Bayes' Theorem: Understanding Probability and Learning from Evidence

A journey through one of mathematics' most powerful tools for reasoning under uncertainty

What is Bayes' Theorem?

Bayes' Theorem is a mathematical formula that allows us to update the probability of a hypothesis as new evidence becomes available. It elegantly expresses conditional probability through the relationship:

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

Named after Reverend Thomas Bayes, an 18th-century mathematician and Presbyterian minister who developed the foundational concepts of probabilistic inference.

About Bayes Theorem

- Bayes' theorem is also known as **Bayes' rule, Bayes' law**, or **Bayesian reasoning**, which determines the probability of an event with uncertain knowledge.
- Bayes' theorem was named after the British mathematician **Thomas Bayes**.
- It is a way to **calculate the value of P(B|A) with the knowledge of P(A|B).**
- Bayes' theorem allows updating the probability prediction of an event by observing new information of the real world.

Bayes Theorem Formula

Bayes' Theorem gives the conditional probability of an event A given another event B has occurred

Bayes Theorem

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

where:

$P(A|B)$ = Conditional Probability of A given B

$P(B|A)$ = Conditional Probability of B given A

$P(A)$ = Probability of event A

$P(B)$ = Probability of event A

About Bayes Theorem Formula

- **P(A|B)** is known as **Posterior**, which we need to calculate, and it will be read as Probability of hypothesis A when we have **occurred an evidence B.**
- **P(B|A)** is called the **Likelihood**, in which we consider that hypothesis is true, then we calculate the **probability of evidence**.
- **P(A)** is called the **Prior Probability**, probability of hypothesis **before considering the evidence**
- **P(B)** is called **Marginal Probability**, **pure probability of an evidence.**

$$\underbrace{P(A|B)}_{\text{Posterior}} = \frac{\overbrace{P(B|A)}^{\text{Likelihood}}.\overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Evidence}}}$$

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

$$P(King|Face) = \frac{P(Face|King)P(King)}{P(Face)}$$

Example of Bayes Theorem

- From the deck of the cards, find the probability of the card being picked is king given that it is the face card.
- This can be represented as : **P(King|Face)**

- There are total 52 cards in a deck (n=52), from which 12 cards are face cards; king, queen and jack with club, diamond, heart and spade
- There is a set of face cards with 12 members.
- There is a subset of king cards with 4 members.
- So 4 face cards out of 12 face cards (4/12) are king cards.

According to Bayes' theorem: $P(King|Face) = \frac{P(Face|King)P(King)}{P(Face)}$

- 4 members are king cards out of 52 cards: $P(King) = \frac{4}{52} = \frac{1}{13}$
- 12 members are face cards out of 52 cards: $P(Face) = \frac{12}{52} = \frac{3}{13}$
- Probability of being face given king: $P(Face|King) = 1$

Here we have prior knowledge that if the card is king, it is face card only. It is 100% sure so it is always 1.

Example of Bayes Theorem

- **Putting all in Bayes' formula:**

$$P(King|Face) = \frac{P(Face|King)P(King)}{P(Face)}$$

$$= \frac{1 * \frac{1}{13}}{\frac{3}{13}}$$

$$= \frac{1}{3}$$

➢Bayes' Theorem proof:

$$P(King|Face) = P(King \cap Face)/P(Face)$$

So from the image above

$$P(King \cap Face) = 4$$

$$P(Face) = 12$$

$$P(King|Face) = 4/12 = 1/3$$

$$\therefore 4/12 = 1/3$$

Problem 1: Detection of Disease

1% of people have a certain disease.
A test detects the disease correctly 95% of the time.
5% false positives.
If a person tests positive, what is the chance they actually have the disease?

Given:

- P(Disease) = 0.01
- P(No Disease) = 0.99
- P(Positive | Disease) = 0.95
- P(Positive | No Disease) = 0.05

Step 1: Compute overall probability of testing positive

$$P(Positive) = (0.95 \cdot 0.01) + (0.05 \cdot 0.99)$$

$$= 0.0095 + 0.0495 = 0.059$$

Step 2: Apply Bayes

$$P(Disease|Positive) = \frac{0.95 \times 0.01}{0.059} \approx 0.161$$

Conclusion:

Even after testing positive, the chance of disease is only **16%**!

Problem 2: Email Spam Filter

A spam filter knows:

- 30% of emails are spam → $P(S) = 0.3$
- Word "lottery" appears in:
  - 40% of spam emails → $P(L \mid S) = 0.40$
  - 2% of non-spam emails → $P(L \mid S') = 0.02$

Question:

If an email contains the word lottery, what is the probability it is spam?

Solution

Step 1: Compute P(L)

$$P(L) = (0.40)(0.30) + (0.02)(0.70)$$

$$= 0.12 + 0.014 = 0.134$$

Step 2: Apply Bayes

$$P(S|L) = \frac{0.40 \cdot 0.30}{0.134} = \frac{0.12}{0.134} \approx 0.896$$

Answer:

89.6% chance the email is spam.

Problem 3: Student Test Accuracy

A test for a subject has:

- Pass rate = 60% → $P(\text{Pass}) = 0.6$
- Students who know the subject answer correctly = 90% → $P(\text{Correct} \mid \text{Know}) = 0.9$
- Students who don't know answer correctly = 20% → $P(\text{Correct} \mid \text{Not Know}) = 0.2$

Question:

If a student answered correctly, what is the probability that the student actually knows the subject?

Solution

P(Know) = 0.6

P(Not Know) = 0.4

$$P(Correct) = (0.9)(0.6) + (0.2)(0.4) = 0.54 + 0.08 = 0.62$$

$$P(Know|Correct) = \frac{0.9 \cdot 0.6}{0.62} = \frac{0.54}{0.62} \approx 0.871$$

Answer:

87.1%, the student likely knows the subject.

Problem 4: Defective Items in Factory

Two machines (A and B) manufacture items.

- Machine A produces 70% of items → $P(A)=0.7$
- Machine B produces 30% → $P(B)=0.3$
- Defect rates:
  - A: 2% defective → $P(D \mid A)=0.02$
  - B: 6% defective → $P(D \mid B)=0.06$

Question:
If an item selected at random is defective, what is the probability it came from Machine B?

Solution

$$P(D) = (0.02)(0.7) + (0.06)(0.3) = 0.014 + 0.018 = 0.032$$

$$P(B|D) = \frac{0.06 \cdot 0.3}{0.032} = \frac{0.018}{0.032} \approx 0.5625$$

Answer:

56.25%, it is more likely the defective item came from Machine B.

Problem 5: Medical Test with False Positives

A disease affects 2% of the population.

A test has:

- True positive = 95%
- False positive = 10%

If a patient tests positive, what is the probability they actually have the disease?

Solution

P(D) = 0.02

P(Not D) = 0.98

P(Pos | D) = 0.95

P(Pos | Not D) = 0.10

$$P(Pos) = (0.95)(0.02) + (0.10)(0.98) = 0.019 + 0.098 = 0.117$$

$$P(D|Pos) = \frac{0.95 \cdot 0.02}{0.117} = 0.162$$

Answer:

16.2% chance of actually having the disease.

Problem 6: Rain and Weather Forecast Accuracy

In a city:

- It rains 20% of days → P(R) = 0.2
- Weather forecast is correct 80% of rainy days → P(Forecast Rain | R) = 0.8
- It wrongly predicts rain on 10% of dry days → P(Forecast Rain | No R) = 0.1

If the forecast says “Rain Today”, what is the probability it will really rain?

Solution

$$P(FR) = (0.8)(0.2) + (0.1)(0.8) = 0.16 + 0.08 = 0.24$$

$$P(R|FR) = \frac{0.8 \cdot 0.2}{0.24} = \frac{0.16}{0.24} = 0.6667$$

Answer:

66.67% chance of rain if the forecast says so.

Problem 7: Student Performance

40% of students study regularly → P(S) = 0.4
60% do not → P(N) = 0.6
Probability of scoring above 80%:

- If studied: 70%

- If not studied: 10%

If a student scored above 80%, what is the probability the student studied?

Solution

$$P(Score) = (0.7)(0.4) + (0.1)(0.6) = 0.28 + 0.06 = 0.34$$

$$P(S|Score) = \frac{0.7 \cdot 0.4}{0.34} = \frac{0.28}{0.34} \approx 0.823$$

Answer:

82.3% chance the student studied.

Problem 7: Machine Learning Classification Example

A Naïve Bayes classifier sees:

•40% of documents are “Sports”

•60% are “Politics”

Word “goal” appears:

•30% in Sports → P(goal | Sports)=0.30

•5% in Politics → P(goal | Politics)=0.05

A new document contains the word “goal”. What is the probability the document is Sports?

Solution

$$P(goal) = (0.30)(0.40) + (0.05)(0.60) = 0.12 + 0.03 = 0.15$$

$$P(Sports|goal) = \frac{0.30 \cdot 0.40}{0.15} = \frac{0.12}{0.15} = 0.8$$

Answer:

80% chance the document is Sports.

Bayes' Theorem and Concept Learning

Since Bayes theorem provides a principled way to calculate the posterior probability of each hypothesis given the training data, we can use it as the basis for a straightforward learning algorithm that calculates the probability for each possible hypothesis, then outputs the most probable.

Concept Learning

The process of refining hypotheses about categories and patterns based on observed data and examples

Bayesian Framework

Formalizes how to systematically update belief in a concept given new examples and evidence

Adaptive Intelligence

Enables both machines and humans to improve predictions by incorporating prior knowledge with fresh evidence

Relationship Between Bayes’ Theorem and Concept Learning

•Bayes’ Theorem provides the probabilistic basis for selecting the best concept (hypothesis) in ML.

•Concept learning aims to find a hypothesis that explains the training data.

•Bayes’ theorem computes the probability of each hypothesis after seeing data.

•The most probable hypothesis is the one that has:
✔ High prior probability (initial belief)
✔ High likelihood (fits the data well)

•Therefore, learning becomes belief updating, and noise/uncertainty can be handled naturally.

•Naïve Bayes classifier is a direct application of Bayesian concept learning.

Concept Learning = Bayesian Updating over Hypotheses

llustrative Example

Goal: Learn the concept edible fruit.

Training data:
Apple = edible, Mango = edible, Strawberry = edible,
Mushroom = not edible.

Possible hypotheses:

- **H1:** Red items are edible
- **H2:** All fruits are edible
- **H3:** Everything except mushrooms is edible

Step 1: Assign priors

Maybe H2 has the highest P(H) because it seems biologically meaningful.

Step 2: Calculate how well data fits (likelihood P(D|H))

- H1 fails because mushroom isn't red
- H3 matches all but is too general
- H2 matches perfectly (all fruits in dataset are edible)

Step 3: Compute posterior

$$P(H|D) \propto P(D|H) \cdot P(H)$$

Posterior is highest for H2.

Thus the concept chosen = All fruits are edible.

This is Bayesian concept learning.

Brute-Force Bayes Concept Learning

In particular, assume the learner considers some finite hypothesis space H defined over the instance space X, in which the task is to learn some target concept $c : X \rightarrow \{0,1\}$. As usual, we assume that the learner is given some sequence of training examples $((x_1, d_1) \ldots (x_m, d_m))$ where x_i is some instance from X and where d_i is the target value of x_i (i.e., $d_i = c(x_i)$).

To simplify, we assume the sequence of instances $(x_1 \ldots x_m)$ is held fixed, so that the training data D can

be written simply as the sequence of target values $D = (d_1 \ldots d_m)$

A straightforward concept learning algorithm can be designed to output the maximum a posteriori hypothesis, based on Bayes theorem, as follows:

BRUTE-FORCE MAP LEARNING algorithm

1. For each hypothesis h in H, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2. Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

Maximum A Posteriori (MAP) Hypothesis

- More precisely, we will say that h_{MAP} **is a MAP hypothesis provided**

$$
\begin{aligned}
h_{MAP} &\equiv \underset{h \in H}{\operatorname{argmax}}\, P(h|D) \\
&= \underset{h \in H}{\operatorname{argmax}}\, \frac{P(D|h)\,P(h)}{P(D)} \\
&= \underset{h \in H}{\operatorname{argmax}}\, P(D|h)\,P(h)
\end{aligned}
$$

Brute-Force Bayes Concept Learning

BRUTE-FORCE MAP LEARNING algorithm

1. For each hypothesis h in H, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$$

2. Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\; P(h|D)$$

Brute-Force Bayes Concept Learning

BRUTE-FORCE MAP LEARNING algorithm

- This algorithm may require significant computation, because it applies Bayes theorem to each hypothesis in H to calculate $P(h|D)$.
  - While this is impractical for large hypothesis spaces,
  - The algorithm is still of interest because it provides a standard against which we may judge the performance of other concept learning algorithms.

Brute-Force Bayes Concept Learning

BRUTE-FORCE MAP LEARNING algorithm

- Brute Force MAP learning algorithm must specify values for P(h) and P(D|h).
- P(h) and P(D|h) can be chosen to be consistent with the assumptions:

1. The training data D is noise free.
2. The target concept c is contained in the hypothesis space H
3. We have no a priori reason to believe that any hypothesis is more probable than any other.

Brute-Force Bayes Concept Learning

- Given these assumptions, what values should we specify for P(h)?
- Given no prior knowledge that one hypothesis is more likely than another, it is reasonable to assign the same prior probability to every hypothesis h in H.
- Furthermore, because we assume the target concept is contained in H we should require that these prior probabilities sum to 1.
- Together these constraints imply that we should choose

$$P(h) = \frac{1}{|H|} \quad \text{for all } h \text{ in } H$$

Brute-Force Bayes Concept Learning

- What choice shall we make for P(D|h)?
- P(D|h) is the probability of observing the target values D = <d1 . . .dm> for the fixed set of instances <X1 . . . Xm>.
- Since we assume noise-free training data, the probability of observing classification di given h is just 1 if di = h(xi) and 0 if di != h(xi).
- Therefore,

$$P(D|h) = \begin{cases} 1 \text{ if } d_i = h(x_i) \text{ for all } d_i \text{ in } D \\ 0 \text{ otherwise} \end{cases}$$

- In other words, the probability of data D given hypothesis h is 1 if D is consistent with h, and 0 otherwise.

Brute-Force Bayes Concept Learning

- **Let us consider the** first step of this algorithm, which uses Bayes theorem to compute the posterior probability **P(h|D) of each hypothesis h given the observed training data D.**

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- First consider the case where h is inconsistent with the training data D.
- We know that P(D|h) to be 0 when h is inconsistent with D, we have,

$$P(h|D) = \frac{0 \cdot P(h)}{P(D)} = 0 \text{ if } h \text{ is inconsistent with } D$$

Brute-Force Bayes Concept Learning

- Now consider the case where **h is consistent** with D.
- We know that P(Dlh) to be 1 when h is consistent with D, we have

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$$P(h|D) = \frac{1 \cdot \frac{1}{|H|}}{P(D)}$$

$$= \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}}$$

$$= \frac{1}{|VS_{H,D}|} \text{ if } h \text{ is consistent with } D$$

Brute-Force Bayes Concept Learning

- To summarize, Bayes theorem implies that the posterior probability P(h|D) under our assumed P(h) and P(D|h) is,

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases}$$

Version Space in Machine Learning

Definition

A Version Space is the set of all hypotheses in the hypothesis space H that are consistent with the available training examples.

In simple words:

Version Space = All hypotheses that fit the data perfectly

Think of it like a filtering process:

- Start with all possible hypotheses in H
- Remove hypotheses that conflict with any training example
- The remaining ones form the Version Space

Suppose you want to learn the concept of "Edible Fruits."

Hypothesis space H contains:

- H1: All fruits are edible
- H2: All red fruits are edible
- H3: Only apples are edible
- H4: Anything sweet is edible
- H5: Nothing is edible

Training data:

- Apple → edible
- Strawberry → edible
- Banana → edible

Remove hypotheses that contradict data:

- H3: Only apples edible → ❌ fails (banana, strawberry not included)
- H5: Nothing edible → ❌ fails

Remaining consistent hypotheses:

- H1, H2, H4

These form the **Version Space**.

$$VS = \{H1, H2, H4\}$$

Version Space is the subset of hypotheses in the hypothesis space that are consistent with all training examples. It represents the current set of possible target concepts. Mitchell represented Version Space using the S (most specific) and G (most general) boundaries.

MAP Hypotheses and Consistent Learners

⭐ 1. MAP Hypothesis (Maximum A Posteriori Hypothesis)

Definition

A **MAP hypothesis** is the hypothesis **H** that has the **highest posterior probability** after observing data **D**.

$$H_{\text{MAP}} = \arg\max_{H} P(H|D)$$

Using Bayes' Theorem:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Since **P(D)** is the same for all H:

$$H_{\text{MAP}} = \arg\max_{H} \left(P(D|H) \cdot P(H)\right)$$

Interpretation

- Considers **how well the hypothesis explains the data** (likelihood)
- AND how **plausible** the hypothesis seemed before seeing data (prior)

Thus, MAP **balances prior belief** and **fit to data**.

⭐ Simple Example

Hypotheses to classify fruits as "edible":

- H1: All red fruits are edible → P(H1)=0.5
- H2: All fruits are edible → P(H2)=0.3
- H3: Only strawberries are edible → P(H3)=0.2

Training data: Apple (edible), Mango (edible), Banana (edible)

Compute posterior (intuitively):

- H1 fails (banana is not red) → lower likelihood
- H3 fails (apple and mango not included) → very low likelihood
- H2 matches all data → high likelihood

Even though H2 had lower prior than H1,

the posterior probability becomes highest.

👉 **MAP hypothesis = H2**

MAP Hypotheses and Consistent Learners

⭐ 2. Consistent Learners

Definition

A learner is **consistent** if it always returns a hypothesis that:

$$\text{Agrees with (fits) all training examples}$$

i.e., it **makes zero errors** on the data.

Key Idea

- Only hypotheses that **perfectly match the data** are considered
- Prior beliefs or probabilities **are not used**
- As long as the hypothesis is consistent, it is acceptable

Consistent learning is the basis of **Version Space** in ML.

⭐ Simple Example

You have training data:

Fruit	Edible?
Apple	Yes
Mango	Yes
Banana	Yes

A consistent learner might pick any hypothesis that fits all:

- H1: All fruits are edible
- H2: All yellow or red fruits are edible
- H3: Anything sweet is edible
- H4: Anything grown on trees is edible

All these **are consistent with the examples**, even if some are overly general.

Thus, consistent learners often return **multiple valid hypotheses**.

MAP Hypotheses and Consistent Learners

⭐ Relationship Between MAP Hypothesis & Consistent Learners

Aspect	Consistent Learner	MAP Hypothesis
Basis	Fits **all** training examples perfectly	Maximizes **posterior probability**
Uses Priors	❌ No	✔ Yes (P(H))
Handles Noise?	❌ Poorly	✔ Very well
Multiple hypotheses?	Yes (version space)	No (chooses the best one)
Robustness	Sensitive to contradictions	Robust; weighs data and prior
Approach	Logical	Probabilistic

Conclusion:

MAP hypothesis selection is **more flexible**, **probabilistic**, and often **more realistic** than strict consistency.

MAP Hypotheses and Consistent Learners

⭐ 3. When MAP = Consistent? (Special Case)

MAP hypothesis reduces to a **consistent learner** when:

1. **All hypotheses have equal prior probability**, and
2. **There is zero noise** in the data.

Then:

$$H_{\text{MAP}} = \arg\max_{H} P(D|H)$$

And if likelihood is 1 only for hypotheses consistent with data:

$$H_{\text{MAP}} = \text{any consistent hypothesis}$$

So consistent learners are a **special case** of MAP learners.

⭐ 4. Real-World Analogy (Easy for Students)

Suppose students are learning the rule for "What counts as a healthy snack".

Consistent Learner

Chooses any rule that fits examples exactly:

- "All fruits are healthy"
- "All green food is healthy"
- "Anything not fried is healthy"

(As long as it matches the examples given)

MAP Learner

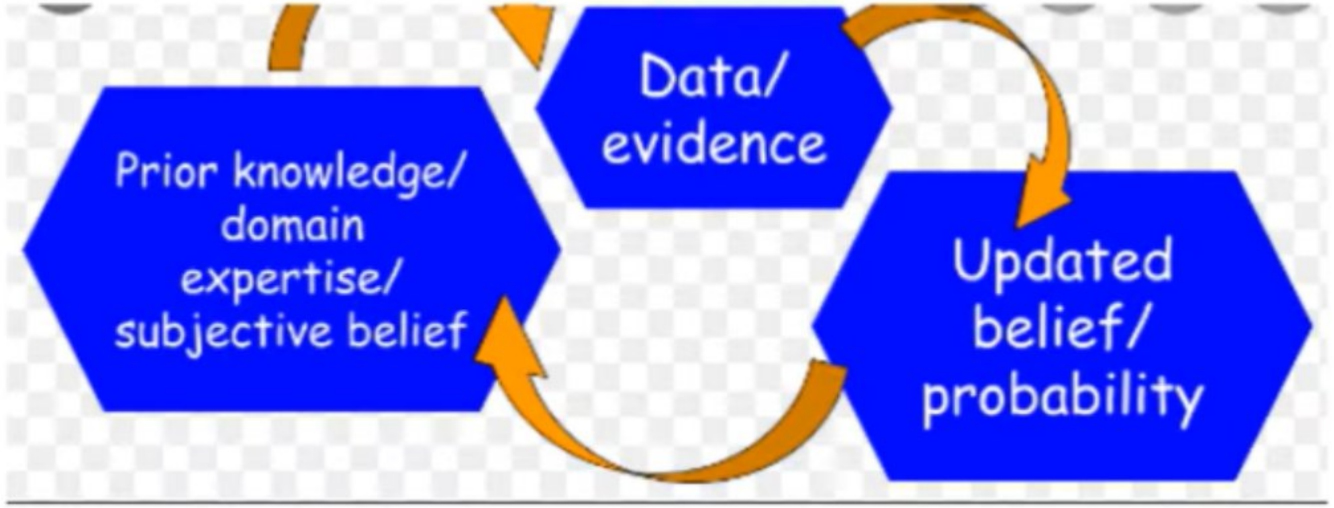
Chooses the **most reasonable** rule:

- Looks at prior belief (general nutritional knowledge)
- Looks at how well rule fits examples

MAP → "Fruits and vegetables are healthy"
which is logical and explains data.

Applications of Bayes Theorem

1. It is used to calculate the next step of the robot when the already executed step is given.
2. Bayes' theorem is helpful in weather forecasting.
3. Solving logical puzzles & games.



Example 1: Medical Diagnosis

The Scenario

01

Disease prevalence (prior)

Only 1% of the population has the disease

02

Test sensitivity

90% accuracy: P(positive test | disease present)

03

False positive rate

5% of healthy patients test positive incorrectly

The Surprising Result: Even with a positive test, the probability the patient actually has the disease is only approximately 15%. This counterintuitive result demonstrates why understanding Bayes' Theorem is critical in medical decision-making.

Example 2: Spam Email Filtering

Prior Probability

20% of incoming emails are spam

Word Evidence

Word "discount" appears in 70% of spam, but only 5% of legitimate emails

Bayesian Classification

Calculate P(spam | "discount") to accurately filter messages

Modern spam filters use Bayes' Theorem to analyze word patterns and classify emails with remarkable accuracy, learning continuously from new data to adapt to evolving spam tactics.

Example 3: Weather Forecasting

25%

Rain Frequency

Base rate of rainy days

15%

Cloud Frequency

Days with sunrise clouds

50%

Conditional

Example 4: DNA Evidence in Forensics

1

Low Prior

Initial probability suspect is guilty: approximately 1% based on circumstantial evidence

2

DNA Match

Probability of DNA match if guilty: 99.9% (highly sensitive test)

3

False Positive

Probability of DNA match if innocent: 0.1% (extremely rare)

4

Posterior Probability

Bayes' Theorem quantifies the dramatically increased likelihood of guilt given DNA evidence

Bayes' Theorem provides the mathematical framework that courts use to evaluate the strength of DNA evidence, helping juries understand how powerful genetic matching truly is in criminal cases.

Example 5: Tea Tasting Game

A Classic Concept Learning Problem

1

Prior Distribution

90% of teas are green, only 10% are white

2

Taster Accuracy

The taster correctly identifies each type 60% of the time

3

Observation

Taster declares the tea is "white"

4

Bayesian Result

True probability it's actually white: only ~14%

This counterintuitive result highlights why prior probabilities matter so much. Even with a confident identification, the rarity of white tea means it's still more likely to be green tea misidentified!

Real-World Applications of Bayes' Theorem

Spam Filtering

Classifying emails by analyzing word patterns and sender behaviors to protect inboxes worldwide

Medical Diagnosis

Updating disease likelihood with test results to guide treatment decisions and save lives

Financial Forecasting

Assessing investment risks and opportunities by incorporating new market data continuously

Engineering Diagnostics

Identifying root causes of system failures from observed symptoms and sensor data

Drug Testing

Evaluating treatment effectiveness in clinical trials by updating confidence with patient outcomes

Machine Learning

Powering AI algorithms that learn from data and improve predictions over time

Why Bayes' Theorem Matters

Better Decisions

Empowers clearer thinking and better decision-making under uncertainty

Evidence Integration

Bridges prior knowledge and new evidence for continuously refined predictions

AI Foundation

Provides the mathematical foundation for machine learning, AI, and scientific reasoning

Probabilistic Thinking

Encourages sophisticated probabilistic thinking in everyday life and complex problems

---

Start applying Bayes' Theorem today to make smarter, data-driven choices! Whether you're evaluating medical tests, filtering information, or making business decisions, this powerful tool will transform how you reason about uncertainty and evidence.