Foundations of Mathematics for Data Science: Linear Algebra, Calculus, and Probability

Jihun Yun

Contents

1 Linear Algebra

Linear algebra studies vectors, matrices, and linear transformations. It is a fundamental language in data science, machine learning, and statistics. This section summarizes key ideas such as vectors, matrices, linear mappings, eigenvalues, eigenvectors, rank, and how these concepts connect to statistical methods like PCA.

Vectors and Matrices

A **vector** is an ordered list of real numbers:

$$\mathbf{v} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}.$$

Vectors represent direction and magnitude in space.

A **matrix** is a rectangular array of numbers:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Matrices encode transformations, data tables, and systems of equations.

Given a matrix A and a vector $\mathbf{x}$, the multiplication $A\mathbf{x}$ produces a new vector. This can be interpreted as stretching, rotating, or shearing the space.

Linear Transformations

A function $T : \mathbb{R}^n \to \mathbb{R}^m$ is a **linear transformation** if:

$$T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y}), \qquad T(c\mathbf{x}) = cT(\mathbf{x}).$$

Every linear transformation can be expressed using a matrix A:

$$T(\mathbf{x}) = A\mathbf{x}.$$

Example of a Linear Transformation

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{v} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \qquad A\mathbf{v} = \begin{pmatrix} 7 \\ 3 \end{pmatrix}.$$

This matrix performs a shear transformation.

Eigenvalues and Eigenvectors

The eigenvalues and eigenvectors reveal the directions in which a transformation is preserved.

A non-zero vector $\mathbf{v}$ is an **eigenvector** of A if:

$$A\mathbf{v} = \lambda\mathbf{v}.$$

The scalar λ is the **eigenvalue**. Geometrically, the transformation stretches or compresses $\mathbf{v}$, but does not change direction.

2x2 Eigenvalue Example

Let

$$A = \begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix}.$$

Solve the characteristic equation:

$$\det(A - \lambda I) = \det \begin{pmatrix} 3-\lambda & 1 \\ 0 & 2-\lambda \end{pmatrix} = (3-\lambda)(2-\lambda) = 0.$$

Thus

$$\lambda_1 = 3, \qquad \lambda_2 = 2.$$

Eigenvector for $\lambda = 3$:

$$(A - 3I)\mathbf{v} = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \mathbf{v} = \mathbf{0} \quad \Rightarrow \quad \mathbf{v} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Eigenvector for $\lambda = 2$:

$$(A - 2I)\mathbf{v} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \mathbf{v} = \mathbf{0} \quad \Rightarrow \quad \mathbf{v} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Inverse, Rank, and Linear Independence

A matrix A is **invertible** if there exists A^{-1} such that:

$$AA^{-1} = A^{-1}A = I.$$

Invertibility requires:

- $\det(A) \neq 0$
- rows and columns are linearly independent
- full rank

Rank is the number of linearly independent columns of A. If columns are linearly dependent, the matrix loses information and may fail to be invertible.

A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots\}$ is **linearly independent** if:

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots = \mathbf{0} \Rightarrow c_1 = c_2 = \cdots = 0.$$

Connection to PCA

In statistics, PCA (Principal Component Analysis) is fundamentally an eigenvalue problem. Given a covariance matrix Σ, PCA finds eigenvectors of Σ:

$$\Sigma\mathbf{v} = \lambda\mathbf{v}.$$

Large eigenvalues correspond to directions with the highest variance in the data. Thus, the entire dimensionality reduction procedure is powered by linear algebra.

2 Calculus

Calculus studies how quantities change. It provides tools to analyze limits, continuity, derivatives, gradients, and integrals. These concepts are essential in optimization methods such as gradient descent and in theoretical models across economics, physics, and statistics.

Limits and Continuity

The **limit** of a function $f(x)$ as x approaches a is written as:

$$\lim_{x \to a} f(x) = L.$$

This expresses that the value of $f(x)$ becomes arbitrarily close to L when x is close to a.

A function is **continuous** at $x = a$ if:

$$\lim_{x \to a} f(x) = f(a).$$

Continuity means no jumps, holes, or breaks in the graph.

Derivative and Rate of Change

The **derivative** of $f(x)$ measures the instantaneous rate of change:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Geometrically, $f'(x)$ is the slope of the tangent line at x.

Example: Derivative of x^2

Let $f(x) = x^2$. Then:

$$f'(x) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h \to 0} (2x + h) = 2x.$$

Thus, the slope increases linearly with x. When $x = 1$, the slope is 2; when $x = 5$, the slope is 10.

Connection to Gradient Descent

In machine learning and optimization, one seeks to minimize a loss function $L(\theta)$. Gradient descent updates parameters in the opposite direction of the derivative:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \frac{dL}{d\theta},$$

where η is the learning rate.

The derivative tells us how steeply the loss changes. If the slope is large, the update should be large. If the slope is small (near zero), the update becomes small and converges to a minimum.

Gradients generalize derivatives to higher dimensions using partial derivatives.

Integral and Area

If the derivative measures instantaneous change, the **integral** measures accumulation. The definite integral of $f(x)$ from a to b is:

$$\int_a^b f(x)\, dx.$$

Geometrically, it represents the area under the curve.

Fundamental Theorem of Calculus

The Fundamental Theorem connects differentiation and integration:

If $F'(x) = f(x)$, then:

$$\int_a^b f(x)\, dx = F(b) - F(a).$$

Thus, differentiation and integration are inverse operations.

Example: Integral of x^2

Since

$$F(x) = \frac{x^3}{3} \quad \text{satisfies} \quad F'(x) = x^2,$$

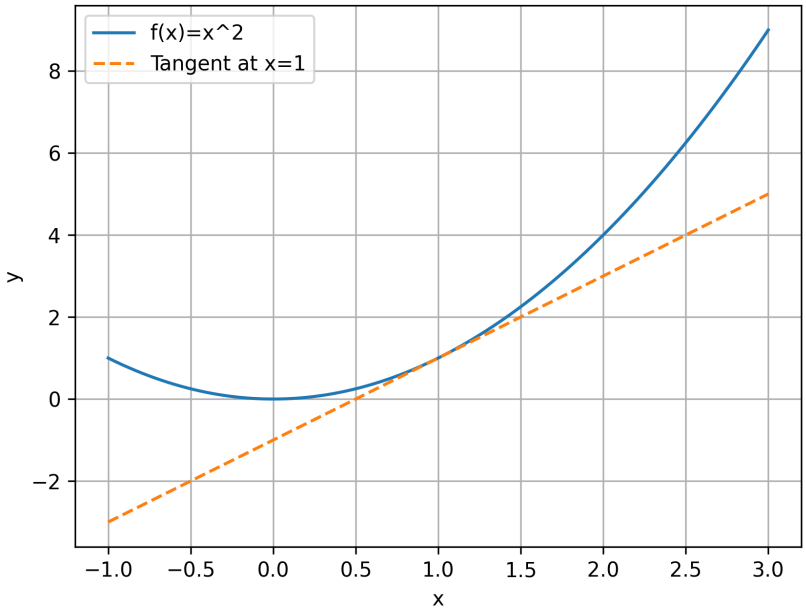
we have:

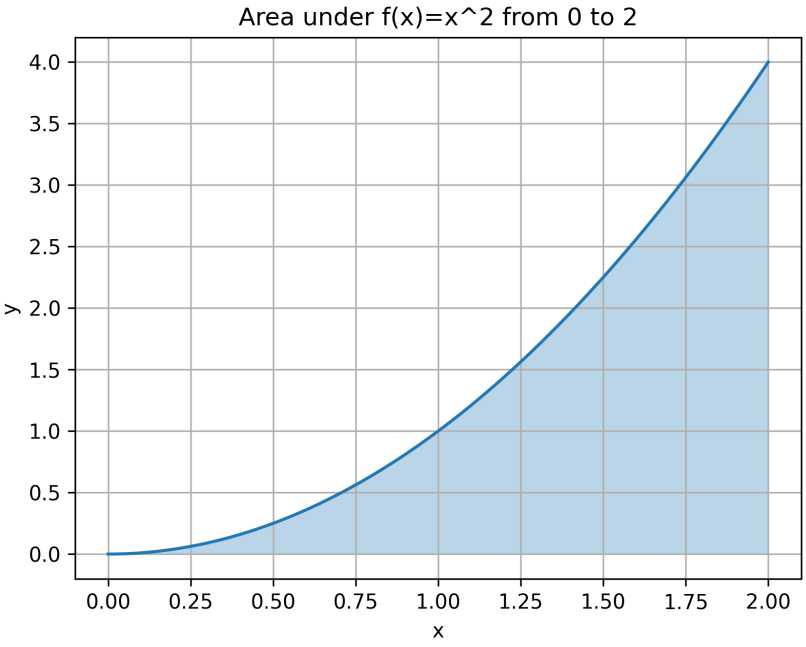
$$\int_0^2 x^2\, dx = \left[\frac{x^3}{3}\right]_0^2 = \frac{8}{3}.$$

Graphical Illustrations

Below are two simple visualizations commonly used in calculus:

- slope of the tangent line for $f(x) = x^2$
- area under the curve for $\int_0^2 x^2\,dx$





These graphs highlight the central ideas of calculus: the **local rate of change** (derivative) and **global accumulation** (integral).

3 Probability

Probability theory provides a mathematical framework for describing uncertainty. It formalizes randomness using probability spaces, conditional probabilities, and distributions such as the Bernoulli, binomial, and normal distributions. These concepts are essential in statistics, machine learning, and empirical modeling such as election forecasting.

Sample Space and Events

A **sample space** Ω is the set of all possible outcomes of an experiment. An **event** A is a subset of Ω.

Example:

$$\Omega = \{\text{Heads}, \text{Tails}\}, \qquad A = \{\text{Heads}\}.$$

A probability measure P assigns each event A a number $0 \leq P(A) \leq 1$, satisfying consistency and additivity.

Conditional Probability

The conditional probability of A given B is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0.$$

This represents the probability of A occurring assuming that B is known to have occurred.

Bayes' Theorem

Bayes' theorem connects prior belief, likelihood, and posterior belief:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

This allows us to update the probabilities when new information arrives.

Example (medical testing): Let A be the event "disease" and B be "positive test".

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}.$$

Although tests may be accurate, rare events can still produce many false positives—a key idea in decision making and evidence interpretation.

Bernoulli and Binomial Distributions

A **Bernoulli** random variable takes values in $\{0, 1\}$ with:

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p.$$

It models a yes/no outcome: click/no-click, vote/non-vote, success/failure.

The expected value and variance are the following.

$$E[X] = p, \qquad \text{Var}(X) = p(1 - p).$$

Binomial Distribution

The sum of n independent Bernoulli trials with success probability p produces a **binomial** random variable:

$$S_n = X_1 + \cdots + X_n \sim \text{Binomial}(n, p).$$

Its probability mass function is the following.

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \ldots, n.$$

The expected value and variance are the following.

$$E[S_n] = np, \qquad \text{Var}(S_n) = np(1-p).$$

Application: Polling and Vote Shares

In a poll of n respondents, let $X_i = 1$ if respondent i supports a candidate and 0 otherwise. Then:

$$X_i \sim \text{Bernoulli}(p), \qquad S_n = \sum X_i \sim \text{Binomial}(n, p).$$

The sample proportion

$$\hat{p} = \frac{S_n}{n}$$

estimates the true support rate p.

Normal Distribution and the Central Limit Theorem

The **normal distribution** is defined by:

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

The standard normal has $\mu = 0$ and $\sigma = 1$.

A remarkable fact is that the distribution of many averages approaches a normal distribution even when the underlying variables are not normal.

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. with mean μ and variance σ^2. Then as $n \to \infty$,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \;\Rightarrow\; \mathcal{N}(0, 1).$$

This explains why sample means, polling proportions, and many aggregated statistics behave approximately normally for large n.

Example: Polling Margin of Error

For a sample of size n, the approximate standard error of $\hat{p}$ is:

$$\text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

A 95% confidence interval for p is:

$$\hat{p} \pm 1.96 \cdot \text{SE}(\hat{p}).$$

This is widely used in election reporting: for $n = 1000$ and $\hat{p} = 0.50$, the margin of error is approximately:

$$1.96 \times \sqrt{\frac{0.5(0.5)}{1000}} \approx 0.031.$$

Thus, the reported support would be $50\% \pm 3.1\%$.

Summary

Probability theory formalizes uncertainty and provides distributional models for binary outcomes (Bernoulli, binomial), continuous phenomena (normal), and aggregated statistics (CLT). These foundations support practical applications in estimation, hypothesis testing, regression modeling, and election forecasting.