Notes on *Mathematics For Machine Learning*

Jonathan Chen

September 13, 2025

Contents

1 Introduction and Motivation

- Machine learning designs algorithms that **automatically** extract valuable information from data. "Automatic" emphasizes general-purpose methodologies that can be applied across diverse datasets, producing meaningful outputs without heavy domain-specific customization.

- **Data**
  - ML is inherently data-driven; data forms the basis of every method.
  - Goal: uncover patterns and structure directly from datasets with minimal prior knowledge.
  - Example: topic modeling in large document corpora (Hoffman et al., 2010).

- **Model**
  - Represents the process generating the data (e.g., regression maps inputs to real-valued outputs).
  - Mitchell (1997): a model learns if its task performance improves after exposure to data.
  - Strong models must not only fit observed data but also **generalize** to unseen cases, which is essential for future applications.

- **Learning**
  - The process of optimizing model parameters to capture patterns and relationships in data.
  - Enables adaptability across tasks and datasets, reducing the need for manual rule design.

- **Mathematical Foundations**
  - Provide clarity on the principles underlying complex ML systems.
  - Enable creation of new methods beyond existing software packages.
  - Support debugging and evaluation of current approaches.
  - Reveal assumptions and limitations, which is crucial for reliable and responsible deployment in practice.

1.1 Finding words for intuition

- In machine learning, concepts and terms can be ambiguous; the same word may have different meanings depending on context.
  - Example: **algorithm**
    * As predictor: a system making predictions from input data.

    * As training procedure: a system adapting parameters so the predictor performs well on unseen data.

- The three main components of an ML system are **data**, **models**, and **learning**.
  - **Data**: represented as vectors.
    * Computer science view: array of numbers.
    * Physics view: arrow with direction and magnitude.
    * Mathematics view: object obeying addition and scaling.
  - **Model**: simplified version of the data-generating process, capturing aspects relevant for prediction and enabling exploration of hidden patterns.
  - **Learning**: training a model means optimizing its parameters with respect to a utility function measuring predictive performance.
    * Analogy: climbing a hill to maximize performance.
    * Training accuracy may only reflect memorization; the real goal is generalization to unseen data.

1.2 Two Ways to Read this Book

- **Two strategies for learning mathematics for ML**
  - Bottom-up: build from foundational to advanced concepts.
    * Advantage: solid grounding, each step relies on previous knowledge.
    * Disadvantage: foundations may feel abstract or unmotivated.
  - Top-down: start from practical needs, drill down into required math.
    * Advantage: clear motivation, direct path to applications.
    * Disadvantage: knowledge may rest on weak foundations.
- **Book structure**
  - Modular design: can be read bottom-up or top-down.
  - Part I: mathematics foundations.
  - Part II: machine learning applications (regression, dimensionality reduction, density estimation, classification).
- **Mathematical foundations (Part I)**
  - Linear algebra: vectors, matrices, data representation.
  - Analytic geometry: similarity and distance between vectors.
  - Matrix decomposition: structure and efficient computation.
  - Vector calculus: gradients for optimization.
  - Probability theory: quantifying uncertainty and noise.

- Optimization: finding maxima/minima using gradients.

- **Applications (Part II)**
  - Regression: functions mapping inputs to outputs; MLE, MAP, Bayesian linear regression.
  - Dimensionality reduction: compact representations (e.g., PCA).
  - Density estimation: probability distributions for data (e.g., Gaussian mixtures).
  - Classification: discrete labels (e.g., support vector machines).

2 Linear Algebra

- **Algebra**: a set of objects (symbols) and rules for manipulating them.
- **Linear algebra**: study of vectors and the rules for combining them.
- **Vectors**: abstract objects that can be added and scaled (closure property). Any object satisfying these rules is a vector.
  - Geometric vectors: arrows with direction and magnitude; addition and scalar multiplication preserve vector form.
  - Polynomials: closed under addition and scalar multiplication; abstract but valid vectors.
  - Audio signals: represented as sequences of numbers; addition and scaling produce new signals.
  - Elements of $\mathbb{R}^n$: n-tuples of real numbers; focus of this book. Operations are defined component-wise.
- **Practical viewpoint**: vectors in $\mathbb{R}^n$ correspond to arrays in computer implementations. Many languages support array operations, enabling efficient ML algorithms.
- **Closure and vector spaces**: the set of all possible vectors generated by addition and scaling forms a vector space. Vector spaces and their properties underpin much of ML.
- **Role in ML**:
  - Chapter 3: analytic geometry for similarity and distances.
  - Chapter 5: matrix operations for vector calculus.
  - Chapter 9: linear regression solved via least squares.
  - Chapter 10: dimensionality reduction with projections (PCA).
  - Chapter 12: classification methods relying on linear algebra.

2.1 Systems of Linear Equations

- Many problems in linear algebra can be formulated as **systems of linear equations**. Linear algebra provides systematic tools to solve them.
- **General form**:

  $$a_{11}x_1 + \cdots + a_{1n}x_n = b_1, \quad \ldots, \quad a_{m1}x_1 + \cdots + a_{mn}x_n = b_m$$

  where $a_{ij}, b_i \in \mathbb{R}$ and $x_1, \ldots, x_n$ are unknowns.
  - Solutions are n-tuples $(x_1, \ldots, x_n) \in \mathbb{R}^n$ that satisfy all equations.
  - A system can have no solution, exactly one solution, or infinitely many solutions.

- **Examples**:
  - No solution: equations contradict each other.
  - Unique solution: $(1, 1, 1)$ solves one example system.
  - Infinitely many solutions: free variables parameterize solution sets.
- **Geometric interpretation**:
  - Two variables: each equation is a line in the x_1x_2-plane; solutions = intersection of lines.
  - Three variables: each equation is a plane; intersections may yield a plane, line, point, or empty set.
- **Matrix form**:
$$A\mathbf{x} = \mathbf{b}, \quad A \in \mathbb{R}^{m\times n},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{b} \in \mathbb{R}^m$$
where A collects coefficients a_{ij}, $\mathbf{x}$ collects unknowns, and $\mathbf{b}$ collects constants.

2.2 Matrices

- **Matrices**: central objects in linear algebra.
  - Represent systems of linear equations compactly.
  - Represent linear mappings (to be discussed later).
- **Definition**: A real-valued (m, n) matrix A is an ordered $m{\cdot}n$-tuple of elements $a_{ij} \in \mathbb{R}$, arranged in m rows and n columns:
$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{R}$$
- **Special cases**:
  - $(1, n)$-matrix: row (row vector).
  - $(m, 1)$-matrix: column (column vector).
- **Notation**: $\mathbb{R}^{m\times n}$ is the set of all real-valued (m, n) matrices.
- **Vectorization**: A matrix $A \in \mathbb{R}^{m\times n}$ can be re-shaped as a vector $a \in \mathbb{R}^{mn}$ by stacking its n columns.

2.2.1 Matrix addition and multiplication

- **Matrix addition**: For $A, B \in \mathbb{R}^{m \times n}$,

$$A + B = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

- **Matrix multiplication**: For $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times k}$,

$$C = AB \in \mathbb{R}^{m \times k}, \quad c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}$$

  - Multiply i-th row of A with j-th column of B and sum.
  - Defined only when the inner dimensions match.
  - In general, $AB \neq BA$.

- **Hadamard product**: element-wise multiplication $c_{ij} = a_{ij} b_{ij}$, distinct from matrix multiplication.

- **Identity matrix**: For $n \in \mathbb{N}$,

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

  - Property: $I_m A = A I_n = A$, for $A \in \mathbb{R}^{m \times n}$.

- **Algebraic properties**:
  - Associativity: $(AB)C = A(BC)$
  - Distributivity: $(A + B)C = AC + BC, \quad A(C + D) = AC + AD$

2.2.2 Inverse and Transpose

- **Inverse of a square matrix**: For $A \in \mathbb{R}^{n \times n}$, if there exists $B \in \mathbb{R}^{n \times n}$ such that

$$AB = I_n = BA,$$

  then B is called the inverse of A, denoted A^{-1}.

  - A is **invertible / nonsingular / regular** if A^{-1} exists.
  - A is **singular / noninvertible** if A^{-1} does not exist.
  - If A^{-1} exists, it is unique.

- **2×2 case**: For

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

the inverse exists iff $\det(A) = a_{11}a_{22} - a_{12}a_{21} \neq 0$, and

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

- **Example**:

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 4 & 5 \\ 6 & 7 & 7 \end{bmatrix}, \quad B = \begin{bmatrix} -7 & -7 & 6 \\ 2 & 1 & -1 \\ 4 & 5 & -4 \end{bmatrix}$$

satisfy $AB = I = BA$, so $B = A^{-1}$.

- **Transpose**: For $A \in \mathbb{R}^{m\times n}$, the transpose $A^\top \in \mathbb{R}^{n\times m}$ is defined by $(A^\top)_{ij} = a_{ji}$.
  - Obtained by writing columns of A as rows of $A^\top$.
- **Properties**:

$$AA^{-1} = I = A^{-1}A$$
$$(AB)^{-1} = B^{-1}A^{-1}, \quad (A+B)^{-1} \neq A^{-1} + B^{-1}$$
$$(A^\top)^\top = A, \quad (AB)^\top = B^\top A^\top, \quad (A+B)^\top = A^\top + B^\top$$

- **Symmetric matrices**: $A \in \mathbb{R}^{n\times n}$ is symmetric if $A = A^\top$.
  - Only square matrices can be symmetric.
  - If A is invertible, then $A^\top$ is invertible and $(A^{-1})^\top = (A^\top)^{-1} = A^{-\top}$.
  - Sum of symmetric matrices is symmetric.
  - Product of symmetric matrices is not necessarily symmetric.

2.2.3 Multiplication by scalar

- For $A \in \mathbb{R}^{m\times n}$ and $\lambda \in \mathbb{R}$, scalar multiplication is defined as

$$(\lambda A)_{ij} = \lambda a_{ij}.$$

Practically, each entry of A is scaled by λ.

- **Properties:** For $\lambda, \psi \in \mathbb{R}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{n\times k}$:
  - **Associativity:** $(\lambda\psi)C = \lambda(\psi C)$
  - Compatible with matrix multiplication:

$$\lambda(BC) = (\lambda B)C = B(\lambda C) = (BC)\lambda$$

- Transpose: $(\lambda C)^\top = \lambda C^\top$
- **Distributivity:**

$$(\lambda + \psi)C = \lambda C + \psi C, \quad \lambda(B + C) = \lambda B + \lambda C$$

- **Example:** For $C = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $\lambda, \psi \in \mathbb{R}$,

$$(\lambda + \psi)C = \begin{bmatrix} \lambda + \psi & 2(\lambda + \psi) \\ 3(\lambda + \psi) & 4(\lambda + \psi) \end{bmatrix} = \lambda C + \psi C$$

2.2.4 Compact Representations of Systems of Linear Equations

- A system of linear equations can be expressed using matrix notation.
- Example:

$$\begin{aligned} 2x_1 + 3x_2 + 5x_3 &= 1 \\ 4x_1 - 2x_2 - 7x_3 &= 8 \\ 9x_1 + 5x_2 - 3x_3 &= 2 \end{aligned}$$

can be written as

$$\begin{bmatrix} 2 & 3 & 5 \\ 4 & -2 & -7 \\ 9 & 5 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 8 \\ 2 \end{bmatrix}.$$

- General form:

$$A\mathbf{x} = \mathbf{b}, \quad A \in \mathbb{R}^{m \times n},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{b} \in \mathbb{R}^m$$

- Interpretation: The product $A\mathbf{x}$ is a **linear combination** of the columns of A, with coefficients given by the components of $\mathbf{x}$.

2.3 Solving Systems of Linear Equations

- General form of a linear system:

$$a_{11}x_1 + \cdots + a_{1n}x_n = b_1, \quad \ldots, \quad a_{m1}x_1 + \cdots + a_{mn}x_n = b_m$$

where $a_{ij}, b_i \in \mathbb{R}$ are known constants and x_j are unknowns.

- Compact matrix form:

$$A\mathbf{x} = \mathbf{b}, \quad A \in \mathbb{R}^{m \times n},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{b} \in \mathbb{R}^m$$

- Matrices provide a concise framework to represent and manipulate linear systems, enabling the use of algebraic operations.
- Goal: focus on **solving** systems of linear equations and introduce an algorithm for computing the inverse of a matrix as part of the solution process.

2.3.1 Particular and General Solution

- Example system:

$$\begin{bmatrix} 1 & 0 & 8 & -4 \\ 0 & 1 & 2 & 12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 42 \\ 8 \end{bmatrix}.$$

- Since the system has two equations and four unknowns, we expect infinitely many solutions.

- **Particular solution**:

$$\mathbf{x}_p = \begin{bmatrix} 42 \\ 8 \\ 0 \\ 0 \end{bmatrix},$$

  since $b = 42c_1 + 8c_2$ (with c_i denoting the i-th column).

- **Solutions to $Ax = 0$ (homogeneous system)**:

$$\lambda_1 \begin{bmatrix} 8 \\ 2 \\ -1 \\ 0 \end{bmatrix}, \quad \lambda_2 \begin{bmatrix} -4 \\ 12 \\ 0 \\ -1 \end{bmatrix}, \quad \lambda_1, \lambda_2 \in \mathbb{R}.$$

- **General solution**:

$$\mathbf{x} = \begin{bmatrix} 42 \\ 8 \\ 0 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 8 \\ 2 \\ -1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} -4 \\ 12 \\ 0 \\ -1 \end{bmatrix}, \quad \lambda_1, \lambda_2 \in \mathbb{R}.$$

- **General method**:
  1. Find a particular solution of $Ax = b$.
  2. Find all solutions of $Ax = 0$.
  3. Combine both to obtain the general solution.

- General systems are not usually in this convenient form, so we use **Gaussian elimination** to reduce them into a form where steps (1)–(3) can be applied.

2.3.2 Elementary Transformations

- **Elementary transformations** simplify a system of linear equations without changing its solution set:
  - Exchange of two equations (row swap)

- Multiplication of a row by $\lambda \in \mathbb{R} \setminus \{0\}$
- Addition of a multiple of one row to another row

- Systems are often written in **augmented matrix form**: $(A|b)$ compactly represents $A\mathbf{x} = b$.

- Example: Transforming a system via row operations leads to an **augmented matrix** in **row-echelon form (REF)**.

$$\begin{bmatrix} 1 & -2 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 3 & -2 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & a+1 \end{bmatrix}$$

corresponds to

$$\begin{aligned} x_1 - 2x_2 + x_3 - x_4 + x_5 &= 0 \\ x_3 - x_4 + 3x_5 &= -2 \\ x_4 - 2x_5 &= 1 \\ 0 &= a+1 \end{aligned}$$

- From REF:
  - Only solvable if $a = -1$.
  - A particular solution is $\begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}$.
  - The general solution is

$$\mathbf{x} = \begin{bmatrix} 2 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ -1 \\ 2 \\ 1 \end{bmatrix}, \quad \lambda_1, \lambda_2 \in \mathbb{R}.$$

- **Row-echelon form (REF):**
  - All zero rows are at the bottom.
  - Each pivot (first nonzero entry in a row) is strictly to the right of the pivot above it.
  - This creates a staircase structure.

- **Basic vs. free variables:**
  - Pivot columns $\rightarrow$ basic variables.

- Non-pivot columns $\to$ free variables.
- Example: in (2.45), x_1, x_3, x_4 are basic; x_2, x_5 are free.

- **Reduced row-echelon form (RREF):**
  - Matrix is in REF.
  - Every pivot is 1.
  - Pivot is the only nonzero entry in its column.

- **Gaussian elimination:** algorithm that applies elementary transformations to bring a system into RREF, enabling direct solution construction.

2.4 The Minus-1 Trick

- **Minus-1 Trick:** A method to read solutions of $A\mathbf{x} = 0$ when A is in reduced row-echelon form (RREF).
  - Extend $A \in \mathbb{R}^{k\times n}$ to $\tilde{A} \in \mathbb{R}^{n\times n}$ by adding rows of the form

$$(0 \ \cdots \ 0 \ -1 \ 0 \ \cdots \ 0),$$

    so that the diagonal entries of $\tilde{A}$ are either 1 or -1.
  - Columns of $\tilde{A}$ with -1 on the diagonal form a basis of the solution space (kernel/null space).

- **Example (Minus-1 Trick):**

$$A = \begin{bmatrix} 1 & 3 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 9 \\ 0 & 0 & 0 & 1 & -4 \end{bmatrix} \quad \Rightarrow \quad \tilde{A} = \begin{bmatrix} 1 & 3 & 0 & 0 & 3 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 9 \\ 0 & 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

- Solutions:

$$\mathbf{x} = \lambda_1 \begin{bmatrix} 3 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 3 \\ 0 \\ 9 \\ -4 \\ -1 \end{bmatrix}, \quad \lambda_1, \lambda_2 \in \mathbb{R}.$$

- **Inverse Calculation:** To compute A^{-1} for $A \in \mathbb{R}^{n\times n}$, solve

$$AX = I_n.$$

  Write the augmented matrix

$$(A \,|\, I_n),$$

  and perform Gaussian elimination until

$$(A \,|\, I_n) \implies (I_n \,|\, A^{-1}).$$

- **Example (Inverse):** For

$$A = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

  row reduction of $(A|I_4)$ yields

$$A^{-1} = \begin{bmatrix} -1 & 2 & -2 & 2 \\ 1 & -1 & 2 & -2 \\ 1 & -1 & 1 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}.$$

  Verification: $AA^{-1} = I_4$.