Tutorial 0: Mathematical Prerequisites for Machine Learning

Scalars, Vectors, Matrices, Probability, and Statistics

ES335 - Machine Learning
IIT Gandhinagar

July 22, 2025

Abstract

This tutorial covers essential mathematical concepts that are fundamental to understanding machine learning algorithms. We start with basic building blocks (scalars, vectors, matrices) and progress to more advanced topics (norms, rank, probability, statistics). Each concept is explained with intuitive examples from real-world scenarios, followed by both simple and challenging exercises to test your understanding.

Contents

1 Introduction: Why These Math Concepts Matter

Imagine you're developing a recommendation system for Netflix. You have:

- **Scalars**: A single user's rating (4.2 stars) for a movie
- **Vectors**: All ratings by one user [4.2, 3.1, 5.0, 2.8, ...]
- **Matrices**: Ratings by all users for all movies (millions $\times$ millions grid)
- **Norms**: How "similar" are two users' preferences?
- **Probability**: What's the chance this user will like a sci-fi movie?
- **Statistics**: What's the average rating for action movies?

Understanding these concepts deeply will help you build better ML models and debug them when things go wrong.

2 Scalars: The Building Blocks

2.1 What Are Scalars?

A **scalar** is just a single number. In machine learning contexts:

- Learning rate: $\alpha = 0.001$
- Temperature: $T = 23.5°C$
- Accuracy: $acc = 0.94$
- Loss value: $L = 2.317$

Example #1: House Price Prediction

You're predicting house prices. Here are some scalars in your model:

- House size: 1,850 square feet
- Number of bedrooms: 3
- Predicted price: $425,000
- Model's confidence: 0.87

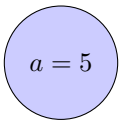
Each of these is a scalar because it's a single number representing one quantity.

2.2 Scalar Operations

When working with scalars, you can perform basic arithmetic:

$$a + b = 5 + 3 = 8 \tag{1}$$
$$a \times b = 5 \times 3 = 15 \tag{2}$$
$$a^b = 5^3 = 125 \tag{3}$$
$$\sqrt{a} = \sqrt{25} = 5 \tag{4}$$

Scalar



Single number

3 Vectors: Lists of Related Numbers

3.1 Understanding Vectors Intuitively

A **vector** is an ordered list of numbers. Think of it as:

- Your location in GPS coordinates: [latitude, longitude]
- Stock prices: [AAPL: \$150, GOOGL: \$2800, MSFT: \$300]
- Image pixels: [red_value, green_value, blue_value] for each pixel

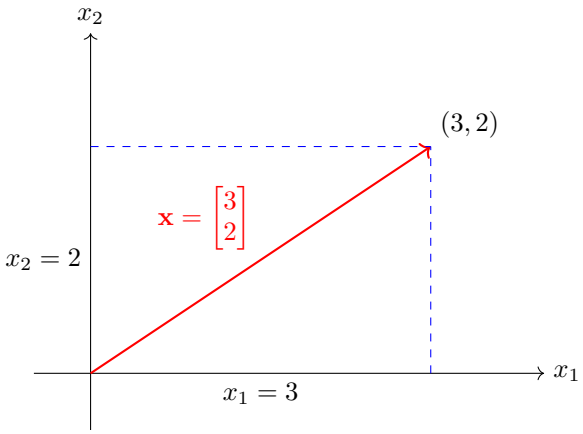
Example #2: Student Performance Vector

A student's performance across different subjects:

$$\mathbf{x} = \begin{bmatrix} 85 \\ 92 \\ 78 \\ 96 \\ 88 \end{bmatrix} \leftarrow \begin{matrix} \text{Math} \\ \text{Physics} \\ \text{Chemistry} \\ \text{Computer Science} \\ \text{English} \end{matrix}$$

This vector $\mathbf{x} \in \mathbb{R}^5$ captures the student's complete academic profile in one mathematical object.

3.2 Vector Visualization



3.3 Common Vector Operations

1. Vector Addition (Element-wise)

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}$$

2. Scalar Multiplication

$$3 \cdot \mathbf{x} = 3 \cdot \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 9 \end{bmatrix}$$

3. Dot Product (Inner Product)

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

Example #3: Movie Recommendation Dot Product

User A's movie ratings: $\mathbf{x} = [5, 3, 4, 2, 5]$ (Action, Comedy, Drama, Horror, Sci-Fi)
User B's movie ratings: $\mathbf{y} = [4, 4, 5, 1, 4]$
Similarity (dot product): $\mathbf{x} \cdot \mathbf{y} = 5 \times 4 + 3 \times 4 + 4 \times 5 + 2 \times 1 + 5 \times 4 = 74$
Higher dot product suggests similar tastes!

4 Matrices: Tables of Numbers

4.1 Matrix Intuition

A **matrix** is a rectangular array of numbers. Think of it as:

- Spreadsheet with rows and columns
- Collection of vectors stacked together
- Transformation that changes vectors

	Col 1	Col 2	Col 3
Row 1	a_{11}	a_{12}	a_{13}
Row 2	a_{21}	a_{22}	a_{23}

$$\mathbf{A} \in \mathbb{R}^{2 \times 3}$$

Example #4: Student Grades Matrix

Grades for 3 students across 4 subjects:

$$\mathbf{X} = \begin{bmatrix} 85 & 92 & 78 & 96 \\ 90 & 88 & 85 & 94 \\ 82 & 95 & 80 & 91 \end{bmatrix} \begin{matrix} \text{Alice} \\ \leftarrow \text{Bob} \\ \text{Carol} \end{matrix}$$

$$\uparrow \quad \uparrow \quad \uparrow \quad \uparrow$$

Math Physics Chem CS

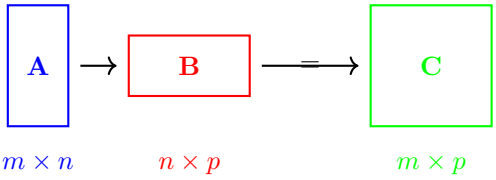
Each row is a student's performance vector. Each column is all students' performance in one subject.

4.2 Matrix Operations

Matrix Multiplication (Most Important!)

$$\mathbf{AB} = \mathbf{C} \text{ where } c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$$

Key Rule: $(m \times n) \times (n \times p) = (m \times p)$



5 Vector and Matrix Norms

5.1 What Are Norms?

A **norm** measures the "size" or "length" of a vector/matrix. Think of it as:

- Distance from origin to a point
- Magnitude of a force vector
- "How big" is this mathematical object?

5.2 Common Vector Norms

1. L1 Norm (Manhattan Distance)

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

Like walking in Manhattan - you can only go up/down, left/right.

2. L2 Norm (Euclidean Distance)

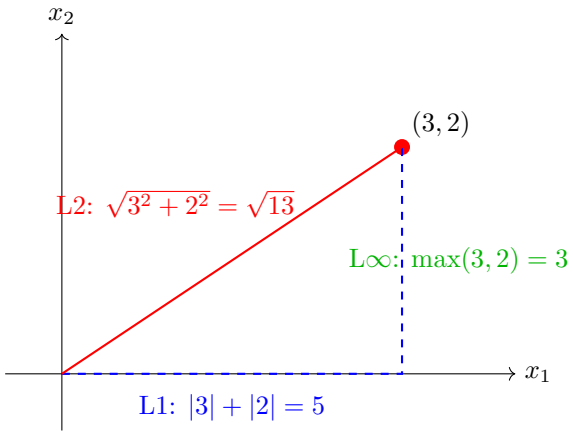
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Straight-line distance "as the crow flies."

3. L∞ Norm (Maximum Norm)

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

Just the largest component.



Example #5: Error Vector Norms

Your model's prediction errors on 4 test samples: errors $= [2, -1, 3, -4]$

- L1 norm: $\|errors\|_1 = 2 + 1 + 3 + 4 = 10$ (total absolute error)
- L2 norm: $\|errors\|_2 = \sqrt{4 + 1 + 9 + 16} = \sqrt{30}$ (root mean square error)
- L∞ norm: $\|errors\|_\infty = 4$ (worst single error)

Different norms emphasize different aspects of your model's performance!

6 Matrix Rank and Linear Independence

6.1 Understanding Rank Intuitively

The **rank** of a matrix tells you:

- How many "truly different" rows/columns it has
- The dimension of the space it spans
- Whether the matrix is invertible

Example #6: Student Survey Matrix

Survey responses from 3 students on 3 questions (1-5 scale):

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & 3 \\ 4 & 3 & 2 \\ 3 & 2 & 1 \end{bmatrix}$$

Notice: Row 2 = Row 1 - 1, Row 3 = Row 1 - 2
This matrix has rank 1 because all rows are multiples of the first row!

6.2 Key Properties

- For $\mathbf{A} \in \mathbb{R}^{m \times n}$: $\text{rank}(\mathbf{A}) \leq \min(m, n)$
- **Full rank**: $\text{rank}(\mathbf{A}) = \min(m, n)$
- Square matrix is **invertible** iff it has full rank

- Low rank often means redundant information

7 Probability Fundamentals

7.1 Why Probability in ML?

Machine learning is fundamentally about dealing with uncertainty:

- Will this email be spam? (Classification uncertainty)
- What's the range of possible house prices? (Regression uncertainty)
- How confident is my model? (Model uncertainty)

7.2 Basic Probability Concepts

Sample Space (Ω): All possible outcomes **Event** (A): A subset of the sample space **Probability** ($P(A)$): How likely an event is

Properties:

$$0 \leq P(A) \leq 1 \tag{5}$$
$$P(\Omega) = 1 \tag{6}$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{7}$$

Example #7: Email Classification

Your spam filter processes 1000 emails:

- 200 are spam
- 800 are not spam
- Of spam emails, 180 contain word "free"
- Of non-spam emails, 50 contain word "free"

Probabilities:

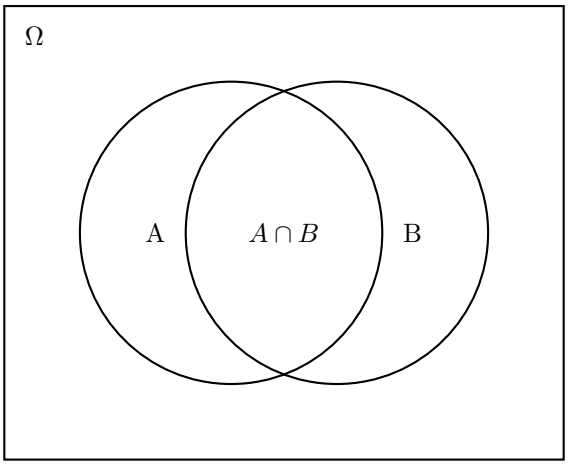
$$P(\text{spam}) = \frac{200}{1000} = 0.2 \tag{8}$$
$$P(\text{"free"}) = \frac{180 + 50}{1000} = 0.23 \tag{9}$$
$$P(\text{"free"}|\text{spam}) = \frac{180}{200} = 0.9 \tag{10}$$

7.3 Conditional Probability and Bayes' Theorem

Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Bayes' Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

This is the foundation of Naive Bayes classifiers!



8 Statistics: Making Sense of Data

8.1 Descriptive Statistics

Central Tendency:

$$\text{Mean: } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{11}$$

$$\text{Median: middle value when sorted} \tag{12}$$

$$\text{Mode: most frequent value} \tag{13}$$

Variability:

$$\text{Variance: } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{14}$$

$$\text{Standard Deviation: } \sigma = \sqrt{\sigma^2} \tag{15}$$

Example #8: Model Performance Analysis

Your model's accuracy across 10 different test sets:

$$\text{accuracies} = [0.85, 0.87, 0.83, 0.89, 0.86, 0.88, 0.84, 0.90, 0.85, 0.87]$$

Statistics:

- Mean: $\mu = 0.864$ (average performance)
- Standard deviation: $\sigma = 0.022$ (consistency)
- Range: $[0.83, 0.90]$ (best/worst case)

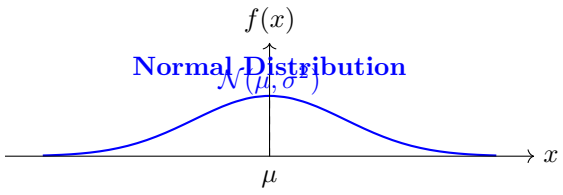
Low standard deviation means your model is consistent!

8.2 Important Distributions

1. Normal Distribution $\mathcal{N}(\mu, \sigma^2)$ - Bell-shaped, symmetric - Many real phenomena follow this - Central Limit Theorem

2. Bernoulli Distribution - Single coin flip: success/failure - Parameter: p (probability of success)

3. Binomial Distribution - Multiple coin flips - Parameters: n (trials), p (success probability)



9 Putting It All Together: A Complete Example

Example #9: Student Performance Analysis System

You're building a system to predict student performance. Here's how all concepts connect:
Data Representation: - Student feature vector: $\mathbf{x} = [study\_hours, attendance, prev\_gpa, sleep\_hours]^T$ - Dataset matrix: $\mathbf{X} \in \mathbb{R}^{1000\times 4}$ (1000 students, 4 features) - Target vector: $\mathbf{y} \in \mathbb{R}^{1000}$ (final exam scores)
Mathematical Operations: - Normalize features: $\mathbf{x}_{norm} = \frac{\mathbf{x}-\mu}{\sigma}$ (statistics) - Compute similarities: $\text{sim}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}$ (dot product, norms) - Linear model: $\hat{y} = \mathbf{w}^T\mathbf{x} + b$ (matrix-vector multiplication)
Analysis: - Check if $\mathbf{X}$ has full rank (are features independent?) - Use probability to handle uncertainty in predictions - Apply statistics to evaluate model performance
This single example uses scalars, vectors, matrices, norms, probability, and statistics!

10 Practice Problems

10.1 Warm-up Problems

Problem #1: Basic Vector Operations

Given vectors $\mathbf{x} = [3, -2, 1]^T$ and $\mathbf{y} = [1, 4, -2]^T$:
a) Calculate $\mathbf{x} + \mathbf{y}$
b) Calculate $\mathbf{x} \cdot \mathbf{y}$
c) Calculate $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_\infty$
d) What is the angle between $\mathbf{x}$ and $\mathbf{y}$? (Hint: $\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}$)
Solutions: a) $\mathbf{x} + \mathbf{y} = [4, 2, -1]^T$ b) $\mathbf{x} \cdot \mathbf{y} = 3(1) + (-2)(4) + 1(-2) = -7$ c) $\|\mathbf{x}\|_1 = 6$, $\|\mathbf{x}\|_2 = \sqrt{14}$, $\|\mathbf{x}\|_\infty = 3$ d) $\cos\theta = \frac{-7}{\sqrt{14}\sqrt{21}} = \frac{-7}{\sqrt{294}} \approx -0.408$, so $\theta \approx 114°$

Problem #2: Matrix Multiplication Dimensions

For each pair of matrices, determine if multiplication $\mathbf{AB}$ is possible. If yes, give the dimensions of the result:
a) $\mathbf{A} \in \mathbb{R}^{3\times 4}$, $\mathbf{B} \in \mathbb{R}^{4\times 2}$
b) $\mathbf{A} \in \mathbb{R}^{5\times 3}$, $\mathbf{B} \in \mathbb{R}^{2\times 7}$
c) $\mathbf{A} \in \mathbb{R}^{100\times 50}$, $\mathbf{B} \in \mathbb{R}^{50\times 1}$
d) If $\mathbf{AB} = \mathbf{C}$ where $\mathbf{C} \in \mathbb{R}^{6\times 8}$ and $\mathbf{A} \in \mathbb{R}^{6\times ?}$, what must be the dimensions of $\mathbf{B}$?
Solutions: a) Yes, result is 3×2 b) No, inner dimensions don't match ($3 \neq 2$) c) Yes, result is 100×1 (column vector) d) $\mathbf{A}$ must be $6 \times k$ and $\mathbf{B}$ must be $k \times 8$ for some k

10.2 Intermediate Problems

Problem #3: Rank and Linear Independence

Consider the matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 3 & 5 \end{bmatrix}$$

a) What is the rank of $\mathbf{A}$?
b) Are the columns linearly independent?
c) Is $\mathbf{A}$ invertible?
d) Can the system $\mathbf{Ax} = \mathbf{b}$ have a unique solution for some vector $\mathbf{b}$?
Solutions: a) Rank is 2 (row 2 = 2 × row 1, but row 3 is independent) b) No, because rank ¡ 3 c) No, because it's not full rank d) No, because the matrix is not invertible

Problem #4: Probability and Bayes

A diagnostic test for a disease has:

- 95% sensitivity (correctly identifies 95% of sick people)
- 90% specificity (correctly identifies 90% of healthy people)
- Disease prevalence in population: 2%

a) If someone tests positive, what's the probability they actually have the disease?
b) If someone tests negative, what's the probability they're actually healthy?
c) Why might this test not be very useful for screening?
Solutions: a) Using Bayes: $P(\text{disease}|+) = \frac{0.95 \times 0.02}{0.95 \times 0.02 + 0.10 \times 0.98} \approx 0.162$ (only 16.2%!) b) $P(\text{healthy}|\text{-}) = \frac{0.90 \times 0.98}{0.90 \times 0.98 + 0.05 \times 0.02} \approx 0.999$ (99.9%) c) Low prevalence leads to many false positives, making positive tests unreliable

10.3 Challenging Problems

Problem #5: Norm Relationships

For any vector $\mathbf{x} \in \mathbb{R}^n$, prove or find counterexamples for:
a) $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$
b) $\|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty$
c) $\|\mathbf{x}\|_1 \leq n\|\mathbf{x}\|_\infty$
d) When do we have equality in part (b)?
Solutions: a) False! Counter: $\mathbf{x} = [1, 1]^T$ gives $1 \leq \sqrt{2} \leq 2$ ✓, but $\mathbf{x} = [1, 0]^T$ gives $1 = 1 > 0$ × Correct: $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$ b) True: $\|\mathbf{x}\|_2^2 = \sum x_i^2 \leq n \max_i |x_i|^2 = n\|\mathbf{x}\|_\infty^2$ c) True: $\|\mathbf{x}\|_1 = \sum |x_i| \leq n \max_i |x_i| = n\|\mathbf{x}\|_\infty$ d) When all components have equal magnitude: $|x_1| = |x_2| = \ldots = |x_n|$

Problem #6: Covariance Matrix Analysis

Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where rows are samples and columns are features:
a) Write the formula for the sample covariance matrix $\mathbf{C}$
b) What does it mean if $\mathbf{C}$ has low rank?
c) If $\mathbf{C}$ is diagonal, what does this tell us about the features?
d) How would you use SVD to reduce dimensionality based on $\mathbf{C}$?
This connects linear algebra with statistics - think about what each mathematical property means for your data!
Solutions: a) $\mathbf{C} = \frac{1}{n-1}(\mathbf{X} - \boldsymbol{\mu})^T(\mathbf{X} - \boldsymbol{\mu})$ where $\boldsymbol{\mu}$ is mean vector b) Features are highly correlated/redundant; data lies in lower-dimensional subspace c) Features are uncorrelated (but not necessarily independent) d) Use SVD to find principal components; keep top-k components based on explained variance

10.4 Thought-Provoking Problems

Problem #7: The Curse of Dimensionality

Consider points uniformly distributed in a unit hypercube in d dimensions.
a) What fraction of the volume is within distance 0.1 from the boundary?
b) As d increases, what happens to the "typical" distance between random points?
c) How does this affect the effectiveness of nearest neighbor methods?
d) What does this suggest about the L2 vs L1 norm in high dimensions?
Hint: Think about how volume scales with dimension, and how distances behave in high-dimensional spaces.
Discussion Points: - In high dimensions, most volume is near the boundary - All pairwise distances become similar (distance concentration) - Nearest neighbors become less meaningful - L1 norm often works better than L2 in high dimensions

Problem #8: Simpson's Paradox in ML

A machine learning model shows these results:
Group A: 80% accuracy on 1000 samples
Group B: 70% accuracy on 1000 samples
Overall: 60% accuracy on 2000 samples
a) How is this possible?
b) What does this teach us about evaluation metrics?
c) How might this relate to fairness in ML models?
d) What additional information would you want to properly evaluate this model?
This problem connects probability, statistics, and critical thinking about ML evaluation!
Discussion: This is an example of Simpson's Paradox - aggregate statistics can be misleading when there are confounding variables. In ML, this highlights the importance of stratified evaluation and understanding your data distribution.

11 Summary and Next Steps

You've now covered the essential mathematical foundations:

- **Scalars, Vectors, Matrices**: The building blocks of ML
- **Norms**: Measuring size and distance
- **Rank**: Understanding linear independence

- **Probability**: Handling uncertainty
- **Statistics**: Making sense of data

These concepts appear everywhere in machine learning:

- Linear regression uses matrix operations and statistics
- Neural networks rely on vector operations and probability
- SVD and PCA use rank and matrix decompositions
- Regularization uses different norms
- Bayesian methods use probability throughout

Next Tutorial: We'll build on these foundations to understand ML-specific conventions, accuracy metrics, and evaluation methods.

12 Further Reading

- **Linear Algebra**: Gilbert Strang's "Introduction to Linear Algebra"
- **Probability**: Sheldon Ross's "A First Course in Probability"
- **Statistics**: Larry Wasserman's "All of Statistics"
- **ML Math**: Deisenroth, Faisal, Ong's "Mathematics for Machine Learning"
- **Online**: Khan Academy, 3Blue1Brown's "Essence of Linear Algebra"