leq P(A) \leq 1$
  - $P(A \vee B) = P(A) + P(B)$ iff A and B are mutually exclusive
    - that means "iff A and B cannot be true at the same time"
  - $P(True) = 1$

Three axioms is enough

Queen Mary University of London

- These three axioms are all we need to define probabilities, as all other rules can be derived from them
- For example we can show that
  - $P(\neg A) = 1 - P(A)$ because
    - $P(A \vee \neg A) = P(True)$ by law of excluded middle
    - $P(A \vee \neg A) = P(A) + P(\neg A)$ by axiom 2
    - $P(True) = 1$ by axiom 3
    - $P(A) + P(\neg A) = 1$ transitivity of =
  - P(false) = 0 because
    - $False = \neg True$ by definition
    - $P(False) = 1 - P(True)$ by the derived rule above

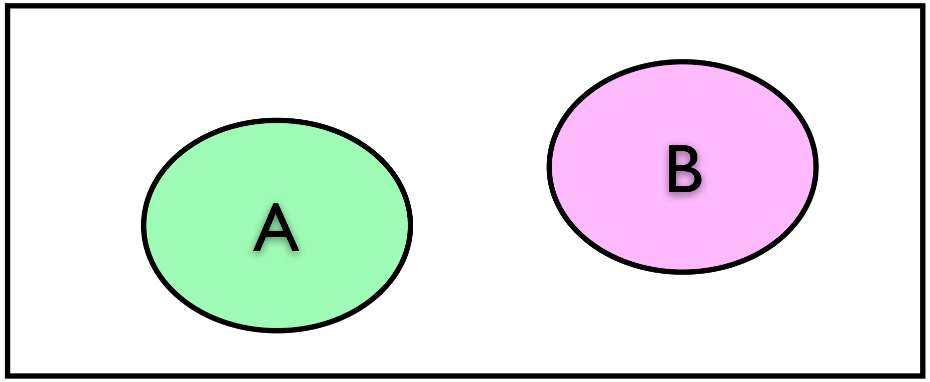
Another way to look at probability



- The entire rectangle is the Universe
  - it contains every possible outcome of every atomic event
- A and B are events
  - the sets in this Venn diagram contain the conditions when they are true
- Here, they are mutually exclusive
  - they do not overlap, so they cannot both occur at the same time
- Probability is represented by the area relative to the universal (which has probability 1)

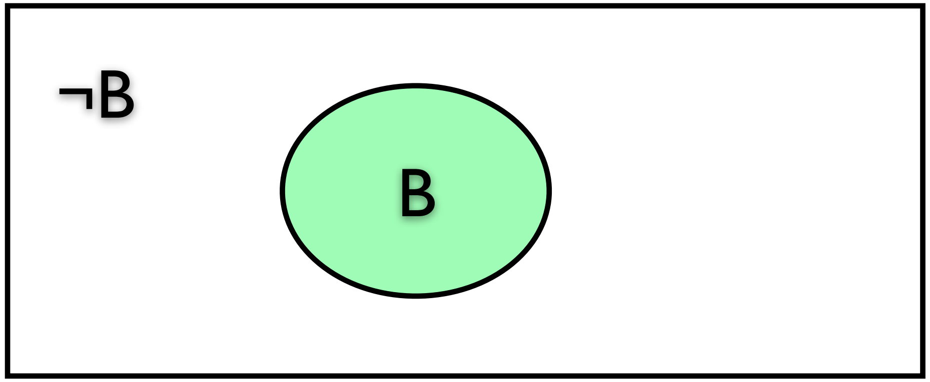
Another way to look at probability

Queen Mary, University of London



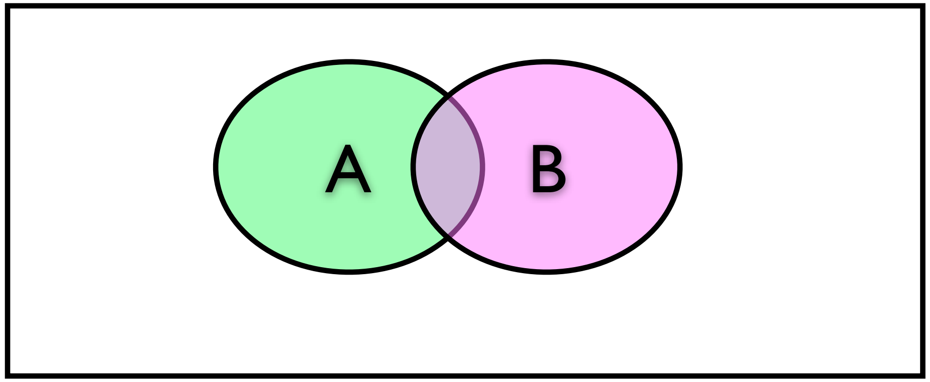
- The axioms:
  - $0 \leq P(A) \leq 1$
    - a set cannot have a negative area; it cannot be bigger than the Universal Set
    - nothing can have less than no chance of happening
    - nothing can have more than absolute certainty of happening
  - $P(A \vee B) = P(A) + P(B)$ iff A and B are mutually exclusive
    - compute the respective probabilities by measuring areas and adding
  - P(True) = 1
    - the Universal Set is the set of all things that are true
    - the area of the Universal Set is that of the Universal Set; $n / n = 1$

Another way to look at probability



- P(¬B) = 1 – P(B)
  - B is the green part
  - ¬B is the rest
  - P(B) + P(¬B) = P(U) = 1

Another way to look at probability



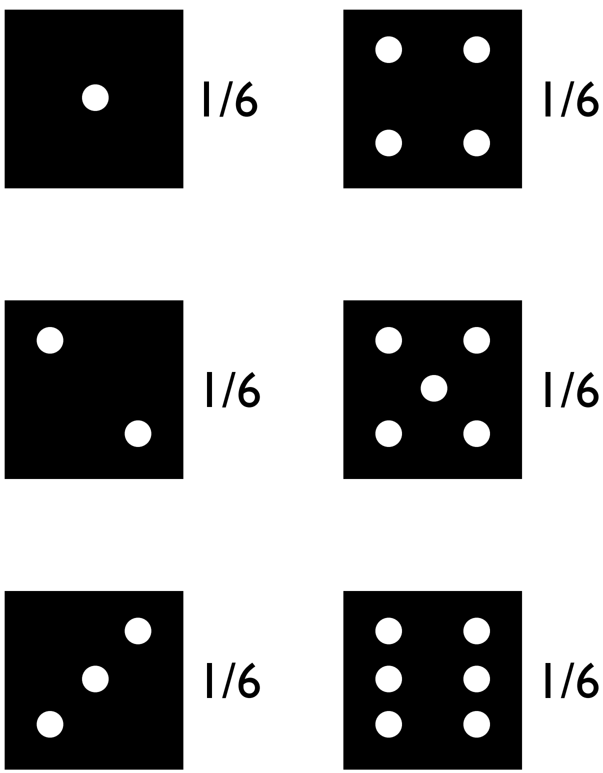
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  - because the intersection area would be included twice, otherwise
- Note, here, the similarity between the connectives and set operators
  - $\wedge$ corresponds with $\cap$
  - $\vee$ corresponds with $\cup$
  - this is not a coincidence

Random variables

- Atomic propositions in PC were assigned from { T, F } by an interpretation
- Here we think of atomic events as being associated with *random variables* ranging over a *domain* of values which are
  - mutually exclusive
    - so a variable can only have one value at a time, or only one possibility can be true at a time
  - exhaustive
    - all possible values are known
  - Each possible outcome has an associated probability
- Examples
    - coin toss: { heads, tails }
    - die roll: { 1, 2, 3, 4, 5, 6 }
    - has measles: { true, false }

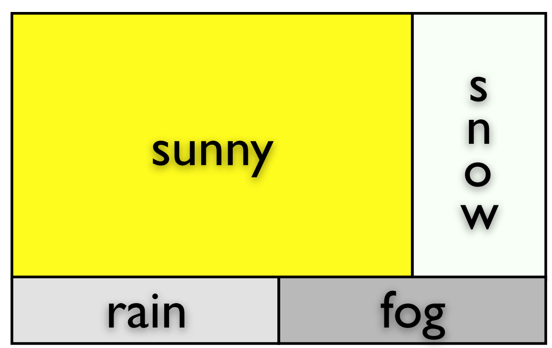
Probability distribution of random variable

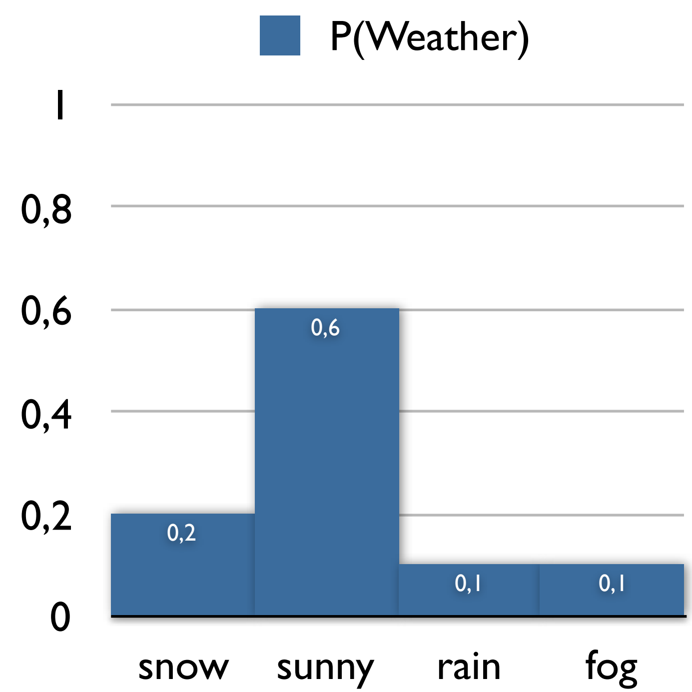
- A probability distribution of a random variable is the set of the probabilities covering every possible value the variable might take



Uniform distribution

Weather	P(Weather)
snow	0,2
sunny	0,6
rain	0,1
fog	0,1





Non-uniform distribution

Joint probability distributions

- A *joint probability distribution* for some random variables is an enumeration of the probabilities for all possible combinations of the *joint outcomes* of the random variables
    - compute the joint outcomes by taking the cross product of the variables' domains
- NB. These probabilities are about co-occurrence, not about cause!!

Repairs and traffic on a bridge		
Road repair	Bad traffic	Probability
T	T	0,3
T	F	0,2
F	T	0,1
F	F	0,4

Age and weight of humans		
Age	Weight	Probability
Young	Light	0,5
Young	Heavy	0,1
Old	Light	0,1
Old	Heavy	0,3

Joint probability distributions

- Alternative representation as multi-dimensional table

	Road repairs	¬ Road repair
Bad traffic	0,3	0,1
¬ Bad traffic	0,2	0,4

Age \ Weight	Light	Heavy
Young	0,5	0,1
Old	0,1	0,3

Repairs and traffic on a bridge		
Road repair	Bad traffic	Probability
T	T	0,3
T	F	0,2
F	T	0,1
F	F	0,4

Age and weight of humans		
Age	Weight	Probability
Young	Light	0,5
Young	Heavy	0,1
Old	Light	0,1
Old	Heavy	0,3

Joint probability distributions

- Notation
  - All these statements are equivalent
    - P(Cavity = true, Insulated = false) = 0.08
    - P(Cavity, ¬Insulated) = 0.08
    - P(Cavity ∧ ¬Insulated) = 0.08
  - so we can use , (comma) as “and”, and mix’n’match PC propositions with variables
    - you will see these and other notations used
    - the important thing is to be consistent in your own usage
- Joint probabilities are useful because they allow us to study the interaction between the component events

Marginal probabilities

Queen Mary University of London

	Road repairs	¬ Road repair	Marginal
Bad traffic	0,3	0,1	0,4
¬ Bad traffic	0,2	0,4	0,6
Marginal	0,5	0,5	1

- P(Bad Traffic) = P(Bad Traffic ∧ Road repairs) + P (Bad Traffic ∧ ¬ Road repairs)
  = 0.3 + 0.1
  = 0.4

- NB: the table must be consistent with the axioms of probability
  - 0 ≤ each value in the table ≤ 1
  - the values in the whole table should add up to 1
  - the different outcomes must be mutually exclusive

Probabilistic inference

	Road repairs	¬ Road repair	Marginal
Bad traffic	0,3	0,1	0,4
¬ Bad traffic	0,2	0,4	0,6
Marginal	0,5	0,5	1

- Now we can compute more subtle inferences
  - P(Bad traffic ∨ Repairs) = 0.3 + 0.1 + 0.2 = 0.6
  - P(Repairs → Bad traffic) = P(¬ Repairs ∨ Bad Traffic)
    = 0.1 + 0.4 + 0.3
    = 0.8

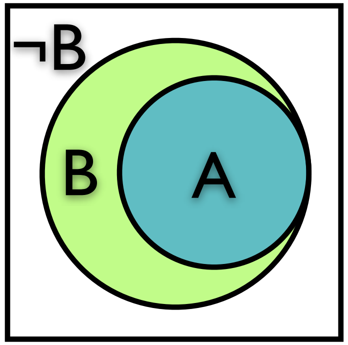
Prior probability

- The *prior probability*, or *unconditional probability*, of an event is the probability assigned to an event in the absence of knowledge supporting its occurrence and absence, that is, the probability of the event prior to any evidence.
- Example
  - P(Toothache = true) = 0.1
  - means that the *unconditional* or prior probability that a patient has a cavity is 0.1
    - *In the absence of other information*, there is 10% chance someone has toothache
- This is the kind of probability we have been dealing with so far, where we view the world as unchanging
  - but what happens if it can change?

Conditional or posterior probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \quad \text{where } P(B) > 0$$

- P(A | B) denotes the *conditional* (or *posterior*) *probability* of A, given that we know B and B is all we know
  - If we receive evidence concerning a proposition, prior probabilities are no longer applicable
  - We need to assess the conditional probability of that proposition given the available evidence
  - Note that for this to be different from P(A), A and B are no longer independent



- Here, P(A) = 0.25, P(B) = 0.5
- A is contained in B, so $P(A \wedge B) = 0.25$
- Therefore, P(A | B) = 0.25 / 0.5 = 0.5

Something to try

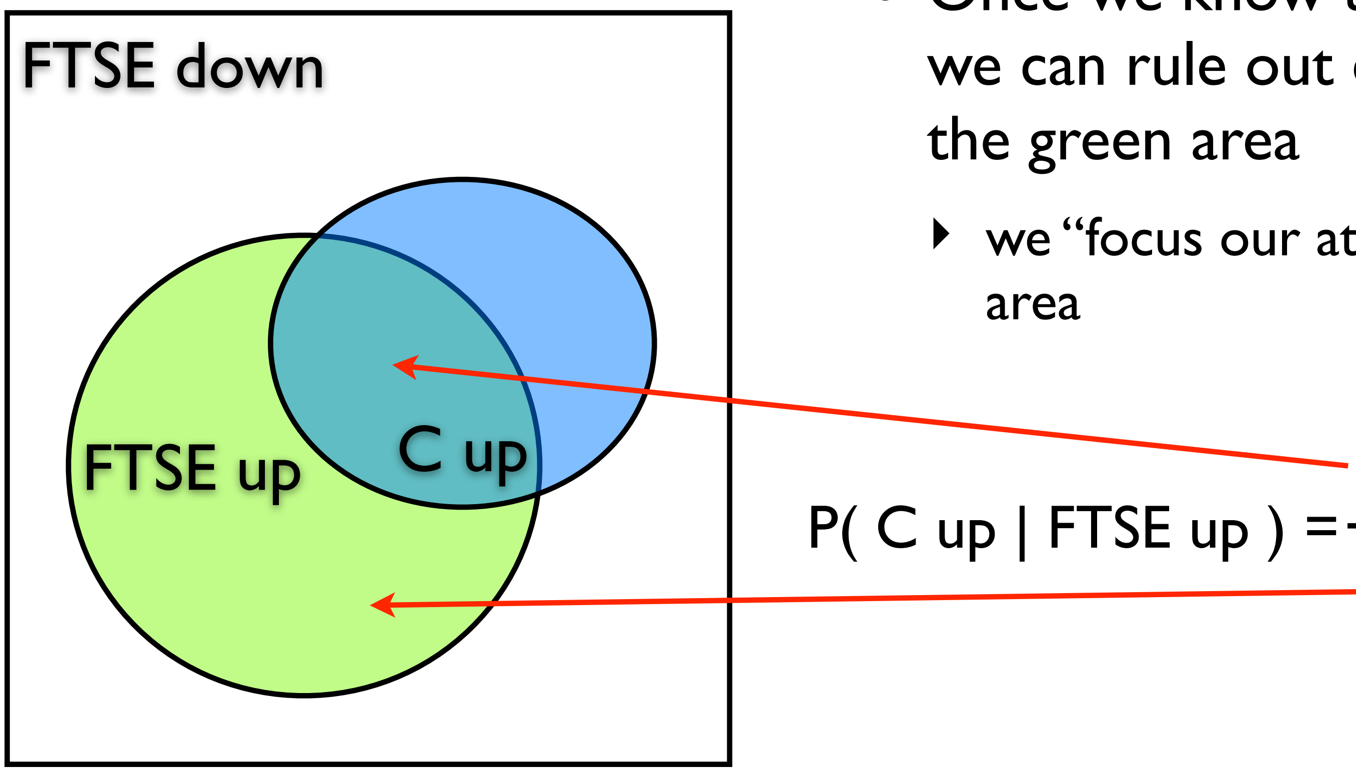
$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \quad \text{where } P(B) > 0$$

- What are the values for the following?
  - P(Heads | Heads)
  - P(Ace of hearts | Ace of spades)
  - P(Bad traffic | Road repairs)
  - P(Road repairs | Bad traffic)

	Road repairs	¬ Road repair	Marginal
Bad traffic	0,3	0,1	0,4
¬ Bad traffic	0,2	0,4	0,6
Marginal	0,5	0,5	1

Another example

- Suppose we are interested in the probability that the stock price of company C will increase, given that we know about the FTSE index from the BBC news

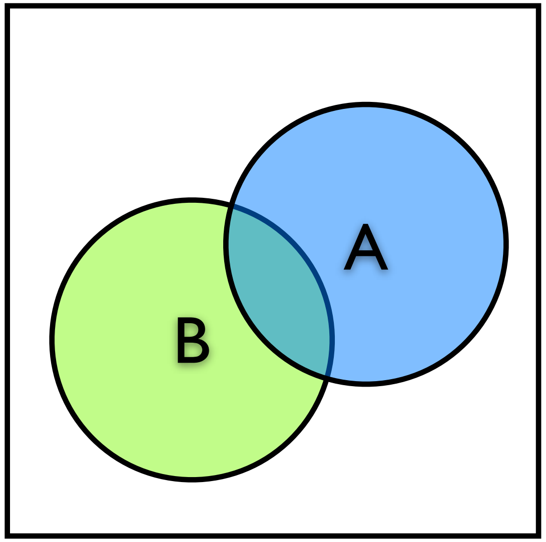


- Once we know that the FTSE is up, we can rule out everything outside the green area
    - we "focus our attention" on the FTSE area

$$P(\text{ C up } \mid \text{ FTSE up }) = \frac{P(\text{ C up } \wedge \text{ FTSE up })}{P(\text{ FTSE up })}$$

Total probability of an event

- A convenient way to calculate P(A) is with the following formula
  - produced using the conditional probability rule
  - $P(A) = P(A \wedge B) + P(A \wedge \neg B)$
    $= P(A \mid B)P(B) + P(A \mid \neg B)P(\neg B)$



- Event A is composed of those occasions when
  - A and B co-occur
  - A and ¬B co-occur
  - Because compound events "A and B" and "A and ¬B" are mutually exclusive, the probability of A must be the sum of these two probabilities

Another example

Queen Mary
University of London

- P(Cavity | Toothache) = 0.8 is a conditional probability
- Maybe we have other evidence
  - P(Cavity | Toothache, Earthquake) = P(Cavity | Toothache)
    - because the Earthquake information is not relevant
- P(Cavity | Toothache, Cavity) = 1
    - because the non-conditional probability overrides

Rules of conditional probability

- Product rule
  - $P(A \wedge B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$
    - An alternative definition of conditional probability
- Chain rule (successive application of product rule)
  - $P(X_n \mid X_1 \wedge \ldots \wedge X_{n-1}) = P(X_1) \prod^{n}_{i=2} P(X_i \mid X_1 \wedge \ldots \wedge X_{i-1})$
  - Proof
    - $P(X_1 \wedge \ldots \wedge X_n) = P(X_1 \wedge \ldots \wedge X_{n-1})\, P(X_n \mid X_1 \wedge \ldots \wedge X_{n-1})$

      $= P(X_1 \wedge \ldots \wedge X_{n-2})\, P(X_{n-1} \mid X_1 \wedge \ldots \wedge X_{n-2})\, P(X_n \mid X_1 \wedge \ldots \wedge X_{n-1})$

      etc.
  - Example
    - $P(X_1 \wedge X_2 \wedge X_3) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1 \wedge X_2)$
- NB. $P(A \rightarrow B)$ is not usually the same as $P(B \mid A)$

Rules of conditional probability

- Conditioning Rule
  - If we know an exhaustive set of conditional probabilities for an event, we can compute its prior probability
    - like the marginal probability calculation from the table earlier
    - $P(Y) = \sum_z P(Y \mid Z)\, P(Z)$
    - This sums all the outcomes of all the variables that *condition* Y
- Conditionalised Product Rule
  - $P(A \wedge B \mid E) = P(A \mid B \wedge E)\, P(B \mid E) = P(B \mid A \wedge E)\, P(A \mid E)$
  - Proof

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

Bayes' Rule

- Derivation

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

  - Product rule
    - $P(A \wedge B) = P(A \mid B)\, P(B)$
  - Commutativity of $\wedge$
    - $P(A \wedge B) = P(B \mid A)\, P(A)$
  - Transitivity of =
    - $P(B \mid A)\, P(A) = P(A \mid B)\, P(B)$
  - Divide both sides by $P(A)$ to get the rule
- Bayes' rule
  - allows us to reverse a conditional probability, if we know the probabilities of its components
    - P(Cause | Effect) = P(Effect | Cause)P(Cause) / P(Effect)

Probablisitc vs. Subjective View of Probability

- There is nothing about Bayes' Rule that tells you how to interpret the probabilities
  - Nevertheless, Bayes' Rule is often used with semi-subjective probabilities
- Subjective Probabilities (degree of belief)
  - Statistical estimates (frequentist interpretation) for probabilities are fine if the experiment under consideration can be repeated a number of times under similar circumstances. But suppose you want to assign the probability that one political party will beat another. You can determine your own personal probability by seeing what kind of bet you'd be willing to make.
  - A doctor might assign a probability of 0.8 to the statement that a particular patient has a meningitis, given that she (the patient!) has a stiff neck. We may not have statistical data on that one patient for meningitis and stiff necks; however, we might be basing our subjective assignment of probability on our experiences with patients with a similar medical history.

Reasoning with Probabilities

Queen Mary University of London

- Scenario
  - Jane goes to the doctor's for a routine checkup and takes some tests
  - One test for a rare genetic disease comes back positive
  - The disease is potentially fatal
- Jane looks up wikipedia and learns that
  - "rare" means $P(\text{Disease}) = P(D) = 1/10{,}000 = 0.0001$
  - the test is 99% accurate:
    - a very small amount of false positives $P(\text{Test} = + \mid \neg D) = 0.01$
    - no false negatives $P(\text{Test} = - \mid D) = 0$
- Jane would like to compute the probability that she has the disease and take appropriate action

Formulating the situation

- The scenario in probabilities
  - The disease is rare: $P(D) = 0.0001$
  - Reliability of test: $P(\text{Test} = - \mid \neg D) = 0.99$
  - Chance of false positive: $P(\text{Test} = + \mid \neg D) = 0.01$
  - No chance of false negative: $P(\text{Test} = - \mid D) = 0$
  - $\therefore$ Chance of true positive: $P(\text{Test} = + \mid D) = 1$

$$
\begin{aligned}
& P(D \mid \text{Test} = +) \\
= \; & P(D \wedge \text{Test} = +) \,/\, P(\text{Test} = +) \\
= \; & 1 \,/\, (1 + 99.99) \\
= \; & 0.0099 \qquad \text{(not 0.99!!!)}
\end{aligned}
$$

Take 10,000 random people

1 has the disease — 9,999 do not have the disease

1 tests positive — 99.99 test positive — 9,899.01 test negative

Formalising the reasoning

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

- Derivation
  - Bayes' rule
    - $P(D \mid Test = +) = P(Test = + \mid D)\,P(D) / P(Test = +)$
      $= 1 \times 0.0001 / P(Test = +) = 0.0001 / P(Test = +)$
  - Bayes' rule
    - $P(\neg D \mid Test = +) = P(Test = + \mid \neg D)\,P(\neg D) / P(Test = +)$
      $= 0.01 \times 0.9999 / P(Test = +) = 0.009999 / P(Test = +)$
  - Excluded middle and arithmetic
    - $P(D \mid Test = +) + P(\neg D \mid Test = +) = 0.0001 / P(Test = +) + 0.009999 / P(Test = +)$
      $= (0.0001 + 0.009999) / P(Test = +) = 1 \therefore 0.0001 + 0.009999 = 1 \times P(Test = +) = 0.010099$
  - Bayes' Rule
    - $P(D \mid Test = +) = 1 \times 0.0001 / 0.010099 = 0.0099$

Another similar example

- Let S be the proposition that a patient has a stiff neck, and M the proposition that the patient has meningitis
- Suppose a doctor wants to know P(M | S)
  - Probabilities from a medical database
    - P(S | M) = 0.5
    - P(M) = 1 / 50,000 = 0.00002
    - P(S) = 1 / 20 = 0.05
  - Using Bayes' Rule
    - P(M | S) = P(S | M) P(M) / P(S) = 0.5 x 0.00002 / 0.05 = 0.0002
- If the doctor knows this probability, and sees a patient with a stiff neck, he/she can know how strongly/weakly to lean toward a diagnosis of meningitis

How practical is Bayes' rule?

- How reasonable is the supposition that you know P(S | M) but need to calculate P(M | S)?
  - Why not estimate P(M | S) directly by sampling?

This is affected by both the presence/absence of an epidemic and the way meningitis works – complicated.

Reflects the way meningitis works. Unaffected by epidemic.

$$P(M \mid S) = \frac{P(S \mid M)\, P(M)}{P(S)}$$

Unrelated information

Increases when there is an epidemic.

You can't separate these two unrelated factors when estimating P(M | S) by sampling

How practical is Bayes' rule?

- What if the prior probability P(S) in the denominator is too difficult to to determine?
  - Can we avoid direct assessment of the probability of evidence?
    - That is, the denominator of Bayes' rule
  - Yes, with *normalization*
    - (As in example above)

Normalization

Queen Mary University of London

- We can avoid direct assessment of priors of evidence, such as P(S) here, by considering an exhaustive set of hypotheses

$$P(M \mid S) = \frac{P(S \mid M)\,P(M)}{P(S)}$$ by Bayes' rule

$$P(\neg M \mid S) = \frac{P(S \mid \neg M)\,P(\neg M)}{P(S)}$$ by Bayes' rule

$$P(M \mid S) + P(\neg M \mid S) = 1$$ by excluded middle

$$P(S \mid M)\,P(M) + P(S \mid \neg M)\,P(\neg M) = P(S)$$ substitute RHSs and multiply by P(S)

$$P(M \mid S) = \frac{P(S \mid M)\,P(M)}{P(S \mid M)\,P(M) + P(S \mid \neg M)\,P(\neg M)}$$ substitute for P(S) in original rule

By assessing P(S | ¬M), we avoid assessing P(S)

Normalization

Queen Mary University of London

- We can generalise normalisation beyond the $P \vee \neg P$ case
  - use a distribution over multiple, mutually exclusive values of a variable
- Suppose we wish to compute a posterior distribution over A given B = b, and suppose A has disjoint domain $a_1, \ldots, a_m$
- Apply Bayes' rule for each value of A
  - $P(A = a_1 \mid B = b) = P(B = b \mid A = a_1) P(A = a_1) / P(B = b)$
    ...
    $P(A = a_m \mid B = b) = P(B = b \mid A = a_m) P(A = a_m) / P(B = b)$
  - Adding these up, and noting that $\sum_i P(A = a_i \mid B = b) = 1$
  - $1 / P(B = b) = 1 / \sum_i P (B = b \mid A = a_i) P(A = a_i)$
  - This quantity is known as the normalisation factor, denoted by α

How practical is Bayes' rule?

Queen Mary University of London

- Normalisation is one way to make Bayes' rule more practically applicable, but there's another
  - The issue is that Bayes' rule requires us to estimate too many probabilities before we can apply it
  - We want to estimate one conditional probability; we need to know
    - another conditional probability
    - two prior probabilities
  - Why are we willing to make this effort?
    - Because the other probabilities are typically easier to obtain by observation
- Can we reduce the number of probabilities needed to apply Bayes' Rule
  - There is a particular technique for doing this when there are multiple sources of evidence and we can use the idea of *conditional independence*

Independence revisited

Queen Mary University of London

- Unconditional independence
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
  - $P(A \wedge B) = P(A) P(B)$
- Conditional independence
  - independence holds, given a common cause
  - A and B are *independent given C* iff
    - $P(A \mid B \wedge C) = P(A \mid C)$
  - it follows that
    - $P(A \wedge B \mid C) = P(A \mid C) P(B \mid C)$
    - try this derivation, using the conditional probability rule, yourself

Causality and conditional independence

- A Bayesian Belief Network (BBN) is a graph for which the following properties hold:
  - A set of random variables makes up the nodes of the network. Variables may be discrete or continuous. Each node is annotated with quantitative probability information.
  - A set of directed links or arrows connects pairs of nodes. If there is an arrow from node X to node Y, X is said to be a parent of Y.
  - Each node X has a conditional probability distribution P(X | Parents(X)) that quantifies the effect of the parents on the node.
    - note there can be more than one parent
  - The graph has no directed cycles (so it is a directed, acyclic graph, or DAG)
- Belief networks are now the standard technology for expert systems
  - Domain experts generally report it is not too hard to interpret the links and fill in the requisite probabilities
  - Some (e.g., Pathfinder IV) seem to be outperforming the experts who designed them!

Summary

Queen Mary University of London

- Probability theory enables the use of varying degrees of likelihood to represent uncertainty
- A (joint) probability distribution completely describes a (set of) random variable(s)
- Conditional probabilities let us calculate probabilities relative to priors that we know
- Bayes' rule is helpful in relating conditional probabilities and priors
- Independence assumptions let us make intractable problems tractable
- Belief networks are now the standard technology for expert systems