BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

ज्ञानं परमं बलम्

BITS Pilani

DIGITAL

Excellence made yours

BITS Pilani
DIGITAL
Excellence made yours

Probability and Statistics

Week 5 Live Session

Dr. Pushpendra Gupta

Recap

- Basic Probability: $P(A) = \text{Favorable}/\text{Total}$
- Conditional Probability: $P(A|B) = P(A \cap B) / P(B)$
- Multiplication Rule: $P(A \cap B) = P(A) \times P(B|A)$
- Introduction to Bayes' Theorem

Practical Examples:

- Fruit Basket Problem (Green Apple)
- Plant Watering Problem

Today's Focus

- 🧠 Coin Problem (From the Recorded Session)
- 🧠 Bayes Theorem
- 🎯 Mutually Exclusive vs Independent Events
- ⚖️ Naive Bayes Classification
- 💻 Hands-on Visualization using Google Colab

Coin Problem

Five independent flips of a fair coin are made. Find the probability that

a. The first three flips are the same.
b. Either the first three flips are the same, or the last three flips are the same.

Given Information
5 independent flips of a fair coin
P(Heads) = P(Tails) = 1/2 for each flip
Total possible outcomes = $2^5 = 32$

Part (a): Probability that the first three flips are the same
Step 1: Identify favorable outcomes
The first three flips are the same means either:
HHH** (first three are heads, last two can be anything), OR
TTT** (first three are tails, last two can be anything)

Coin Problem (Contd..)

BITS Pilani DIGITAL
Excellence made yours

a. The first three flips are the same.

Step 2: Count favorable outcomes

Case 1: HHH (First three flips are fixed as HHH)
Last two flips can be anything: HH, HT, TH, TT (4 possibilities)
Total: 4 outcomes

Case 2: TTT (First three flips are fixed as TTT)
Last two flips can be anything: HH, HT, TH, TT (4 possibilities)
Total: 4 outcomes
Total favorable outcomes = 4 + 4 = 8

Step 3: Calculate probability
P(first three same) = $\mathbf{8/32 =}$ **1/4 or 0.25**

Coin Problem (Contd..)

BITS Pilani DIGITAL
Excellence made yours

a. The first three flips are the same.

Alternative Method (using independence):

P(first three same) = P(HHH or TTT)

P(HHH) $= (1/2)^3 = 1/8$

P(TTT) $= (1/2)^3 = 1/8$

P(first three same) $= \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$

Coin Problem (Contd..)

BITS Pilani DIGITAL
Excellence made yours

Part (b): Probability that either the first three flips are the same OR the last three flips are the same

Step 1: Define events

A = first three flips are the same

B = last three flips are the same

We need: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Step 2: Find P(A)

From part (a): $P(A) = 1/4$

Step 3: Find P(B)

By symmetry (same logic as part a):

Last three are HHH: **HHH (4 possibilities for first two)

Last three are TTT: **TTT (4 possibilities for first two)

$$P(B) = 8/32 = 1/4$$

Coin Problem (Contd..)

BITS Pilani DIGITAL
Excellence made yours

Part (b): Probability that either the first three flips are the same OR the last three flips are the same

Step 4: Find P(A ∩ B) Both first three and last three are the same means:
Flips 1, 2, **3** are the same
Flips **3**, 4, 5 are the same
This means ALL five flips must be the same!
Favorable outcomes: HHHHH or TTTTT (2 outcomes)

$$P(A \cap B) = 2/32 = 1/16$$

Step 5: Calculate P(A ∪ B)
P(A ∪ B) = P(A) + P(B) - P(A ∩ B)
P(A ∪ B) = 1/4 + 1/4 - 1/16
P(A ∪ B) = 4/16 + 4/16 - 1/16
P(A ∪ B) = **7/16** or **0.4375**

Bayes Theorem

$$P(A|B) = \frac{[P(B|A) \times\ P(A)]}{P(B)}$$

Thomas Bayes

Or in expanded form:

$$P(A|B)\ =\ [P(B|A)\ \times\ P(A)]\,/\,[P(B|A)\ \times\ P(A)\ +\ P(B|A')\ \times\ P(A')]$$

Named Components:

$P(A|B)$ = Posterior (what we want)
$P(B|A)$ = Likelihood (what we know)
$P(A)$ = Prior (base rate)
$P(B)$ = Marginal Probability (normalizing constant)

Bayes Theorem Applications

We often know P(B|A) but need P(A|B)
Example: We know $P(Symptoms|Disease)$ but need $P(Disease|Symptoms)$
We know P(Word | Spam) but need P(Spam | Word)

Key Insight: Bayes' Theorem lets us "flip" conditional probabilities

Applications:

- Medical Diagnosis
- Spam Detection
- Fraud Detection
- Machine Learning
- Criminal Justice
- Scientific Research

Medical Testing Example

Example: Medical Test for Rare Disease

Given:

P(Disease) = 0.01 (1% of population has it)
P(Positive Test | Disease) = 0.95 (95% accurate)
P(Positive Test | No Disease) = 0.05 (5% false positive)
Question: If someone tests positive, what's the probability they actually have the disease?

Most people guess: 95% Actual answer: We'll calculate!

Medical Testing Calculation

Calculating P(Disease | Positive Test)

Using Bayes' Theorem:

$$P(D|+) = [P(+|D) \times P(D)] / P(+)$$

Step 1: Calculate P(+)

$$P(+) = P(+|D) \times P(D) + P(+|No\ D) \times P(No\ D)$$

$$P(+) = 0.95 \times 0.01 + 0.05 \times 0.99$$

$$P(+) = 0.0095 + 0.0495 = 0.059$$

Step 2: Apply Bayes: $P(D|+) = (0.95 \times 0.01) / 0.059$

$$P(D|+) = 0.0095 / 0.059 = 0.161\ or\ 16.1\%$$

Only 16% chance of actually having the disease!

Visualizing the Medical Test

Visualizing with 10,000 People (Population of 10,000):
Population breakdown:
People WITH disease: 1% of 10,000 = 100 people
People WITHOUT disease: 99% of 10,000 = 9,900 people

Now let's see who tests positive:
From the diseased group (100 people):
95% test positive = 0.95 × 100 = 95 TRUE positives ✓
From the healthy group (9,900 people):
5% test positive = 0.05 × 9,900 = 495 FALSE positives ✗

Total people who test positive = 95 + 495 = 590 people
Out of these 590 positive tests: Only 95 actually have the disease
495 are false alarms   So: 95/590 = 0.161 = 16.1%

Why does this happen?

Even though the false positive rate is small (5%), it's being applied to a HUGE group (9,900 healthy people).

Meanwhile, the true positive rate (95%) is being applied to a tiny group (100 diseased people).

$$5\% \text{ of } 9{,}900 \;>>\; 95\% \text{ of } 100$$

That's why most positive tests are false positives - there are simply way more healthy people being tested!

Test's accuracy (95%): "How good is the test at catching sick people?“
Out of 100 sick people → 95 will test positive
Positive predictive value (16.1%): "If I test positive, what are my chances?“
Out of 590 positive tests → only 95 actually have it

Mutually Exclusive Events

Definition: Events that **CANNOT** occur at the same time

Mathematical Property:

$$P(A \cap B) = 0$$
$$P(A \text{ or } B) = P(A) + P(B)$$

Examples:

Rolling a die: Getting 2 AND getting 5
Flipping a coin: Getting heads AND getting tails
Card draw: Drawing King of Hearts AND Queen of Spades (same draw)

Independent Events

Definition: Events where one occurrence does NOT affect the probability of the other

Mathematical Property:

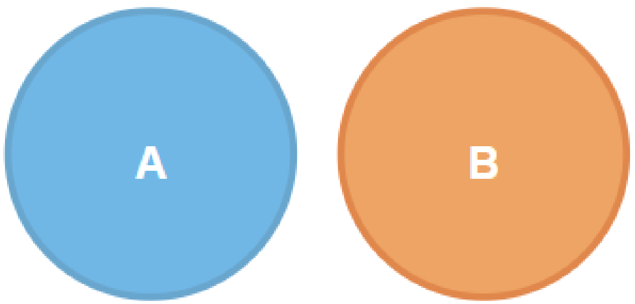
$$P(A \cap B) = P(A) \times P(B)$$
$$P(A \mid B) = P(A)$$

Examples:

Flipping a coin twice: Result of first flip doesn't affect second
Rolling two dice: One die doesn't affect the other
Drawing cards WITH replacement

Mutually Exclusive

Cannot occur at the same time



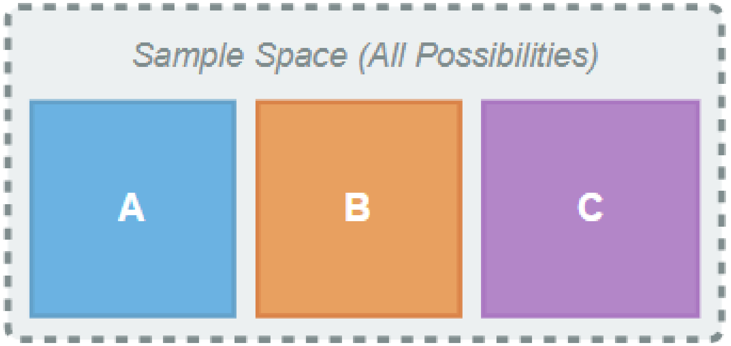
Example:

Coin flip: {Heads} and {Tails}

Cannot get both at once

Exhaustive Events

Cover all possible outcomes

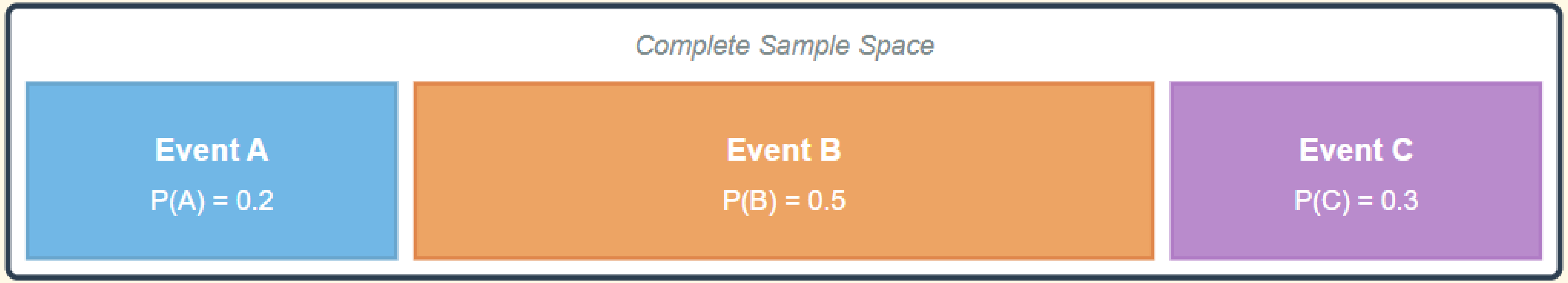


Example:

Die roll: {1,2} {3,4} {5,6}

One of these MUST occur

Mutually Exclusive AND Exhaustive (Partition)



Law of Total Probability

Law of Total Probability

If $B_1, B_2, \ldots, B_{\mathrm{n}}$ are mutually exclusive and exhaustive events:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_{\mathrm{n}})P(B_{\mathrm{n}})$$

Binary Case (most common): $P(A) = P(A|B)P(B) + P(A|B')P(B')$

Purpose: Calculate total probability by considering all possible "pathways" to A

Used in: Denominator of Bayes' Theorem!

Example - Weather and Tennis

Example: Will They Play Tennis?

Given Data (from past 14 days):

Sunny: 5 days (played: 3, didn't play: 2)
Overcast: 4 days (played: 4, didn't play: 0)
Rainy: 5 days (played: 2, didn't play: 3)
Calculate:

P(Play) $= 9/14 \approx 0.643$
P(No Play) $= 5/14 \approx 0.357$

Question: If weather is Sunny, what's $P(Play|Sunny)$?

on sunny days they played 3 out of 5 times, so: P(Play|Sunny) = 3/5 = 0.6 or 60%

Multi-Feature Example

Day	Outlook	Temp	Humidity	Windy	Play Tennis
1	Sunny	Hot	High	False	NO
2	Sunny	Hot	High	True	NO
3	Overcast	Hot	High	False	YES
4	Rainy	Mild	High	False	YES
5	Rainy	Cool	Normal	False	YES

Today's Conditions: Outlook = Sunny, Temp = Cool, Humidity = High, Windy = True

Question: Will they play Tennis?

To answer this properly, we need to find: P(Play | Sunny, Cool, High, True)

The Naïve Bayes Assumption

The Problem: P(Play | Outlook, Temp, Humidity, Windy) is hard to estimate

The Solution: Assume features are conditionally independent given the class

Mathematical Expression: $P(X_1, X_2, \ldots, X_\mathrm{n} \mid Y) = P(X_1|Y) \times P(X_2|Y) \times \ldots \times P(X_\mathrm{n}|Y)$

Why "Naive"? Real features often ARE related, but we assume they're not!

Surprising Fact: Works well in practice despite the assumption!

P(Outlook, Temp, Humidity, Windy | Play) can be broken down into:
P(Outlook|Play) × P(Temp|Play) × P(Humidity|Play) × P(Windy|Play)

Calculating Probabilities from Data

Step 1: Calculate Priors: P(Play = Yes) = 9/14 and P(Play = No) = 5/14
Step 2: Calculate Likelihoods (Example) From 14 days of data:

$P(Outlook = Sunny \mid Play = Yes) \; = \; 2/9$
$P(Outlook = Sunny \mid Play = No) \; = \; 3/5$
$P(Temp = Cool \mid Play = Yes) \; = \; 3/9$
$P(Temp = Cool \mid Play = No) \; = \; 1/5$
(Similarly for Humidity and Windy...)

Example Calculation - Play Tennis?

Today: Sunny, Cool, High Humidity, Windy

For $P(Yes \mid Today)$:

$$P(Yes) \times P(Sunny|Yes) \times P(Cool|Yes) \times P(High|Yes) \times P(Windy|Yes)$$
$$= (9/14) \times (2/9) \times (3/9) \times (3/9) \times (3/9) = 0.0053$$

For $P(No \mid Today)$:

$$P(No) \times P(Sunny|No) \times P(Cool|No) \times P(High|No) \times P(Windy|No)$$
$$= (5/14) \times (3/5) \times (1/5) \times (4/5) \times (3/5) = 0.0206$$

$$P(No \mid Today) > P(Yes \mid Today)$$

Prediction: Will NOT play tennis

Why Multiply Probabilities?

Understanding the Multiplication

Each probability is "evidence":

$P(Sunny|No) = 0.6$ → Sunny is moderate evidence for No
$P(High|No) = 0.8$ → High humidity is strong evidence for No
$P(Cool|No) = 0.2$ → Cool is weak evidence for No

Multiplying combines the evidence:
Strong evidence × Strong evidence = Very strong conclusion
Strong evidence × Weak evidence = Moderate conclusion

Logarithms in practice:
Often use log probabilities: log(a × b) = log(a) + log(b)
Avoids numerical underflow!

Advantages of Naive Bayes

✓ Simple and fast to train
✓ Works well with limited data
✓ Handles high-dimensional data well
✓ No complex parameter tuning needed
✓ Naturally handles multi-class classification
✓ Provides probabilistic predictions
✓ Scales well to large datasets

Extra Slide Added for Explaining the Advantages

Simple and Fast: Training is just counting. No iterative optimization, no gradient descent, no hyperparameters to tune. Classification is just multiplication (or addition if using log probabilities). This makes it one of the fastest machine learning algorithms.

Works with Limited Data: Because each feature is considered independently, we don't need to see every combination of features. Even with just a few examples, we can estimate the individual probabilities. Other algorithms might need much more data.

Handles High Dimensions: Text classification might have 10,000+ different words (features), but Naive Bayes handles this easily. Each word's probability is estimated independently. Contrast this with algorithms that need to consider feature interactions - they'd need exponentially more data.

Extra Slide Added for Explaining the Advantages (Contd..)

No Tuning Required: Unlike neural networks or Support Vector Machines, there are very few parameters to tune. The main choice is whether to use Laplace smoothing and what smoothing parameter to use. That's it!

Multi-class: Need to classify into more than two categories? Naive Bayes handles this naturally. Just calculate probabilities for each class and pick the highest. No need to train multiple binary classifiers.

Probabilistic: Unlike algorithms that just give you a classification, Naive Bayes gives you probabilities. You can see not just that an email is spam, but that it's 95% likely to be spam. This helps you set thresholds and understand confidence.

Scalable: Training can be done incrementally. As new data arrives, just update your counts. This makes it perfect for online learning.

It's particularly excellent for text classification. If you're building a spam filter, sentiment analyzer, or document categorizer, Naive Bayes should be one of your first choices. It's the baseline that other algorithms have to beat.

Limitations of Naive Bayes

Limitations and When Not to Use

Disadvantages:

✗ Strong independence assumption (often violated)
✗ Can't learn feature interactions
✗ Sensitive to irrelevant features
✗ Not great for regression tasks
✗ Probability estimates can be extreme

Poor Choice For:

Tasks where feature relationships matter
Small number of highly correlated features
Need for precise probability estimates

Extra Slide Added to Explain the Limitations

No algorithm is perfect, and Naive Bayes has significant limitations you should know about.

Independence Assumption: This is the big one. The algorithm assumes features are independent given the class, but this is often false. In spam detection, if you see "free", you're more likely to also see "offer" or "prize". In medical diagnosis, symptoms are often related. Naive Bayes ignores these relationships.

Can't Learn Interactions: Because it treats features independently, it can't learn patterns like "if feature A is high AND feature B is low, then class X." It only learns "feature A is high → class X" and "feature B is low → class X" separately. Sometimes the interaction is what matters.

Sensitive to Irrelevant Features: If you have many features that are actually irrelevant to the classification task, they add noise to the probability estimates. Other algorithms can learn to ignore irrelevant features, but Naive Bayes treats all features equally.

Extra Slide Added to Explain the Limitations (Contd..)

BITS Pilani DIGITAL
Excellence made yours

Not for Regression: Naive Bayes is a classification algorithm. If you need to predict a continuous value (like house price or temperature), you need a different approach.

Extreme Probabilities: Because we multiply many probabilities, the final values can be very close to 0 or 1. The relative ranking is usually correct (spam vs ham), but the actual probability estimates can be overconfident. If you need well-calibrated probabilities, you might need to apply calibration techniques.

When to avoid it:

- If you have just a few features that are highly correlated, other algorithms might work better
- If understanding feature interactions is crucial to the task
- If you need precise probability estimates for decision-making (though the classifications themselves are often accurate)

Despite these limitations, Naive Bayes is still incredibly useful. For many tasks, especially text classification, the advantages outweigh the disadvantages. It's often the first algorithm to try, and frequently it's good enough that you don't need anything more complex.

Naive Bayes in Practice

Real-World Applications

Email Spam Filtering (Gmail, Outlook)

Each word is a feature
98%+ accuracy achievable

Sentiment Analysis

Positive, Negative, Neutral reviews
Twitter sentiment tracking

Document Classification

News article categorization
Medical record classification

More Real-World Applications (Contd...)

Recommendation Systems
User preference prediction
Content filtering

Real-time Prediction

Fast enough for online systems
Fraud detection

Thank You