Mathematical Foundations for Machine Learning

Week 0 (Optional Reference)

Daniel E. Acuna
Associate Professor, University of Colorado Boulder

Mathematical Foundations for Machine Learning

Contents of This Video

In this reference material, we will cover:

- Linear algebra fundamentals (scalars, vectors, matrices)
- Matrix operations and properties
- Calculus concepts for optimization
- Probability theory basics
- Common probability distributions
- Essential statistics

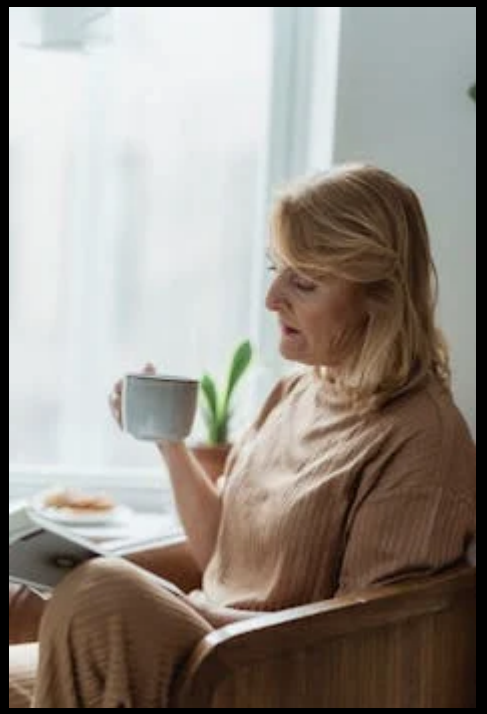
Contents of This Video

In this reference material, we will cover:

- Linear algebra fundamentals (scalars, vectors, matrices)
- Matrix operations and properties
- Calculus concepts for optimization
- Probability theory basics
- Common probability distributions
- Essential statistics

How to Use This Reference Material

This mathematical review:



- Serves as a **reference resource** for the entire ML specialization
- Covers foundations needed across all three courses:
  - Supervised Learning
  - Unsupervised Learning
  - Deep Learning
- Contains concepts that may not be used immediately
- Is designed for you to **return to** whenever needed

Remember, you don’t need to master all of these concepts right now. Some will become more relevant as you progress through specific topics in the specialization. For instance, the probability concepts will be especially important when you reach unsupervised learning techniques like clustering and dimensionality reduction. :::

Linear Algebra Review

Scalars, Vectors, and Matrices

Scalars

- Represented by Greek letters α, β, γ or regular letters
- Single numerical values
- Example: $\alpha = 0.1, \beta = 10^{-10}$

Vectors

- Collections of scalars arranged in a column
- Default representation is a column vector
- Example: $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

Matrix Notation

A matrix $\mathbf{X}$ with n rows and p columns:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- n observations (rows)
- p features/variables (columns)
- Each element x_{ij} is the value of feature j for observation i

Rows and Columns of a Matrix

Row Representation

- The i-th row: $x_i^T = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}$
- Each row is a single observation/sample

Column Representation

- The j-th column: $\mathbf{X}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$
- Each column is a single feature/variable

Matrix Operations

- **Scalar multiplication**: $\alpha\mathbf{A} = (\alpha \times a_{ij})_{ij}$
- **Matrix addition**: $\mathbf{A} + \mathbf{B}$ (element-wise addition)
- **Matrix multiplication**: $\mathbf{AB}$ (requires $\#\text{cols}_A = \#\text{rows}_B$)

$$\mathbf{AB} = \left(\sum_z a_{iz} b_{zj} \right)_{ij}$$

- **Matrix transposition**: $\mathbf{A}^T = (a_{ij})_{ji}$

Special Matrices and Properties

- **Identity matrix**: $I = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$
  - Product with any matrix returns the original: $\mathbf{A}I = I\mathbf{A} = \mathbf{A}$
- **Matrix inverse**: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = I$
- **Properties**:
  - Matrix addition is commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
  - Matrix multiplication is NOT commutative: $\mathbf{AB} \neq \mathbf{BA}$ (generally)
  - Transpose of a product: $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

Dimensions and Notation

To indicate dimensions, we use notation like:

- Scalar: $a \in \mathbb{R}$
- Vector of length n: $\mathbf{a} \in \mathbb{R}^n$
- Matrix of size $r \times s$: $\mathbf{A} \in \mathbb{R}^{r \times s}$

In machine learning:

- Bold capitals ($\mathbf{X}$) for matrices
- Bold lowercase ($\mathbf{y}$) for vectors of length n
- Regular lowercase (x) for other vectors or scalars

Linear Models as Matrix Operations

For a linear model:

$$y = b_0 + \sum_{j=1}^{p} b_j x_j$$

We can represent this in matrix form:

$$\mathbf{y} = \mathbf{X}\mathbf{b}$$

Where:

- $\mathbf{X}$ includes a column of 1s for the intercept
- $\mathbf{b}$ is the vector of coefficients
- $\mathbf{y}$ is the vector of predictions

Calculus and Optimization

Learning as Optimization

In machine learning, we:

1. Define a model with parameters Θ
2. Define a loss function $L(\Theta)$ measuring prediction error
3. Find parameters that minimize the loss:

$$\hat{\Theta} = \arg\min_{\Theta} L(\Theta)$$

Example: For a simple model $\hat{y} = b_0$ and data $y = \{30000, 40000, 30000\}$

- Using squared error loss: $L(b_0) = \sum_{i=1}^{n} (b_0 - y_i)^2$
- The optimal value is $b_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = 33333.33$

Derivatives and Optimization

Derivatives measure how a function changes as inputs change:

$$\frac{df(x)}{dx} \approx \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

To find a minimum of function $f(x)$, we solve:

$$\frac{df(x)}{dx} = 0$$

Examples:

- $f_1(x) = a + xb \implies \frac{df_1(x)}{dx} = b$
- $f_2(x) = x^2 \implies \frac{df_2(x)}{dx} = 2x$

Common Derivation Rules

- **Constant rule:** $\frac{d(c)}{dx} = 0$
- **Power rule:** $\frac{d(x^n)}{dx} = nx^{n-1}$
- **Constant multiple rule:** $\frac{d(cf(x))}{dx} = c\frac{df(x)}{dx}$
- **Sum rule:** $\frac{d(f(x)+g(x))}{dx} = \frac{df(x)}{dx} + \frac{dg(x)}{dx}$
- **Chain rule:** $\frac{dg(f(x))}{dx} = \frac{dg(f)}{df} \cdot \frac{df(x)}{dx}$
- **Exponential:** $\frac{d(e^x)}{dx} = e^x$
- **Logarithm:** $\frac{d(\log(x))}{dx} = \frac{1}{x}$

Gradient Descent

For functions with multiple parameters, we use the gradient:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Gradient descent algorithm:

1. Start with initial parameters θ_0
2. Update: $\theta_{t+1} = \theta_t - \alpha \nabla f(\theta_t)$
3. Repeat until convergence

Where α is the learning rate that controls step size.

Example: Sigmoid Function Derivative

The sigmoid function is commonly used in logistic regression:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Its derivative can be calculated using the chain rule:

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z))$$

This derivative has a special form that makes it computationally efficient.

Probability

Probability Basics

Probability deals with uncertainty and randomness, with two main interpretations:

- **Frequency**: Relative frequency over many repeated trials
- **Subjective**: Degree of belief that must be updated consistently

Axioms of Probability:

1. Probability is non-negative: $P(A) \geq 0$
2. Probability of the entire sample space is 1: $P(S) = 1$
3. Probability of the union of disjoint events equals the sum of their probabilities

Random Variables and Distributions

A **random variable** is a function mapping outcomes to numerical values.

Types of random variables:

- **Discrete**: Takes on a countable number of values
  - Example: Number of heads in 5 coin tosses
- **Continuous**: Takes on values in a continuous range
  - Example: Height of a randomly selected person

Probability distributions:

- Describe how likely different values are
- Must satisfy: all probabilities ≥ 0 and sum/integrate to 1

Discrete Probability Distributions

Probability Mass Function (PMF) for discrete random variable X:

- Gives probability for each possible value: $P(X = x)$
- Must satisfy: $P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Examples:

- **Bernoulli distribution** (single coin flip):

$$P(X = x) = p^x(1 - p)^{1-x} \text{ for } x \in \{0, 1\}$$

- **Uniform distribution** (equal probability for values a to b):

$$P(X = x) = \frac{1}{b - a + 1} \text{ for } a \leq x \leq b$$

Continuous Probability Distributions

Probability Density Function (PDF) for continuous random variable X:

- Not the probability itself, but a density
- Probability in interval: $P(a \leq X \leq b) = \int_a^b p(x)dx$
- Must satisfy: $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)dx = 1$

Examples:

- **Uniform distribution** on interval [a,b]:

$$p(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

- **Gaussian/Normal distribution**:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Expectation and Variance

The **expectation** or mean of a random variable X:

$$E[X] = \sum_x x \cdot P(X = x) \text{ (discrete)}$$

$$E[X] = \int_x x \cdot p(x)dx \text{ (continuous)}$$

The **variance** measures spread around the mean:

$$Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

Covariance measures how two variables vary together:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Joint, Marginal, and Conditional Probability

Joint probability of events A and B:

$$P(A, B) = P(A \text{ and } B)$$

Marginal probability of A (across all possible B values):

$$P(A) = \sum_B P(A, B) \text{ (discrete)}$$

$$P(A) = \int_B P(A, B) dB \text{ (continuous)}$$

Conditional probability of A given B:

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

Independence and Bayes' Theorem

Independence of events A and B:

$$P(A, B) = P(A) \cdot P(B)$$

$$P(A|B) = P(A)$$

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Where:

- $P(A|B)$ is the posterior probability
- $P(B|A)$ is the likelihood
- $P(A)$ is the prior probability

Summary

Key Mathematical Foundations for Machine Learning {transition="slide" small}

Linear Algebra

- Vectors and matrices for data representation
- Matrix operations for efficient computation
- Linear models as matrix operations

Calculus

- Derivatives and gradients for optimization
- Finding minima of loss functions
- Gradient descent algorithm

Probability

- Random variables and distributions
- Expectation, variance, covariance
- Joint, marginal, and conditional probability
- Bayes' theorem

Connecting Mathematics to Machine Learning

Mathematical concepts directly translate to ML tasks:

Mathematics	Machine Learning Application
Matrix multiplication	Making predictions with linear models
Gradient descent	Training model parameters
Probability distributions	Modeling data and uncertainty
Bayes' theorem	Updating beliefs based on evidence
Derivatives	Calculating how to update parameters

What We've Covered

In this video, we covered:

- **Linear algebra fundamentals** including vectors, matrices, and operations
- **Calculus concepts** for optimization and gradient descent
- **Probability theory** and common distributions
- **Statistical measures** like mean, variance, and correlation
- **Mathematical connections** to machine learning applications

Remember: These mathematical foundations will serve as your toolkit throughout the specialization!

A Reference for Your Machine Learning Journey

Keep this material as a reference:

- Revisit specific sections when needed
- Use it to understand the “why” behind algorithms
- Some concepts will become more relevant in later courses
- Mathematics is the language of machine learning