Mathematics and Statistics for Artificial Intelligence

Introduction

Welcome to the Second Module of our AI course! Today we'll explore the essential mathematical and statistical concepts that power artificial intelligence systems. Don't worry if you don't have a strong background in these areas - we'll explain everything in simple terms and build your understanding step by step.

By the end of this session, you'll understand the key mathematical tools that make AI work and why they matter. Let's get started!

1. Why Mathematics Matters in AI

AI systems might seem magical, but they're built on mathematical foundations. Here's why math is crucial:

- **Mathematics provides the language** to express relationships between data
- **Algorithms need math** to process information and make decisions
- **Statistics helps us understand uncertainty** and make **predictions**
- **Optimization techniques** help AI systems learn and improve

AI in action: Consider these real-world applications:

- Image recognition uses linear algebra to process pixels
- Speech recognition uses probability to handle unclear words
- Recommendation systems use distance metrics to find similar items
- Self-driving cars use calculus to make continuous adjustments

💡 **Interactive Question**: Why do you think we need mathematics in AI? Think about this before continuing.

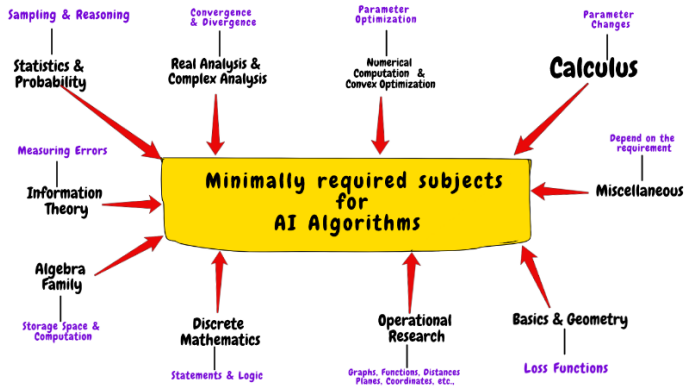


Image Source: https://medium.com/swlh/ai-mathematics-699a9ea2a0d6

2. Essential Linear Algebra Concepts

2.1 Vectors and Matrices

Vectors are **ordered lists** of numbers that can represent:

- **Features** of data (like height, weight, color)
- Positions in space
- Directions and magnitude

Let's start with the basics:

A vector with 3 elements might look like [5, 2, 8].

Each number has a position and meaning. In AI, each number might represent a different feature of something we're analyzing.

Some more examples of vectors are

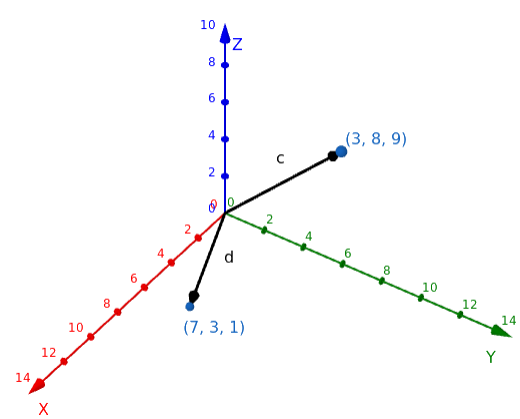
$$\vec{a} = \begin{bmatrix} 1 \\ 33 \\ 445 \\ 89 \end{bmatrix}$$

$$\vec{b} = \begin{bmatrix} 4 & 47 & 89 \end{bmatrix}$$

a is a **column vector** and **b** is **row vector**

$$c = \begin{bmatrix} 3 & 8 & 9 \end{bmatrix}$$

$$d = \begin{bmatrix} 7 \\ 3 \\ 1 \end{bmatrix}$$



c and d represented on **graph**

Image Source: https://medium.com/linear-algebra/part-12-vectors-a99364499121

Matrices are **rectangular arrays** of numbers that can represent:

- Collections of **data points**
- Transformations of data
- Relationships between entities

Think of a matrix as a table of numbers arranged in rows and columns.

Just like a spreadsheet, each position has meaning related to its row and column.

Example: If we have 3 students and their test scores in 2 subjects:

Student	Math	Science
Alice	90	85
Bob	75	80
Charlie	85	90

We can represent this as a 3×2 matrix:

[90 85]

[75 80]

[85 90]

💡 **Interactive Question**: What might each row and column represent in this matrix?

2.2 Matrix Operations

Key operations include:

- **Addition and subtraction**: Combining or finding differences between datasets
- **Multiplication**: Transforming data or **computing relationships**
- **Transposition**: Switching rows and columns
- **Inverse**: Finding the "opposite" transformation

These operations allow AI systems to transform data, calculate distances, and **find patterns**.

Multiple Choice Question #1:

Which matrix operation would you use to switch rows and columns?

A) Addition

B) Multiplication

C) Transposition

D) Inverse

2.3 Distance Metrics

When working with vectors in AI, we often need to measure "**how far apart**" two data points are. Here are common ways to do this:

- **Euclidean Distance**: The **straight-line distance** between two points
  - Formula: square root of the sum of squared differences in each dimension

$$\sqrt{\{(x_2 - x_1)^2 + (y_2 - y_1)^2\}}$$

  - Example: This is what we typically think of as "distance" in everyday life
- **Manhattan Distance**: The **sum of absolute differences** in each dimension
  - Formula: $|x_1 - x_2| + |y_1 - y_2|$
  - Example: Think of it as the distance a taxi would drive in a city with a grid layout (also called "L1 norm" or "taxicab distance").
- **Cosine Similarity**: Measures the **angle between two vectors**
  - Ranges from -1 (pointing in opposite directions) to 1 (pointing in same direction)
  - Used when direction matters more than magnitude
  - Popular for text analysis and recommendation systems

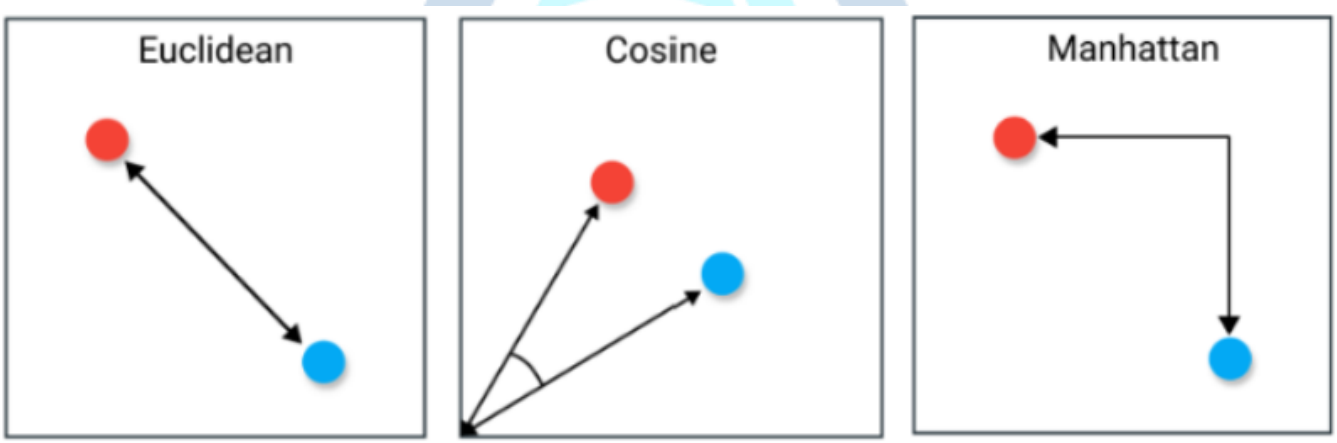


Image Source: https://www.teradata.com/insights/ai-and-machine-learning/large-language-models-meet-teradata-vantage

3. Calculus Fundamentals

3.1 Derivatives and Gradients

Derivatives measure how a **function changes** when its **input changes**. In AI, we use derivatives to:

- Find the direction of steepest improvement
- Adjust model parameters to **minimize errors**
- Understand how changes in input affect the output

Ok - Let's break down **derivatives** in the **simplest way possible.**

A **derivative** tells you **how fast something is changing**.

- Imagine you're driving a car.
- The **speed** on your dashboard is the **derivative** of your **position** over time.
- If you're at 0 km, then 1 km, then 2 km — you're moving!
- The **rate of change** of your position = **your speed** = **derivative**.

Simple Example: A Straight Line

Function: Let's say: **f(x)=2x**. This is a line. Every time x goes up by 1, f(x) goes up by 2.

Derivative: The derivative of this function is: **f'(x)=2**.

Which means: "No matter what x is, the slope (rate of change) is always 2."

Slightly More Interesting: A Curve

Function: $\boldsymbol{f(x) = x2}$ **-** Now the graph is a curve — it changes shape.

At x = 1, f(x) = 1

At x = 2, f(x) = 4

At x = 3, f(x) = 9

You can see the output is changing faster as x increases.

Derivative: **f'(x)=2x -** Now the rate of change depends on x**.**

At x = 1, slope is 2

At x = 2, slope is 4

At x = 3, slope is 6

It's like your car is **accelerating** — the speed (rate of change) increases as time goes on.

Gradients extend this idea to functions with **multiple inputs**. They point in the **direction of greatest increase**.

A gradient is a collection of partial derivatives (one for each input variable).

In Simpler way, A **gradient** is just the **multi-dimensional version** of a derivative.

- A **derivative** is used for functions with **one input (x)**.
- A **gradient** is used for functions with **two or more inputs** — like **x and y**, or **x, y, z**, etc.

One-variable (Derivative):

$$f(x) = x2 \Rightarrow f'(x) = 2x$$

This tells you how **f changes when x changes**.

Two-variable (Gradient):

$$f(x, y) = x2 + y2$$

Now we want to know:

- How does f change when x changes?
- How does f change when y changes?

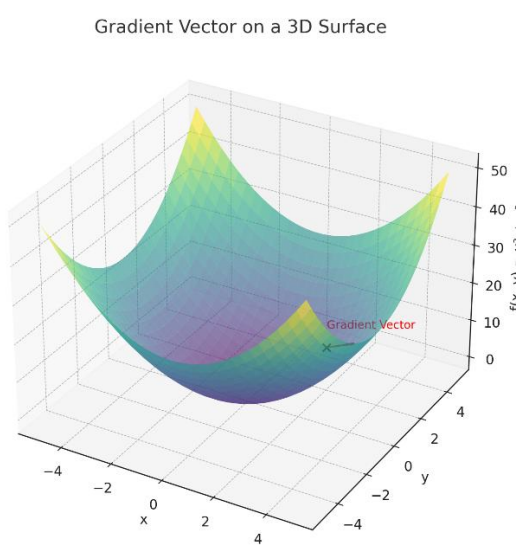
We calculate partial derivatives for each variable.

$$\frac{\partial f}{\partial x} = 2x, \quad \frac{\partial f}{\partial y} = 2y$$

Together, they form the **gradient**:

$$\nabla f = [2x,\ 2y]$$

This is a **vector** that points in the direction of the **steepest increase** of the function.



3.2 Optimization with Calculus

AI systems learn by optimizing – finding the best parameters to make accurate predictions. This involves:

- Defining a **loss function** that **measures error**
- Using **gradient descent** to find the **minimum** of this function
- Taking small steps in the direction that **reduces error**

Gradient Descent Variations

- **Batch Gradient Descent**: Calculates the gradient using the entire dataset
  - **Most accurate** but can be **slow** with large datasets
- **Stochastic Gradient Descent (SGD)**: Updates parameters **after** each training example
  - Faster but more erratic path to the minimum
  - May help escape local minima due to its "noisier" updates
- **Mini-batch Gradient Descent**: Compromise between batch and stochastic
  - Updates parameters after a small batch of examples (e.g., 32 or 64)
  - Combines the stability of batch with the speed of stochastic

When training neural networks, adaptive learning rate methods like Adam, RMSprop, and Adagrad can help reach the minimum faster and more reliably.

Multiple Choice Question #2: In gradient descent, we move in which direction?

A) The direction where the function increases fastest

B) The direction where the function decreases fastest

C) Random directions

D) Perpendicular to the gradient

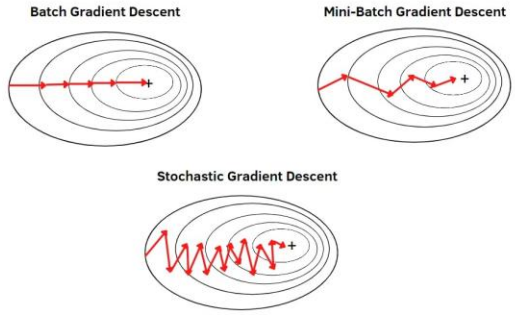


Image Source: https://www.analyticsvidhya.com/blog/2022/07/gradient-descent-and-its-types/

4. Probability and Statistics

4.1 Probability Basics

Probability measures the **likelihood of events**. In AI, we use probability to:

- Handle **uncertainty** in data and predictions
- Make **decisions** with incomplete information
- Estimate confidence in results

Probability is always between 0 and 1:

- 0 means an event is impossible
- 1 means an event is certain
- 0.5 means an event has a 50% chance of happening

Key concepts include:

- **Probability distributions**: Ways to describe how likely different outcomes are
- **Conditional probability**: The chance of one event given another has occurred (written as P(A|B) - "probability of A given B")
- **Bayes' theorem**: A way to update beliefs based on new evidence

Bayes' theorem is especially important in AI and looks like this: P(A|B) = P(B|A) × P(A) / P(B)

This lets us update what we believe (like whether an email is spam) based on new evidence (like what words appear in the email).

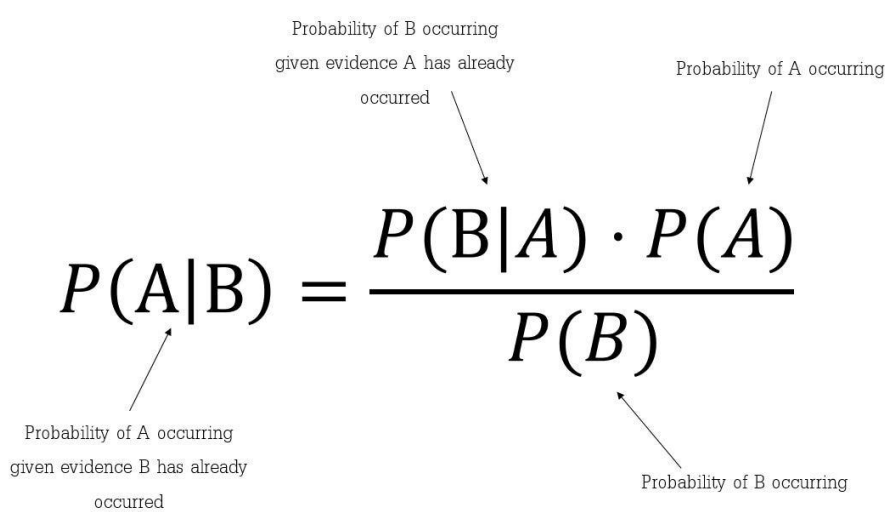


Image Source: https://iamkartikeya.medium.com/machine-learning-bayes-theorem-2f48c33d51e5

4.2 Statistical Measures

These help us understand our data:

- **Mean**: The **average** value (sum all values and divide by the count)
  - Example: Mean of [2, 3, 4, 5, 6] = (2+3+4+5+6)/5 = 20/5 = 4
  - **Sensitive to outliers** (extreme values)
- **Median**: The **middle value** when sorted
  - Example: Median of [2, 3, 4, 5, 6] = 4
  - Example: Median of [2, 3, 4, 5, 6, 100] = (4+5)/2 = 4.5
  - **Less sensitive to outliers** than the mean
- **Mode**: The **most common** value
  - Example: Mode of [2, 2, 3, 4, 5] = 2
  - Can have **multiple modes** or **no mode**

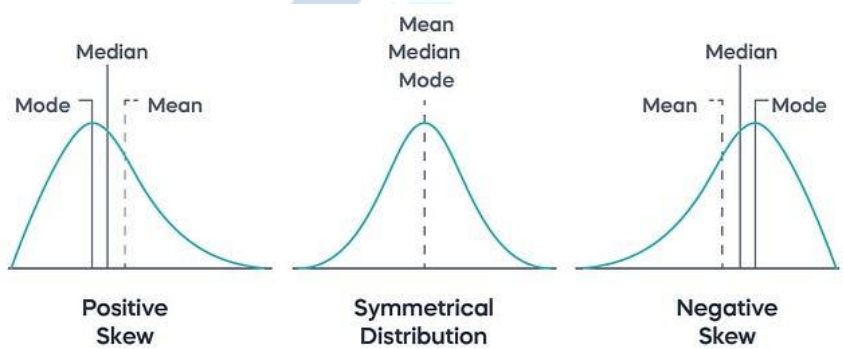


Image Source: https://medium.com/@nitesh.py/understanding-measures-of-central-tendency-mean-median-and-mode-cabb73175b29

- **Variance and Standard Deviation**: Measures of **how spread** out the data is
  - **Variance**: **Average of squared differences from the mean**
  - **Standard Deviation**: **Square root of variance** (more intuitive as it's in the same units as the data)
  - Small values mean data points are close to the mean; large values mean they're widely spread

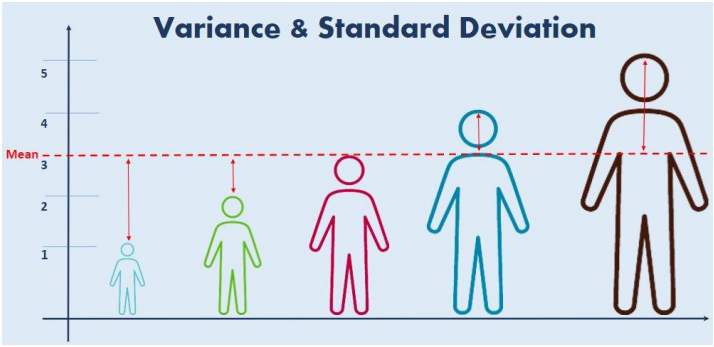


Image Source: https://www.shiksha.com/online-courses/articles/measures-of-dispersion-range-iqr-variance-standard-deviation/

- **Covariance**: **Measures** how **two variables change together**
    - **Positive covariance**: variables tend to increase or decrease together
    - **Negative covariance**: when one increases, the other tends to decrease
    - **Problem**: Scale-dependent, making it hard to interpret the strength
- **Correlation**: How **strongly variables** are **related to each other**
    - Ranges from -1 to 1
    - 1 means perfect positive correlation (when one goes up, the other goes up)
    - -1 means perfect negative correlation (when one goes up, the other goes down)
    - 0 means no correlation
    - Essentially normalized covariance, making it easier to interpret
    - **Important**: **Correlation does not imply causation**!

Multiple Choice Question #3: If a dataset has a high standard deviation, what does this tell us?

A) The data points are mostly close to the mean

B) The data points are widely spread out from the mean

C) The data has a normal distribution

D) The mean is equal to the median

These mathematical foundations enable AI systems to learn from data, make predictions, classify information, and generate new content.

Glossary

- **Correlation**: A measure of the relationship between two variables
- **Derivative**: The rate of change of a function with respect to one of its variables
- **Gradient**: Vector of partial derivatives that points in the direction of steepest increase
- **Gradient Descent**: An optimization algorithm for finding the minimum of a function
- **Loss Function**: A function that measures the difference between predicted and actual values
- **Matrix**: A rectangular array of numbers, symbols, or expressions
- **Mean**: The average of a set of values (sum divided by count)
- **Median**: The middle value in a sorted list of numbers
- **Normal Distribution**: A probability distribution characterized by a bell-shaped curve
- **Probability**: A measure of the likelihood of an event occurring
- **Standard Deviation**: A measure of the amount of variation or dispersion in a set of values
- **Variance**: A measure of how far a set of numbers is spread out from their average value
- **Vector**: A quantity with both magnitude and direction

Multiple Choice Answers

1. C) Transposition
2. B) The direction where the function decreases fastest
3. B) The data points are widely spread out from the mean

Congratulations on completing the Mathematics and Statistics for AI module! These concepts will serve as building blocks for understanding the algorithms we'll explore in the upcoming modules. In our next module, we'll dive into machine learning algorithms that build on these mathematical foundations.

AMEYA