Principles of
Machine Learning

Week 11: Probabilistic Machine Learning (Part 1)

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

- Part 1A: Review of Probability Basics
- Part 1B: Unconditional and Conditional Probability
- Part 1C: Probability Formulae
- Topic Part 1D: Reasoning with Bayes' Rule
- Part 1E: Bayes' Rule with Normalisation

Week 12

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

Principles of Machine Learning

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?

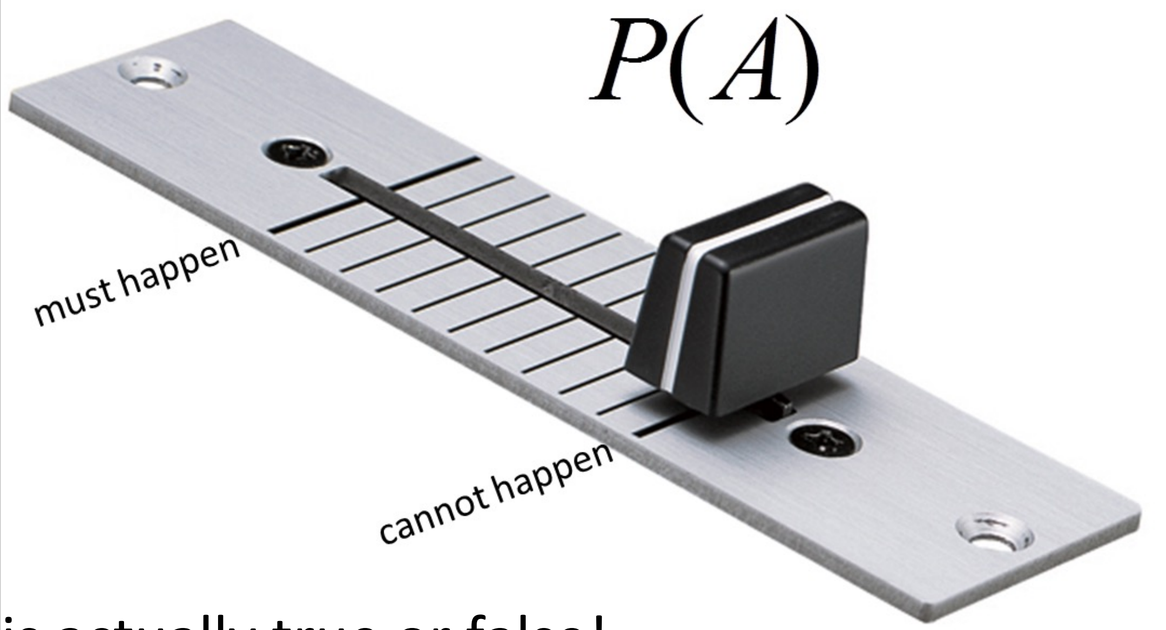
Probability:

- 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(Card = Q♥) = 1/52**

  After looking: **P(Card = Q♥) = 0 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(Cavity=true) is also written as **P(cavity)**
  P(Cavity=false) is also written as **P(¬cavity)**
- General case of **discrete random variable**: values in domain are **mutually exclusive and exhaustive**
  ExamResult: <a, b, c, d, fail>
  P(ExamResult=a) = 0.2, …
  P(ExamResult) = <0.2, 0.3, 0.3, 0.15, 0.05>

Review: Probability Notation (2)

Probabilities of Values in a Domain Sum to 1

- Provided they are mutually exclusive and exhaustive:
  P(Elvis=alive) = .01; P(Elvis=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:

  $Anna\_Here = false \wedge Bob\_Here = false$

  $Anna\_Here = false \wedge Bob\_Here = true$

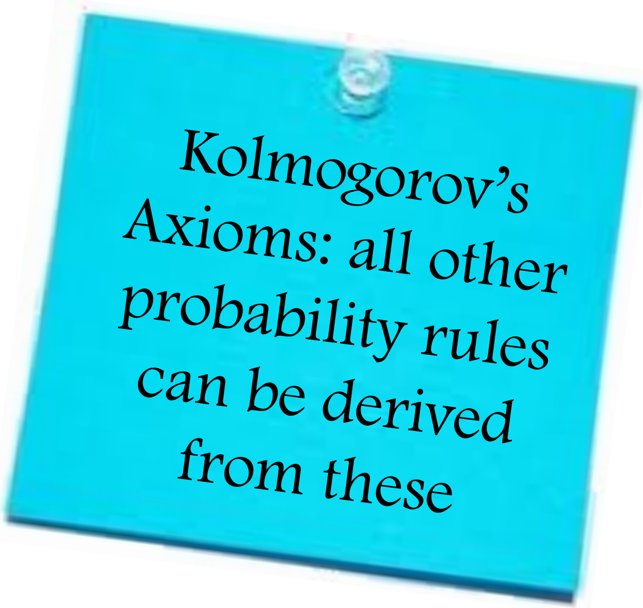
  $Anna\_Here = true \wedge Bob\_Here = false$

  $Anna\_Here = true \wedge Bob\_Here = true$

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

Review: Axioms of Probability (1)

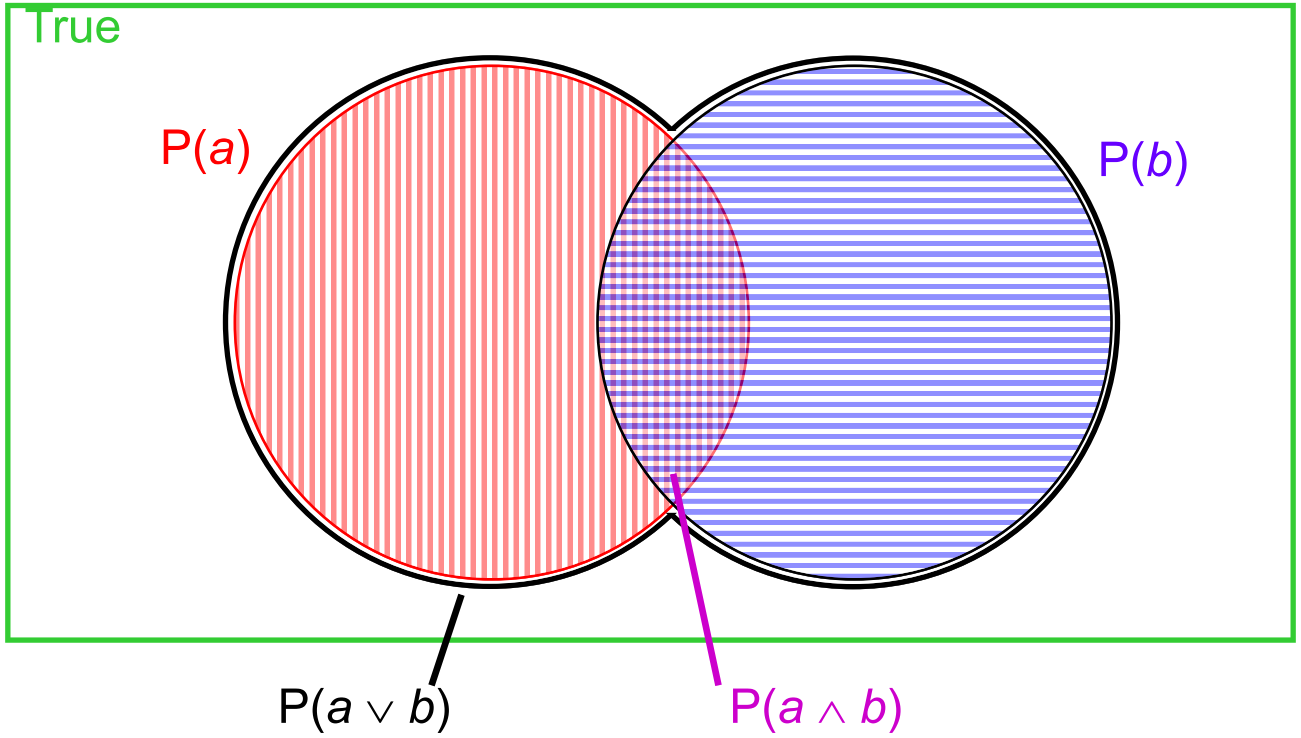
1. Probability of any proposition is between 0 and 1:
   $0 \leq P(a) \leq 1$
2. Necessarily true propositions have probability 1; necessarily false propositions have probability 0:
   $P(\mathit{false}) = 0, \; P(\mathit{true}) = 1$
3. Probability of a disjunction ($a \vee b$) is:
   $P(a \vee b) =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



Review: Axioms of Probability (2)

Illustration of Sum Rule:

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



Principles of Machine Learning

Part 1B: Unconditional and Conditional Probability

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(ExamResult) = <0.2, 0.3, 0.3, 0.15, 0.05>
  P(Anna_Here=true) = .98, P(Anna_Here=false) = .02

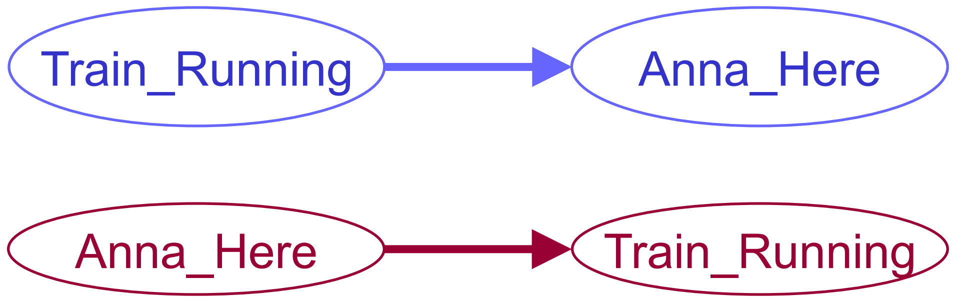
Conditional Probability:

- Also known as **posterior probability**: Belief *post* arrival of new information
- Probability is *conditioned* by other evidence
- **P(Anna_Here=false | Train_Running=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:

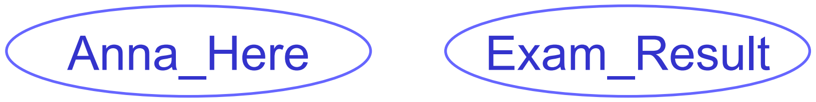
P(Anna_Here=false | Train_Running=False, Exam_Result=a)
= P(Anna_Here=false | Train_Running=False)
= 0.8

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



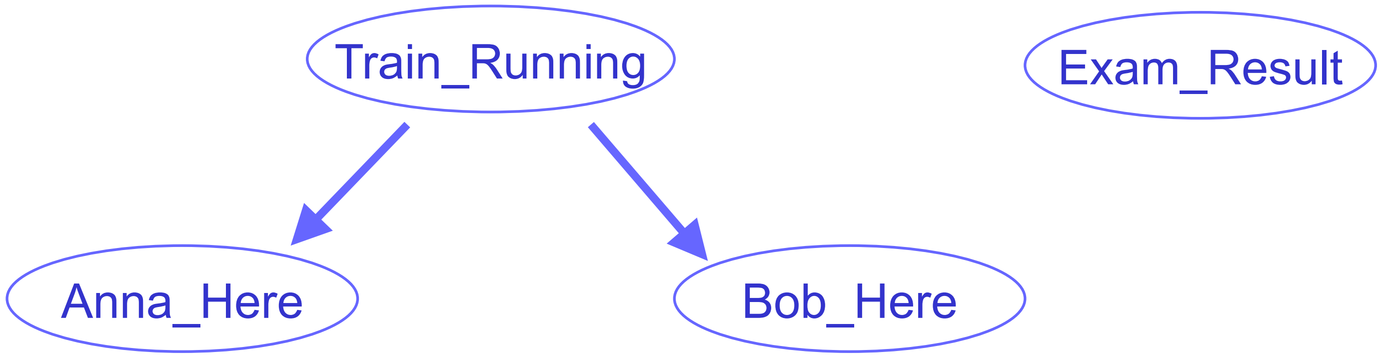
Conditional Independence (1)

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

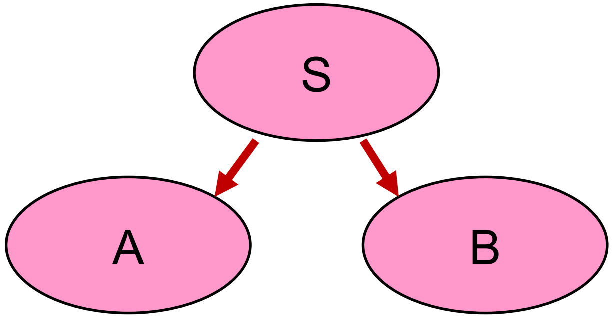
- Anna_Here **conditionally independent of** Bob_Here **given** Train_Running

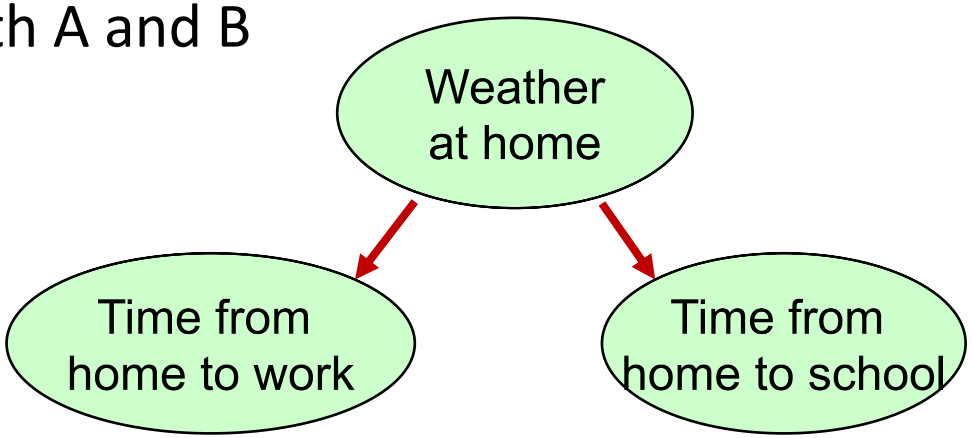
- Graphically:



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





Principles of Machine Learning

Part 1C: Probability Formulae

Probability Formulae (1)

Absolute Independence:

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

  $P(A \wedge B) = P(A)\,P(B) \iff$ A independent of B

- Example:

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

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

Conditional Independence:

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

  $P(A \wedge B \mid X) = P(A \mid X)\,P(B \mid X) \iff$ A cond. indep. of B given X

  If X is known, knowledge of A’s value will not affect opinion of B’s value, so we 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(⚡ | storm) = 0.6, P(storm) = 0.01

=> P(⚡ $\wedge$ 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:

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

Probability Formulae (3)

Theorem of Total Probability:

– If domain of A is <a_1, a_2, …, a_n>, then:

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

P(⚡) = P(⚡ | storm) P(storm) + P(⚡ | ¬ storm) P(¬ storm)

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

Bayes' Rule:

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

– Follows from Product Rule

– Allows us to reason about causes when we have observed effects

Principles of Machine Learning

Part 1D: 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:

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

Divide by $P(a)$: $P(a \mid b)\, P(b) \,/\, P(a) = P(b \mid a)$

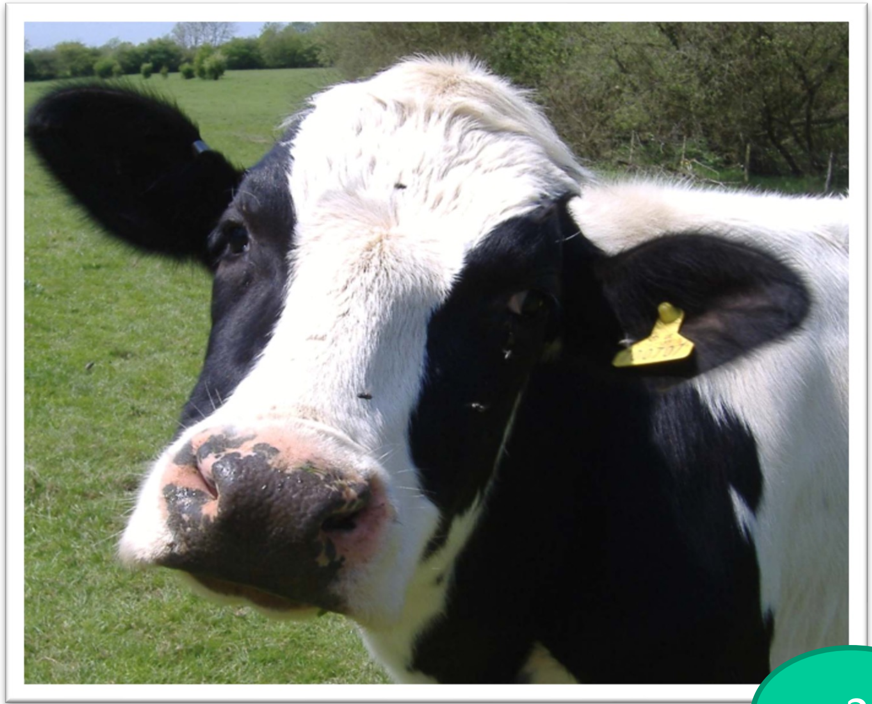
Probably not the Rev. Thomas Bayes 1702-1761

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( *caused by* )

= P( *causes* ) × P() / P()

1

$P($  *caused by*  $)$

$= P($  *causes*  $) \times P($  $)$

$P($  *caused by*  $)$

$= P($  *causes*  $) \times P($  $)$

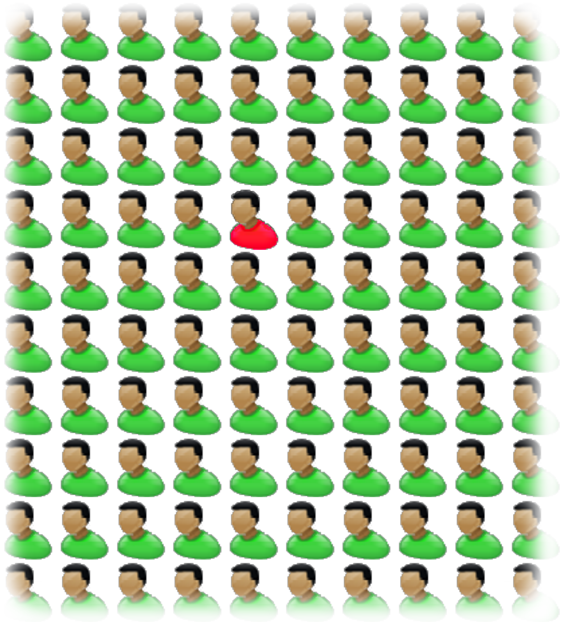
A disease affects 1 in 1000 people

A Person Tests +

Test is 99% accurate

Probably have it?

Probably don't?

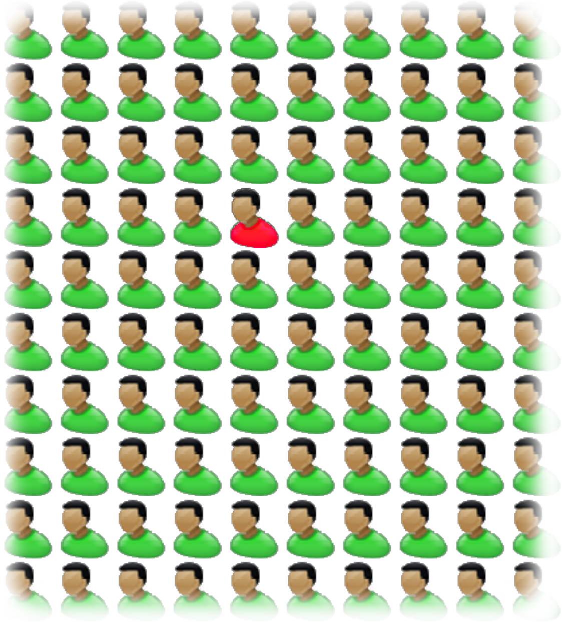


Disease:
1 in 1000

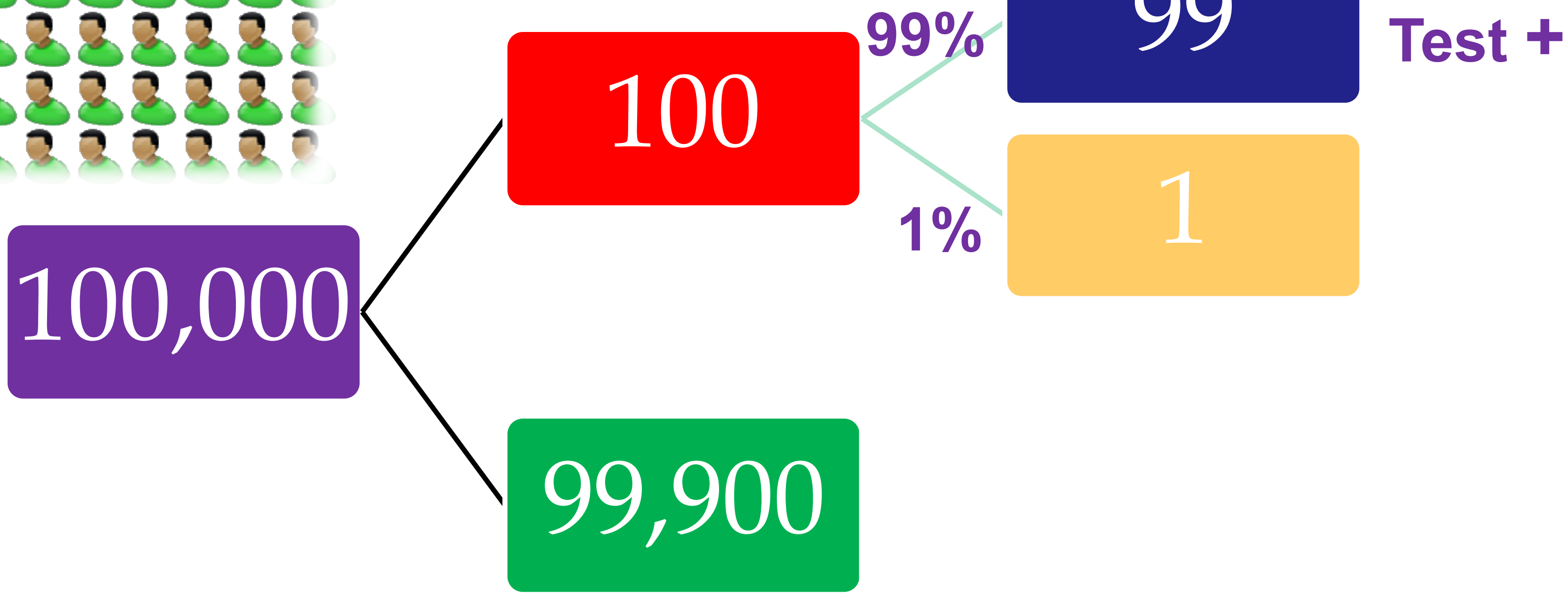
100,000

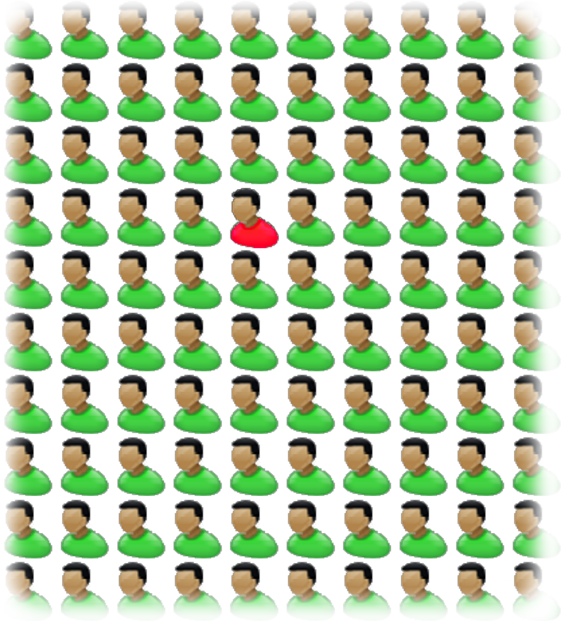
100

99,900



Disease: 1 in 1000





Disease:
1 in 1000

100,000

100

99% → 99 — Test +

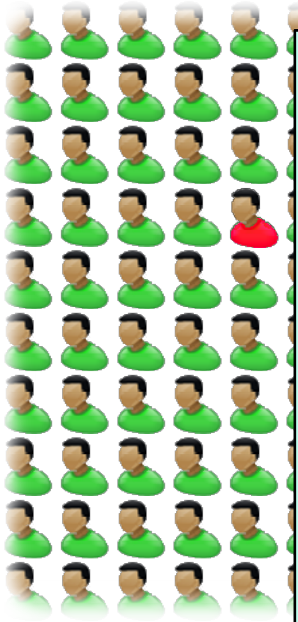
1% → 1

99,900

99% → 98,901

1% → 999 — Test +

1098 people *test* +
99 of them *actually* +
=> P(Disease | Test +) = **9%**



P(b | a) = P(a | b) P(b) / P(a)

a: Test=+
b: Disease=t

P(Disease=t | Test=+) =
P(Test=+ | Disease=t) * P(Disease=t) / P(Test=+)

P(Test=+ | Disease=t) = 0.99 [accuracy of test]
P(Disease=t) = 0.001 [1 in 1000 have disease]
P(Test=+): overall prob of a + result: can get this from info on screen: (99+999) positives from 100,000 tests
=> P(Test=+) = 1098/100000 = 0.01098
(Note: can use Theorem of Total Probability for this.)

Therefore, P(Disease=t | Test=+) = 0.99 * 0.001 / 0.01098
= 0.09164 = 9.064%

This is identical to what we compute from the tree diagram:
P(Disease=t | Test=+) = 99 / 1098 = = 0.09164 = 9.064%

10

01

Test +

Test +

1098

99 of t

=> *P*(Dis

Bayes' Rule: Example

Patient has sore throat: caused by influenza?

- **P(sore_throat | flu) = 0.75**
  75% probability of a patient with flu having a sore throat, based on experience of patients with flu
- **P(sore_throat) = 0.1**:
  1 in 10 patients have sore throat, based on observations at the surgery
- **P(flu) = 0.02**:
  1 in 50 patients have flu

=> P(flu | sore_throat) = .75 × .02 / .1 = 0.15
15% probability based on this evidence

Principles of Machine Learning

Part 1E: Bayes' Rule with Normalisation

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(flu | …)** and **P(¬flu | …)** must sum to 1
- Both times we are dividing by same term, **P(sore_th)**, so we can eliminate it by *normalising* **P(flu | …)** and **P(¬flu | …)**
- Probability of not having flu is easily found: **P(¬flu) = 1 – P(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(sore_th)**)
  P(flu | sore_th) = α ×P(sore_th | flu) ×P(flu)
  P(¬flu | sore_th) = α ×P(sore_th | ¬flu) ×P(¬flu)
- We already have:
  P(sore_throat | flu) = 0.75, P(flu) = 0.02
  P(¬ flu) = 1 – P(flu) = 0.98
- In this case, we also need to know **P(sore_th | ¬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:

$P(flu \mid sore\_th) = \alpha \times 0.75 \times 0.02 = \alpha\ 0.015$

$P(\neg flu \mid sore\_th) = \alpha \times 0.087 \times 0.98 = \alpha\ 0.08526$

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

$P(flu \mid sore\_th) = 0.015 / 0.10026 = \mathbf{0.15}$ **(as before)**

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

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:

$$P(Y \mid X) = \alpha \, P(X \mid Y) \, P(Y)$$ where **α** is normalisation constant

The Product Rule that we saw earlier states that:

$$P(X \wedge Y) = P(X \mid Y) \, P(Y)$$

If we combine both equations, we see that:

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

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(flu | sore_throat)?

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

Rationale: causal probabilities are more robust

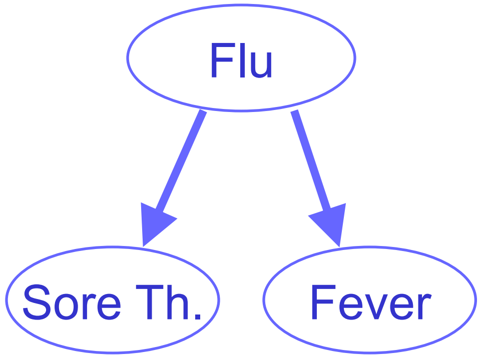
- If there's a flu outbreak, **P(flu)** will increase
- **P(flu | sore_throat)** is difficult to re-assess directly: counts from before outbreak will not be applicable
- However, **P(flu)** can be measured again and **P(sore_throat | flu)** is **unchanged**
- From Bayes' Rule, see that **P(flu | sore_throat)**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

Bayes' Rule: Combining Evidence

Suppose patient also has fever: effect on reasoning?

$P(flu \mid sore\_th \wedge fever) = \alpha\, P(sore\_th \wedge fever \mid flu)\, P(flu)$

- Problem: To apply Bayes' Rule directly, need *joint probability:* **$sore\_th \wedge fever$**



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:

  $P(flu \mid sore\_th \wedge fever) = \alpha\, P(sore\_th \mid flu)\, P(fever \mid flu)\, P(flu)$
- 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(flu)**

  Then, check whether patient has sore throat and compute **P(flu | sore_th)** by applying Bayes' Rule to update **P(flu)**

  Then, check whether patient has a fever and compute **P(flu | sore_th $\wedge$ fever)** by applying Bayes' Rule to **P(flu)**

- It is easy to check that this gives the same result in the case where we assume sore_th and fever are independent