



NUI Galway
OÉ Gaillimh

Machine Learning

Week 11: Probabilistic Machine Learning (Part 1)

Prof. Michael Madden

Chair of Computer Science

Head of Machine Learning & Data Mining Group

National University of Ireland Galway



Learning Objectives

After successfully completing this, you will be able to ...

- Discuss the motivation for handling uncertainty in ML
- Distinguish between prior and conditional probability
- Demonstrate understanding of how to use the axioms of probability and Bayes' rule
- Describe and apply the Naïve Bayes classifier to inductive learning problems
- Show how Bayesian Networks represent influence and independence of variables
- Discuss how BNs can be used for classification & data exploration.



Structure of Videos for Probabilistic ML Topic

Week 11

- Topic Part 1A: Review of Probability Basics
- Topic Part 1B: Reasoning with Bayes' Rule

Week 12

- Topic Part 2A: Probabilistic Classifiers
- Topic Part 2B: Bayesian Networks



Part 1A: Review of Probability Basics



Why Consider Uncertainty? (1)

- In a deterministic domain, is there uncertainty?
- What are the sources of uncertainty?
 - **Incomplete knowledge:** lack of relevant facts, partial observations, inaccurate measurements, incomplete domain theory ...
 - **Inability to process:** too complex to use all possible relevant data in computations, or to consider all possible exceptions and qualifications



Why Consider Uncertainty? (2)

Example: Going to airport – will t minutes be enough?

Problems:

- incomplete observations (road state, other drivers' plans, etc.)
- noisy sensors (traffic reports)
- uncertainty in action outcomes (flat tyre, etc.)
- immense complexity of modeling traffic

Therefore, purely logical approach either:

- Risks falsehood:
“90 minutes will get me there on time”, or
- Leads to conclusions that are too weak for decision making:
“90 minutes will get me there on time, if there's no accident, and it doesn't rain, and my car doesn't break down ...”
“24 hours will get me there on time” – but requires very long wait





Techniques for Handling Uncertainty

Default or Nonmonotonic logic:

- Assume car does not have flat tyre
- Assume 90 minutes is OK unless contradicted by evidence

Issues: What assumptions are reasonable?
How should contradiction be handled?

Rules with Certainty Factors:

- **90 minutes** | $\rightarrow_{0.3}$ **get there on time** (ie 30% certainty)
- **Sprinkler** | $\rightarrow_{0.99}$ **WetGrass**
- **WetGrass** | $\rightarrow_{0.7}$ **Rain**

Issues: Problems with combination: Sprinkler causes rain?

Probability:

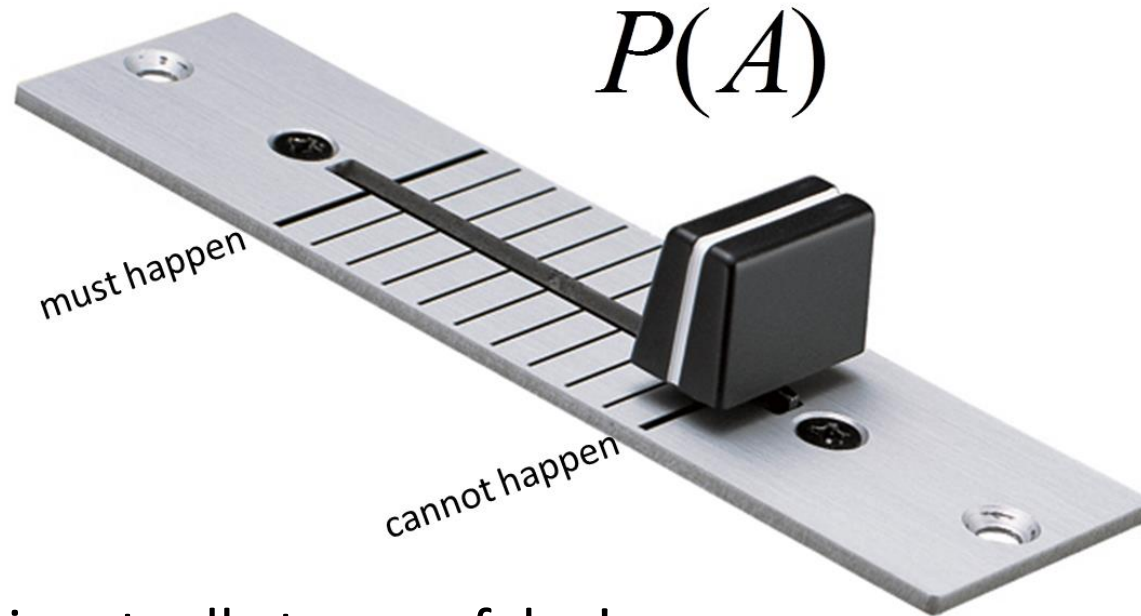
- Model agent's degree of belief
- Given the available evidence,
90 minutes will get me there on time **with probability 0.4**



Review of Probability (1)

Probability: a way of **summarising** uncertainties

$0 \leq P(s) \leq 1$: level of **belief** in truth of statement



Important:

- Statement itself is actually true or false!
- Our beliefs may change when we observe new evidence



Review of Probability (2)

How do we assess probabilities?

- Statistical data, general rules, logical considerations, or combination of evidence
- Although probabilities are personal, they must still be reasonable and **rational**

Example:

- Pick a card ...

Before looking: $P(\text{Card} = \text{Q} \heartsuit) = 1/52$

After looking: $P(\text{Card} = \text{Q} \heartsuit) = 0 \text{ or } 1$



Review: Probability Notation (1)

We will consider discrete random variables:

- **Random** variable: cannot directly control its value, we can just observe it
- **Boolean-valued** random variable: denotes an event, with some degree of uncertainty as to whether it occurs:

Cavity: <true, false>

$P(\text{Cavity}=\text{true})$ is also written as **$P(\text{cavity})$**

$P(\text{Cavity}=\text{false})$ is also written as **$P(\neg \text{cavity})$**

- General case of **discrete random variable**: values in domain are **mutually exclusive and exhaustive**

ExamResult: <a, b, c, d, fail>

$P(\text{ExamResult}=\text{a}) = 0.2, \dots$

$P(\text{ExamResult}) = \langle 0.2, 0.3, 0.3, 0.15, 0.05 \rangle$



Review: Probability Notation (2)

Probabilities of Values in a Domain Sum to 1

- Provided they are mutually exclusive and exhaustive:

$$P(\text{Elvis}=\text{alive}) = .01; P(\text{Elvis}=\text{dead}) = .99$$

Elementary proposition:

- Constructed by assigning a value to a random variable:
- Example 1: *Weather = sunny*
- Example 2: *Cavity = false*



Review: Probability Notation (3)

Complex proposition:

- Formed from **elementary propositions** and logical operators
- Example: *Weather = sunny* \vee *Cavity = false*

Logical operators and notations used to represent them:

AND: $a \wedge b$ (mnemonic: similar shape to an A)

OR: $a \vee b$

NOT: $\neg a$



Review: Probability Notation (4)

Atomic event: complete specification of state of the ‘world’

- World: environment/scenario about which we are reasoning
- E.g.: if world consists of just two Boolean variables, *Anna_Here* and *Bob_Here*, There are 4 distinct atomic events:

Anna_Here = false \wedge Bob_Here = false

Anna_Here = false \wedge Bob_Here = true

Anna_Here = true \wedge Bob_Here = false

Anna_Here = true \wedge Bob_Here = true

- Atomic events are **mutually exclusive and exhaustive**:
only one is true; their probabilities sum to 1



Review: Axioms of Probability (1)

1. Probability of any proposition is between 0 and 1:
 $0 \leq P(a) \leq 1$
2. Necessarily true propositions have probability 1;
necessarily false propositions have probability 0:
 $P(\text{false}) = 0, P(\text{true}) = 1$
3. Probability of a disjunction ($a \vee b$) is:
 $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$
(Sum Rule)

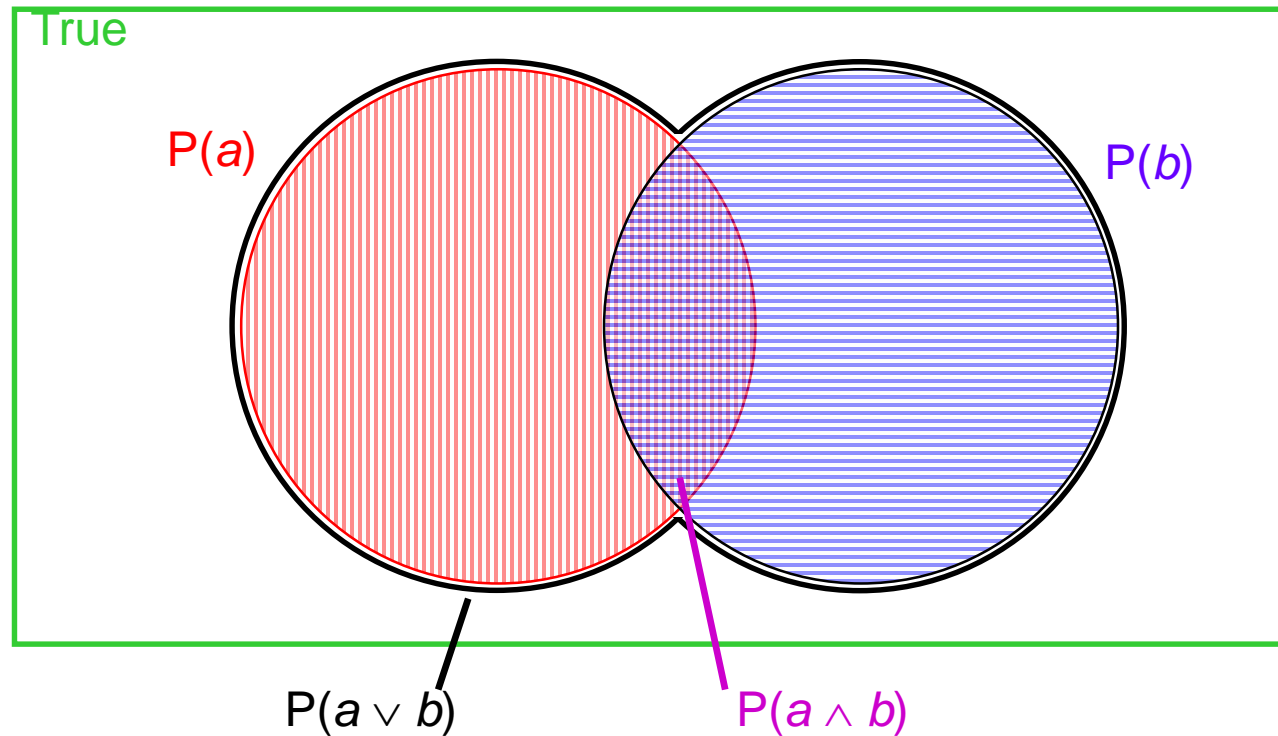
Kolmogorov's
Axioms: all other
probability rules
can be derived
from these



Review: Axioms of Probability (2)

Illustration of Sum Rule:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$





Unconditional and Conditional Probability

Unconditional Probability:

- $P(a)$ = degree of belief in proposition **a** in *absence* of any other information
- Also known as **prior probability**:
Belief *prior* to arrival of any new information
- Specified as a **probability distribution**:

$P(\text{ExamResult}) = \langle 0.2, 0.3, 0.3, 0.15, 0.05 \rangle$

$P(\text{Anna_Here}=\text{true}) = .98, P(\text{Anna_Here}=\text{false}) = .02$

Conditional Probability:

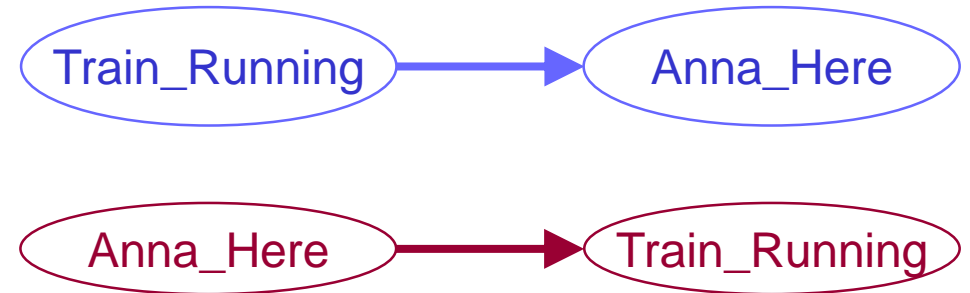
- Also known as **posterior probability**: Belief *post* arrival of new information
- Probability is *conditioned* by other evidence
- **$P(\text{Anna_Here}=\text{false} \mid \text{Train_Running}=\text{false}) = 0.8$**
“Prob. That Anna is NOT Here is 0.8,
given that all you know is that the Train is NOT running.”



Conditional Probability Graphically ...

Train_Running and **Anna_Here** interact

- Knowing about **Train_Running** gives us evidence about **Anna_Here**
- Or vice versa



Diagrams are easiest to understand if we think about causality in drawing them:

- Which causes which?
- Hidden causes?



Joint Probability Distribution

For a set of random variables, **joint probability distribution** gives probability of every **atomic event** on those variables

- For n Boolean variables, size is 2^n (exponential in no. of variables)
- Later: represent more compactly because of independence
- **P(Weather, Tennis)**: 4×2 matrix of values:

	Weather = sunny	Weather = rain	Weather = cloudy	Weather = snow
Tennis = true	0.144	0.02	0.016	0.02
Tennis = false	0.576	0.08	0.064	0.08



A Real-Life Joint Probability Distribution ...



Nate Silver  @NateSilver538 · 1h 

The joint probabilities are as follows, per our Deluxe model.

D Senate + D House: 18%

D Senate + R House: <1%

R Senate + D House: 68%

R Senate + R House: 14%

So still better than a 30% chance that *either* the House or the Senate will result in an upset tonight. Pretty exciting!



Independence

New evidence may be irrelevant:

$$\begin{aligned} &P(\text{Anna_Here}=\text{false} \mid \text{Train_Running}=\text{False}, \text{Exam_Result}=a) \\ &= P(\text{Anna_Here}=\text{false} \mid \text{Train_Running}=\text{False}) \\ &= 0.8 \end{aligned}$$

This indicates independence between variables:

- Exam_Result independent of Anna_Here
- Also known as **absolute independence**
- Diagram: no arc

Such simplifications very important

- Can greatly reduce the number of combinations we need to consider

Anna_Here

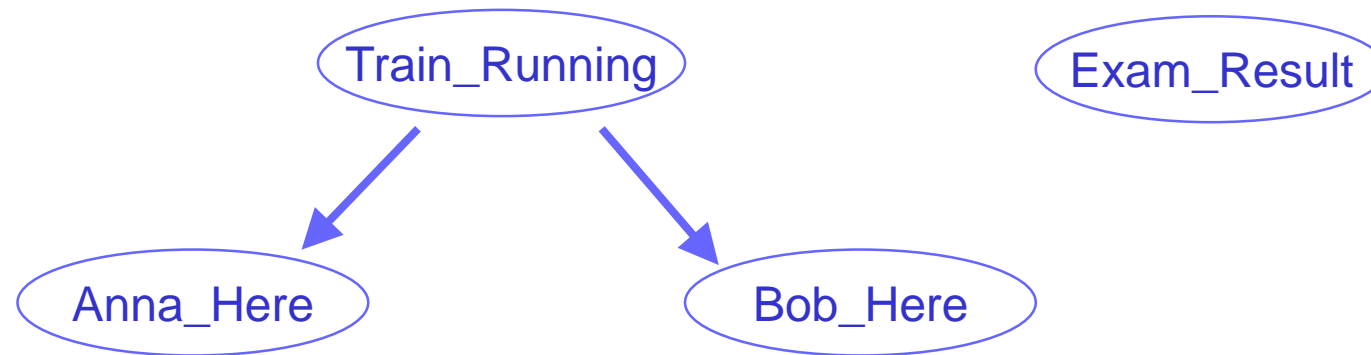
Exam_Result



Conditional Independence (1)

Conditional Independence is different from absolute independence

- **Anna** and **Bob** both take the train, so if Anna is not here, it is **more likely** that Bob is not here
- Bob being here is **not completely independent** of Anna being here; they are both dependent on the Train running
- Anna_Here **conditionally independent of** Bob_Here **given** Train_Running
- Graphically:

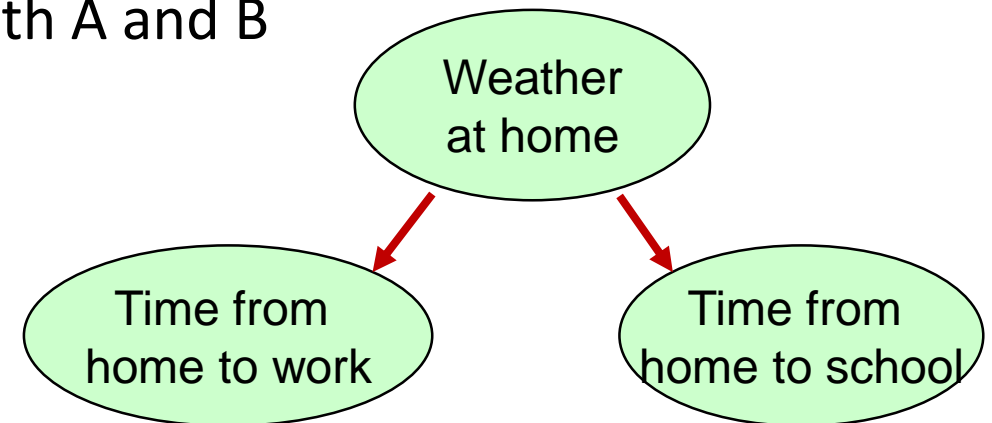
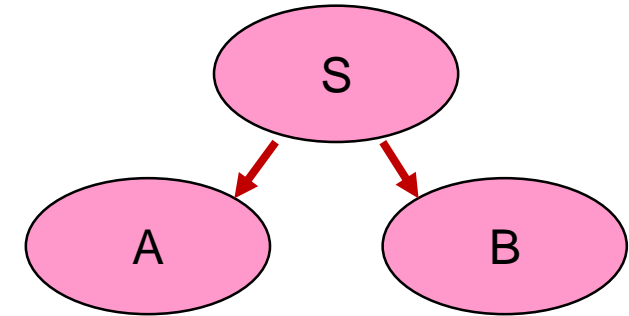




Conditional Independence (2)

- Note also:

- We say that A and B are ***d-separated*** by S (D=directional)
- If we know value of S, then knowing value of A does not give us any additional information about value of B
- However, in the absence of information about the separating variable S, knowing A **does** give us information about B
- Relates to S being the root cause of both A and B





Probability Formulae (1)

Absolute Independence:

- If two variables are **independent**, the probability of **both** happening is the **product** of their individual probabilities

$$P(A \wedge B) = P(A) P(B) \Leftrightarrow A \text{ independent of } B$$

- Example:

$$P(\text{Suit}=\heartsuit) = 1/4. \quad P(\text{Value}=Q) = 1/13.$$

$$P(\text{Suit}=\heartsuit \wedge \text{Value}=Q) = 1/52.$$

Conditional Independence:

- If A is **conditionally independent** of B given X:

$$P(A \wedge B \mid X) = P(A \mid X) P(B \mid X) \Leftrightarrow A \text{ cond. indep. of } B \text{ given } X$$

If X is known, knowledge of A's value will not affect opinion of B's value, so can apply same rule as for absolute independence



Probability Formulae (2)

Product Rule (conjunctions):

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

Example:

$$P(\text{⚡} \mid \text{storm}) = 0.6, P(\text{storm}) = 0.01$$

$$\Rightarrow P(\text{⚡} \wedge \text{storm}) = 0.006$$

If lightning occurs in 60% of storms,
and a storm occurs 1% of the time,
then lightning and storm together occur 0.6% of the time

Conditional Probability of two propositions:

$$P(a \mid b) = P(a \wedge b) / P(b), \text{ for } P(b) > 0$$



Probability Formulae (3)

Theorem of Total Probability:

- If domain of A is $\langle a_1, a_2, \dots, a_n \rangle$, then:

$$P(B) = \sum_n P(B | a_i) P(a_i)$$

$$P(\text{⚡}) = P(\text{⚡} | \text{storm}) P(\text{storm}) + P(\text{⚡} | \neg \text{storm}) P(\neg \text{storm})$$

Total probability of lightning occurring =
prob. of it *with* a storm + prob. of it *without* a storm

Bayes' Rule:

$$P(B | A) = P(A | B) P(B) / P(A)$$

- Follows from Product Rule
- Allows us to reason about causes when we have observed effects



Part 1B: Reasoning with Bayes' Rule



Reasoning with Bayes' Rule

Bayes' Rule:

$$P(b \mid a) = P(a \mid b) P(b) / P(a)$$

Easily derived from Product Rule:

$$P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$$

$$\text{Divide by } P(a): P(a \mid b) P(b) / P(a) = P(b \mid a)$$



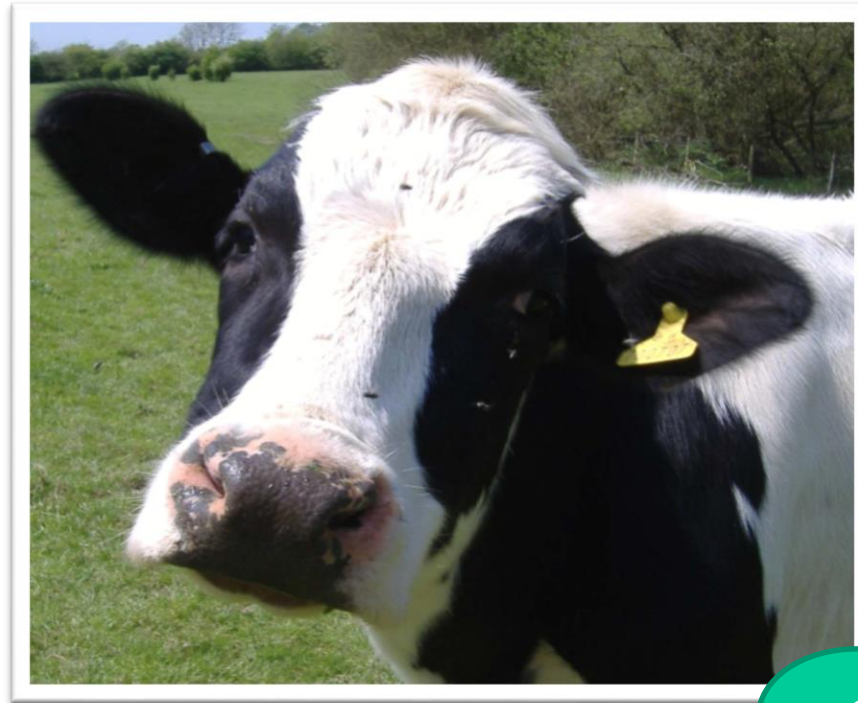
Probably not the
Rev. Thomas Bayes
1702-1761

Applies to variables as well as propositions:

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

What's the point?

Allows us to reason about an **unobserved cause** (B) when we have observed the **effect** (A)



x 20



$$P(\text{cemetery caused by cow}) = P(\text{cow causes cemetery}) \times \frac{P(\text{cow})}{P(\text{cemetery})}$$

1



$$P(\text{coffin caused by cow}) \\ = P(\text{cow causes coffin}) \times P(\text{cow})$$

$$P(\text{coffin caused by shark}) \\ = P(\text{shark causes coffin}) \times P(\text{shark})$$



A disease affects 1 in 1000 people

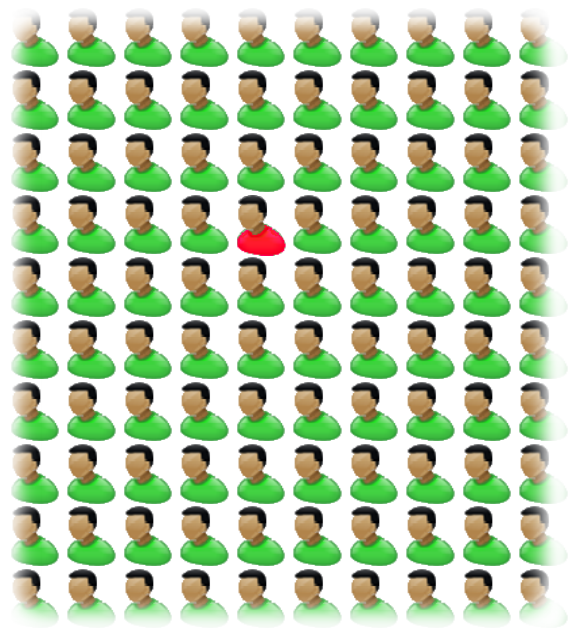
A Person Tests +

Test is 99% accurate

Probably
have it?

Probably
don't?



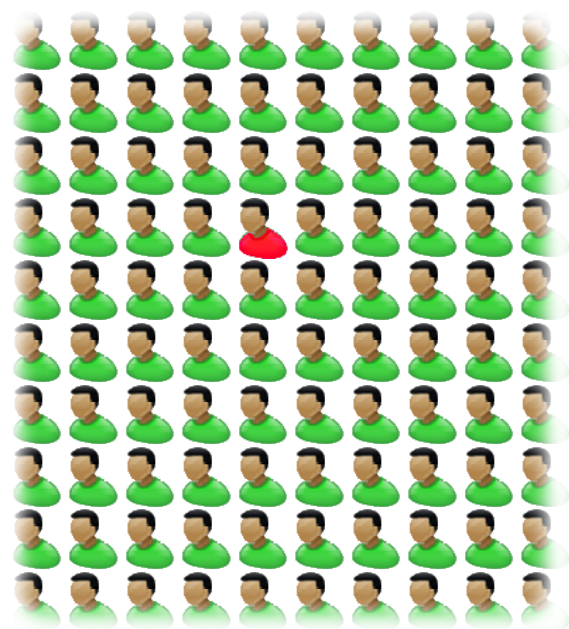


Disease:
1 in 1000

100

100,000

99,900



Disease:
1 in 1000

100,000

100

99%

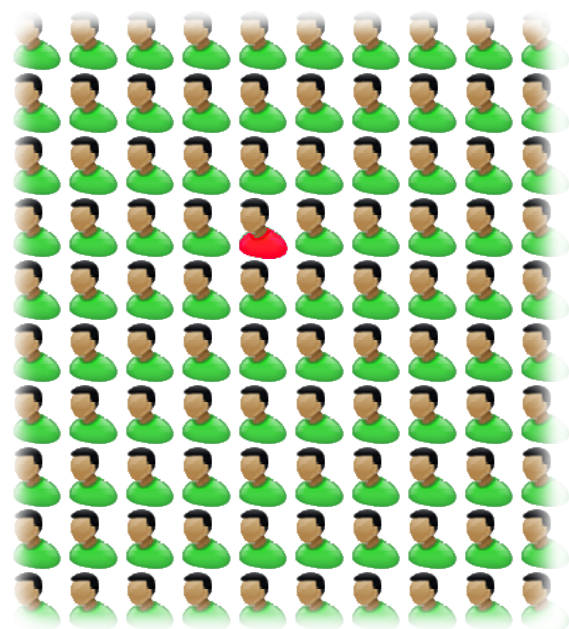
99

Test +

1%

1

99,900



Disease:
1 in 1000

100,000

100

99%

99

Test +

1%

1

99,900

99%

98,901

1%

999

Test +

1098 people *test +*
99 of them *actually +*
 $\Rightarrow P(\text{Disease} \mid \text{Test}+) = 9\%$



Bayes' Rule: Example

Patient has sore throat: caused by influenza?

- **$P(\text{sore_throat} \mid \text{flu}) = 0.75$**

75% probability of a patient with flu having a sore throat, based on experience of patients with flu

- **$P(\text{sore_throat}) = 0.1$:**

1 in 10 patients have sore throat, based on observations at the surgery

- **$P(\text{flu}) = 0.02$:**

1 in 50 patients have flu

$\Rightarrow P(\text{flu} \mid \text{sore_throat}) = .75 \times .02 / .1 = 0.15$

15% probability based on this evidence



Bayes' Rule with Normalisation (1)

Alternative formulation of Bayes' Rule

- Avoid needing to know **prior probability of evidence**:
In this example, **P(sore_throat)**
- Instead, compute posterior probability for each value of **query variable**, and normalise:

$$P(\text{flu} \mid \text{sore_th}) = P(\text{sore_th} \mid \text{flu}) \times P(\text{flu}) / P(\text{sore_th})$$

$$P(\neg \text{flu} \mid \text{sore_th}) = P(\text{sore_th} \mid \neg \text{flu}) \times P(\neg \text{flu}) / P(\text{sore_th})$$



Bayes' Rule with Normalisation (2)

Some observations:

- The probabilities $P(\text{flu} \mid \dots)$ and $P(\neg\text{flu} \mid \dots)$ must sum to 1
- Both times we are dividing by same term, $P(\text{sore_th})$, so we can eliminate it by *normalising*
 $P(\text{flu} \mid \dots)$ and $P(\neg\text{flu} \mid \dots)$
- Probability of not having flu is easily found:
 $P(\neg\text{flu}) = 1 - P(\text{flu})$



Bayes' Rule with Normalisation (3)

Applying this to the Influenza example:

- Use α to denote the **normalisation constant**
(this will be equal to $1/P(\text{sore_th})$)

$$P(\text{flu} \mid \text{sore_th}) = \alpha \times P(\text{sore_th} \mid \text{flu}) \times P(\text{flu})$$

$$P(\neg \text{flu} \mid \text{sore_th}) = \alpha \times P(\text{sore_th} \mid \neg \text{flu}) \times P(\neg \text{flu})$$

- We already have:

$$P(\text{sore_throat} \mid \text{flu}) = 0.75, \quad P(\text{flu}) = 0.02$$

$$P(\neg \text{flu}) = 1 - P(\text{flu}) = 0.98$$

- In this case, we also need to know $P(\text{sore_th} \mid \neg \text{flu})$,
probability of having a sore throat when you don't have flu:
From observations at the surgery, this is **0.087**



Bayes' Rule with Normalisation (4)

Can then calculate:

$$P(\text{flu} \mid \text{sore_th}) = \alpha \times 0.75 \times 0.02 = \alpha 0.015$$

$$P(\neg\text{flu} \mid \text{sore_th}) = \alpha \times 0.087 \times 0.98 = \alpha 0.08526$$

To eliminate α , normalise the numbers (sum is 0.10026):

$$P(\text{flu} \mid \text{sore_th}) = 0.015 / 0.10026 = \mathbf{0.15} \text{ (as before)}$$

$$P(\neg\text{flu} \mid \text{sore_th}) = 0.08526 / 0.10026 = \mathbf{0.85}$$

General form of Bayes' Rule with Normalisation:

$$P(Y \mid X) = \alpha P(X \mid Y) P(Y)$$

where α is constant to make $P(Y|X)$ entries sum to 1



Bayes' Rule with Normalisation (5)

General form of Bayes' Rule with Normalisation:

$$P(Y | X) = \alpha P(X | Y) P(Y) \text{ where } \alpha \text{ is normalisation constant}$$

The Product Rule that we saw earlier states that:

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

If we combine both equations, we see that:

$$P(Y | X) = \alpha P(X \wedge Y)$$

Therefore, to calculate the probability of **X given Y**, we can calculate probability of **X and Y** and normalise result to get the final answer.



Bayes' Rule: Discussion

Why not just measure $P(\text{flu} \mid \text{sore_throat})$?

- This **diagnostic probability** could be assessed from observations, as the other values are

Rationale: causal probabilities are more robust

- If there's a flu outbreak, $P(\text{flu})$ will increase
- $P(\text{flu} \mid \text{sore_throat})$ is difficult to re-assess directly: counts from before outbreak will not be applicable
- However, $P(\text{flu})$ can be measured again and $P(\text{sore_throat} \mid \text{flu})$ is **unchanged**
- From Bayes' Rule, see that $P(\text{flu} \mid \text{sore_throat})$ should increase proportionately with $P(\text{flu})$

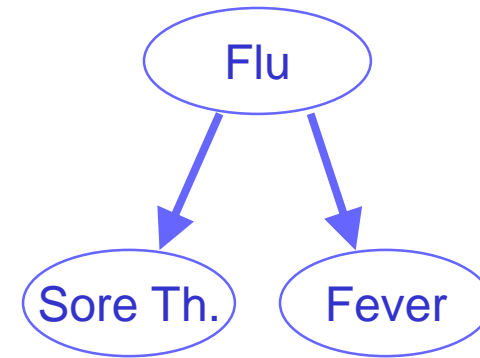


Bayes' Rule: Combining Evidence

Suppose patient also has fever: effect on reasoning?

$$P(\text{flu} \mid \text{sore_th} \wedge \text{fever}) = \alpha P(\text{sore_th} \wedge \text{fever} \mid \text{flu}) P(\text{flu})$$

- Problem: To apply Bayes' Rule directly, need *joint probability*: $\text{sore_th} \wedge \text{fever}$



Observation: sore_th **conditionally independent** of fever given flu

- Can use Conditional Independence formula

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

- Therefore, only need individual conditional probabilities:

$$P(\text{flu} \mid \text{sore_th} \wedge \text{fever}) = \alpha P(\text{sore_th} \mid \text{flu}) P(\text{fever} \mid \text{flu}) P(\text{flu})$$

- Just calculate both using Bayes' Rule and multiply them.



Bayesian Updating

- Bayes' Rule can be applied iteratively, for sequential testing
- Flu example can operate in this way:
Initially, have prior probability of flu: $P(\text{flu})$
Then, check whether patient has sore throat and compute $P(\text{flu} \mid \text{sore_th})$ by applying Bayes' Rule to update $P(\text{flu})$
Then, check whether patient has a fever and compute $P(\text{flu} \mid \text{sore_th} \wedge \text{fever})$ by applying Bayes' Rule to $P(\text{flu})$
- It is easy to check that this gives the same result in the case where we assume sore_th and fever are independent