1. Linear Algebra

Linear algebra is a branch of mathematics that deals with vectors, vector spaces, linear transformations, and systems of linear equations. It is a foundational area of mathematics with applications in various fields including science, engineering, computer science, economics, and more.

Applications of Linear Algebra:

- **Computer Graphics**: Used for rendering images, transformations, and animations.
- **Machine Learning**: Fundamental for algorithms, data representation, and optimizations.
- **Physics**: Essential for quantum mechanics, relativity, and other fields.
- **Engineering**: Applied in systems theory, signal processing, and control systems.
- **Economics**: Used in modelling economic systems, optimization problems, and statistical analysis.

Applications of Linear Algebra in AI

Linear algebra is fundamental to many aspects of artificial intelligence (AI). Here are some key applications:

1. Machine Learning

- **Data Representation**:
  - Datasets are often represented as matrices where each row is a data sample and each column is a feature.
  - Vectors are used to represent individual data points in feature space.
- **Principal Component Analysis (PCA)**:
  - PCA is a dimensionality reduction technique that uses eigenvectors and eigenvalues of the covariance matrix to transform data into a new coordinate system. This is useful for reducing the number of features while retaining the most important information.
- **Linear Regression**:
  - Linear regression models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The solution involves solving a system of linear equations.
- **Support Vector Machines (SVM)**:
  - SVMs are used for classification and regression tasks. They find the optimal hyperplane that separates data into different classes. The optimization problem solved in SVMs heavily relies on linear algebra.

- **Neural Networks**:
    - Operations in neural networks, such as feedforward and backpropagation, involve matrix multiplications and other linear algebra operations.
    - Weight matrices, bias vectors, and input vectors are key components of neural networks.

2. Computer Vision

- **Image Processing**:
    - Images are represented as matrices (2D arrays for grayscale, 3D arrays for color images). Operations like filtering, transformations, and edge detection involve matrix operations.
- **Convolutional Neural Networks (CNNs)**:
    - CNNs use convolution operations, which are linear operations, to process data with a grid-like topology, such as images. This involves matrix multiplications and convolutions.
- **Eigenfaces for Face Recognition**:
    - Eigenfaces use PCA to reduce the dimensionality of face images and represent them in a lower-dimensional space, making face recognition more efficient.

3. Natural Language Processing (NLP)

- **Word Embeddings**:
    - Techniques like Word2Vec and GloVe represent words as vectors in a high-dimensional space. These vectors capture semantic meanings and relationships between words.
- **Transformers**:
    - Transformers, which are the backbone of models like BERT and GPT, use matrix multiplications for their attention mechanisms and transformations.
- **Latent Semantic Analysis (LSA)**:
    - LSA is a technique used to analyze relationships between a set of documents and the terms they contain. It uses singular value decomposition (SVD) to reduce the dimensionality of term-document matrices.

4. Robotics

- **Kinematics and Dynamics**:
    - Robot motion and control involve solving systems of linear equations to determine joint angles, velocities, and forces.

- **Path Planning**:
    - Algorithms for finding the optimal path often use linear algebra to represent and solve spatial relationships and constraints.

5. Optimization

- **Gradient Descent**:
    - Many AI algorithms, including those for training neural networks, use gradient descent to minimize loss functions. This involves linear algebra operations to update parameter vectors.
- **Convex Optimization**:
    - Many problems in AI can be formulated as convex optimization problems, which involve linear algebra to find the optimal solutions.

6. Graph Theory and Network Analysis

- **Graph Representations**:
    - Graphs are represented using adjacency matrices or incidence matrices. Analyzing properties of graphs, such as connectivity and shortest paths, involves linear algebra.
- **Spectral Clustering**:
    - Spectral clustering uses eigenvalues and eigenvectors of the Laplacian matrix of a graph to perform dimensionality reduction before clustering.

Linear algebra provides the mathematical foundation for representing and manipulating data in AI. Its concepts and techniques are integral to many AI algorithms and applications, making it an essential tool for anyone working in the field of artificial intelligence.

Key Concepts in Linear Algebra:

1. **Vectors**:
    - Objects that have both magnitude and direction.
    - Can be represented as ordered lists of numbers (coordinates).
    - Can be added together and multiplied by scalars to produce new vectors.
2. **Vector Spaces**:
    - A set of vectors that can be added together and multiplied by scalars while still remaining within the set.
    - Must satisfy certain axioms such as closure under addition and scalar multiplication, the existence of an additive identity (zero vector), and the existence of additive inverses.

3. **Linear Transformations**:
    - Functions that map vectors to vectors in a way that preserves vector addition and scalar multiplication.
    - Can be represented by matrices, which are rectangular arrays of numbers.
4. **Matrices**:
    - Rectangular arrays of numbers that represent linear transformations.
    - Can be added, multiplied, and subjected to various operations (such as finding determinants and inverses).
5. **Systems of Linear Equations**:
    - Collections of linear equations involving the same set of variables.
    - Can be represented in matrix form and solved using various methods (e.g., Gaussian elimination, matrix inversion).
6. **Eigenvalues and Eigenvectors**:
    - Eigenvectors are vectors that only get scaled (not rotated) when a linear transformation is applied.
    - Eigenvalues are the scalars by which the eigenvectors are scaled.
7. **Inner Product Spaces**:
    - Vector spaces equipped with an inner product, which allows for the definition of angles and lengths (norms).

Fundamental Theorems and Concepts:

- **Rank-Nullity Theorem**: Relates the dimensions of the kernel (null space) and the image (column space) of a matrix.
- **Spectral Theorem**: Pertains to the eigenvalues and eigenvectors of certain types of matrices.
- **Singular Value Decomposition (SVD)**: Factorizes a matrix into three other matrices, revealing important properties about the original matrix.
- **Gram-Schmidt Process**: An algorithm for orthonormalizing a set of vectors in an inner product space.

Linear algebra provides the language and tools for much of modern mathematics and its applications, making it an indispensable area of study.

1.1 Vectors

1.1.1 Definition and Notation

- **Definition**:
  - A vector is a mathematical object that has both magnitude (or length) and direction. Vectors are often represented in a coordinate system by an ordered set of numbers.
  - In two dimensions, a vector **v** can be written $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, as , where $v1$ and $v2$ are the components of the vector.
  - In three dimensions, a vector **v** can be written as $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$.
- **Notation**:
  - Vectors are typically denoted by boldface letters (e.g., **v**) or with an arrow above the letter (e.g., $\vec{v}$).
  - The magnitude (or length) of a vector **v** is denoted by |**v**| or ‖**v**‖ and is calculated as $\sqrt{v_1^2 + v_2^2}$ in 2D or $\sqrt{v_1^2 + v_2^2 + v_3^2}$ in 3D.

1.1.2 Vector Addition and Subtraction

- **Vector Addition**:
  - Vectors are added component-wise. If $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, then their sum $w=u+v$ is $\mathbf{w} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \end{pmatrix}$. .
  - Geometrically, vector addition corresponds to placing the tail of the second vector at the head of the first vector and drawing a new vector from the tail of the first vector to the head of the second vector.
- **Vector Subtraction**:
  - Vectors are subtracted component-wise. If $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, then their difference $w=u-v$ is $\mathbf{w} = \begin{pmatrix} u_1 - v_1 \\ u_2 - v_2 \end{pmatrix}$. .
  - Geometrically, vector subtraction can be visualized as the addition of u and the negative of v, i.e., $u-v=u+(-v)$.

1.1.3 Scalar Multiplication

- **Definition**:
  - Scalar multiplication involves multiplying a vector by a scalar (a real number). If c is a scalar and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, then $c\mathbf{v} = \begin{pmatrix} cv_1 \\ cv_2 \end{pmatrix}$.

1.1.4 Dot Product and Its Properties

- **Definition**:
  - The dot product (or scalar product) of two vectors u and v is a scalar given by $u \cdot v = u1v1 + u2v2$ in two dimensions.
  - In three dimensions, the dot product is $u \cdot v = u1v1 + u2v2 + u3v3$.
- **Properties**:
  - **Commutativity**: $u \cdot v = v \cdot u$.
  - **Distributivity**: $u \cdot (v+w) = u \cdot v + u \cdot w$.
  - **Scalar Multiplication**: $(cu) \cdot v = c(u \cdot v)$.
  - **Orthogonality**: Two vectors are orthogonal (perpendicular) if their dot product is zero: $u \cdot v = 0$.
  - **Magnitude**: The dot product of a vector with itself gives the square of its magnitude: $v \cdot v = |v|^2$.
- **Geometric Interpretation**:
  - The dot product can be used to find the angle θ between two vectors: $u \cdot v = |u||v| \cos\theta$.
  - If $u \cdot v > 0$, the angle between u and v is acute.
  - If $u \cdot v < 0$, the angle between u and **v** is obtuse.

These concepts form the foundational elements of vector algebra and are crucial for understanding more advanced topics in linear algebra and vector calculus.

1.2 Matrices

1.2.1 Definition and Notation

- **Definition**:
  - A matrix is a rectangular array of numbers arranged in rows and columns. Each number in the matrix is called an element or entry.
  - The size of a matrix is defined by the number of rows and columns it contains, denoted as m×n, where m is the number of rows and n is the number of columns.
- **Notation**:
  - Matrices are typically denoted by uppercase boldface letters (e.g., **A, B**).
  - An element in the ith row and jth column of matrix A is denoted as aij.

- **Example**:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

1.2.2 Matrix Addition and Subtraction

- **Matrix Addition**:
  - Two matrices of the same size can be added by adding their corresponding elements.
  - If A and B are both m×n matrices, their sum C=A+B is an m×n matrix where cij=aij+bij .
- **Matrix Subtraction**:
  - Similar to addition, two matrices of the same size can be subtracted by subtracting their corresponding elements.
  - If A and B are both m×n matrices, their difference0 C=A-B is an m×n matrix where cij=aij-bij .
- **Example**:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} 1-5 & 2-6 \\ 3-7 & 4-8 \end{pmatrix} = \begin{pmatrix} -4 & -4 \\ -4 & -4 \end{pmatrix}$$

1.2.3 Scalar Multiplication

- **Definition**:
  - Scalar multiplication involves multiplying every element of a matrix by a scalar (a real number).
  - If c is a scalar and A is an m×n matrix, the result B = cA is also an m×n matrix where bij=c·aij .
- **Example**:

$$c = 2, \quad \mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$c\mathbf{A} = 2 \cdot \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 6 & 8 \end{pmatrix}$$

1.2.4 Matrix Multiplication and Properties

- **Matrix Multiplication**:
  - The product of two matrices A of size m×n and B of size n×p is another matrix C of size m×p.
  - The element cij in C is calculated as the dot product of the i-th row of A and the j-th column of B

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

- **Properties**:
  - **Associativity**: A(BC)=(AB)C
  - **Distributivity**: A(B+C)=AB+AC
  - **Identity Matrix**: AI=A and IA=A, where I is the identity matrix with 1's on the diagonal and 0's elsewhere.
  - **Non-Commutativity**: In general, AB≠BA
- **Example**:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 2 & 0 \\ 1 & 3 \end{pmatrix}$$

$$\mathbf{C} = \mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} (1 \cdot 2 + 2 \cdot 1) & (1 \cdot 0 + 2 \cdot 3) \\ (3 \cdot 2 + 4 \cdot 1) & (3 \cdot 0 + 4 \cdot 3) \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 10 & 12 \end{pmatrix}$$

1.3 Matrix Operations

1.3.1 Transpose of a Matrix

- **Definition**:
  - The transpose of a matrix A is denoted by A^T. It is obtained by flipping the matrix over its diagonal, i.e., the element at position (i,j) in A is placed at position (j,i) in A^T.
- **Notation**:
  - If A is an m×n matrix, then A^T is an n×m matrix.
- **Example**:

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

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

1.3.2 Inverse of a Matrix

- **Definition**:
    - The inverse of a square matrix A is denoted by A^{-1}. It is the matrix that, when multiplied with A, yields the identity matrix I.
    - A matrix must be square (n×n) and have a non-zero determinant to have an inverse.
- **Notation**:
    - $A \cdot A^{-1} = I$
- **Example**:

$$\mathbf{A} = \begin{pmatrix} 4 & 7 \\ 2 & 6 \end{pmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{(4 \cdot 6 - 7 \cdot 2)} \begin{pmatrix} 6 & -7 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 0.6 & -0.7 \\ -0.2 & 0.4 \end{pmatrix}$$

1.3.3 Determinant of a Matrix

- **Definition**:
    - The determinant is a scalar value that can be computed from the elements of a square matrix. It provides important properties of the matrix, such as whether it is invertible (a non-zero determinant means the matrix is invertible).
- **Notation**:
    - The determinant of matrix A is denoted as det(A) or |A|.
- **Example**:
    - For a 2×2 matrix:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\det(\mathbf{A}) = ad - bc$$

- For a 3×3 matrix:

$$\mathbf{A} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

$$\det(\mathbf{A}) = a(ei - fh) - b(di - fg) + c(dh - eg)$$

1.3.4 Eigenvalues and Eigenvectors

- **Definition**:
  - Eigenvalues and eigenvectors are properties of a square matrix. For a given square matrix A, an eigenvector v is a non-zero vector that, when the matrix is multiplied by v, results in a scalar multiple of v. This scalar is called the eigenvalue **λ**.
- **Notation**:
  - **Av = λv**
  - **v** is the eigenvector, and **λ** is the eigenvalue.
- **Example**:
  - For the matrix:

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

  - To find the eigenvalues, solve the characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$:

$$\det \begin{pmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{pmatrix} = 0$$

$$(4-\lambda)(3-\lambda) - 2 = \lambda^2 - 7\lambda + 10 = 0$$

  - The solutions λ are the eigenvalues.

- **Properties**:
  - Eigenvalues can be real or complex numbers.
  - The number of eigenvalues (counting multiplicities) of an n×n matrix is n.
  - Eigenvectors corresponding to distinct eigenvalues are linearly independent.

1.4 Applications in Machine Learning

1.4.1 Linear Transformations

- **Definition**:
  - A linear transformation is a mapping between two vector spaces that preserves the operations of vector addition and scalar multiplication. In the context of

machine learning, linear transformations are often represented as matrix multiplications.

- **Application**:
  - Linear transformations are fundamental in neural networks, where weights of the layers apply linear transformations to the input data.
  - In feature engineering, linear transformations can be used to project data into different spaces for better analysis or to meet certain algorithmic requirements.
- **Example**:
  - If A is a matrix representing a linear transformation and x is a vector, then $y=Ax$ represents the transformed vector y.

1.4.2 Principal Component Analysis (PCA)

- **Definition**:
  - PCA is a statistical technique used for dimensionality reduction. It transforms the data to a new coordinate system such that the greatest variances of the data are projected onto the first coordinates (principal components).
- **Application**:
  - PCA is widely used for reducing the dimensionality of large datasets while preserving as much variability as possible. This is useful for visualization, noise reduction, and improving the efficiency of other algorithms.
  - In image processing, PCA can be used to compress images.
- **Steps in PCA**:
  1. Standardize the data.
  2. Compute the covariance matrix of the data.
  3. Compute the eigenvalues and eigenvectors of the covariance matrix.
  4. Sort the eigenvectors by decreasing eigenvalues and select the top kkk eigenvectors.
  5. Transform the original data using the selected eigenvectors to get the principal components.

1.4.3 Singular Value Decomposition (SVD)

- **Definition**:
  - SVD is a factorization of a real or complex matrix. It generalizes the “eigendecomposition” of a square matrix to any $m \times n$ matrix. SVD expresses the matrix as the product of three matrices: $A=U\Sigma V^T$ where:
    - U is an $m \times m$ orthogonal matrix.

- Σ is an m×n diagonal matrix with non-negative real numbers on the diagonal.
- V is an n×n orthogonal matrix.

- **Application**:
  - SVD is used in dimensionality reduction, noise reduction, and data compression. In recommendation systems, SVD is employed to predict user preferences.
  - SVD is also used in image compression and in solving linear inverse problems.
- **Steps in SVD**:
  1. Compute the SVD of the matrix A.
  2. Use the singular values to determine the rank and to compress the matrix.
  3. Reconstruct the matrix using a reduced number of singular values to approximate the original matrix.

---

2. Calculus

2.1 Derivatives

Derivatives are a fundamental concept in calculus that describe the rate at which a function is changing at any given point. They are essential in understanding and modelling change in various contexts, from physics to economics to machine learning.

2.1.1 Definition and Notation

- **Definition**:
  - The derivative of a function f(x) at a point x is defined as the limit of the average rate of change of the function over an interval as the interval approaches zero. Formally, the derivative f′(x) is given by:

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

- **Notation**:
  - $f'(x)$: Lagrange's notation
  - $\frac{d}{dx} f(x)$: Leibniz's notation
  - $D_x f(x)$: Euler's notation
  - For higher-order derivatives, $f''(x)$, $f^{(n)}(x)$, or $\frac{d^n}{dx^n} f(x)$ are used.

2.1.2 Basic Differentiation Rules

- **Sum Rule**:
  - The derivative of the sum of two functions is the sum of their derivatives.
  - If $f(x) = g(x) + h(x)$, then $f'(x) = g'(x) + h'(x)$.
- **Product Rule**:
  - The derivative of the product of two functions is given by:

$$(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$$

- **Quotient Rule**:
  - The derivative of the quotient of two functions is given by:

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{[g(x)]^2}$$

2.1.3 Chain Rule

- **Definition**:
  - The chain rule is used to differentiate composite functions. If $y = f(u)$ and $u = g(x)$, then the derivative of y with respect to x is:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

  - In Leibniz's notation, if $y = f(g(x))$, then:

$$\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x)$$

2.1.4 Higher-Order Derivatives

- **Definition**:
  - Higher-order derivatives are derivatives of derivatives. The second derivative, denoted as $f''(x)$ or $\frac{d^2}{dx^2} f(x)$, describes the rate of change of the first derivative.
  - Similarly, the third derivative $f'''(x)$ or $\frac{d^3}{dx^3} f(x)$, and so on for higher orders.
- **Notation**:
  - The n-th derivative is denoted as $f^{(n)}(x)$ or $\frac{d^n}{dx^n} f(x)$.

2.2 Partial Derivatives

2.2.1 Definition and Notation

Definition: Partial derivatives are a type of derivative where we hold all variables except one constant, and then find the derivative with respect to the one variable that is changing. If we have a multivariable function f(x,y,z,…), the partial derivative of f with respect to x is denoted as ∂f / ∂x.

Notation:

- For a function $f(x, y)$, the partial derivatives are denoted as:
  - $\frac{\partial f}{\partial x}$ or f_x for the derivative with respect to x.
  - $\frac{\partial f}{\partial y}$ or f_y for the derivative with respect to y.

2.2.2 Gradient Vectors

Definition:
The gradient of a function f is a vector that contains all of its partial derivatives. It points in the direction of the greatest rate of increase of the function.

Notation:

- For a function $f(x, y, z, \ldots)$, the gradient is denoted as ∇f or $\operatorname{grad} f$ and is defined as:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}, \ldots \right)$$

Example:
If $f(x, y) = x^2 + y^2$, the gradient is:

$$\nabla f = \left(\frac{\partial f}{\partial x} \frac{\partial f}{\partial y} \right) = (2x, 2y)$$

2.2.3 Jacobian and Hessian Matrices

Jacobian Matrix: The Jacobian matrix is a matrix of all first-order partial derivatives of a vector-valued function. It generalizes the gradient to vector-valued functions.

Hessian Matrix: The Hessian matrix is a square matrix of second-order partial derivatives of a scalar-valued function. It is used in optimization to describe the local curvature of a function.

2.3 Integrals

2.3.1 Definition and Notation

Definition: Integration is a fundamental concept in calculus that represents the accumulation of quantities, such as areas under a curve, total distance travelled given a velocity function, or the mass of an object with a varying density. There are two main types of integrals: definite and indefinite.

Indefinite Integral:

The indefinite integral of a function $f(x)$ is a function $F(x)$ such that $F'(x) = f(x)$. It is denoted as $\int f(x)\,dx = F(x) + C$, where C is the constant of integration.

Definite Integral:

The definite integral of a function $f(x)$ from a to b is the limit of a sum of areas of rectangles approximating the area under the curve $f(x)$ between $x = a$ and $x = b$. It is denoted as $\int_a^b f(x)\,dx$ and can be evaluated using the Fundamental Theorem of Calculus:

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

where F is an antiderivative of f, i.e., $F'(x) = f(x)$.

2.3.2 Basic Integration Rules

Some basic rules for integration include:

1. **Power Rule:**

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C \quad (\text{for } n \neq -1)$$

2. **Constant Multiple Rule:**

$$\int k \cdot f(x)\,dx = k \int f(x)\,dx$$

3. **Sum Rule:**

$$\int [f(x) + g(x)]\,dx = \int f(x)\,dx + \int g(x)\,dx$$

4. **Difference Rule:**

$$\int [f(x) - g(x)]\,dx = \int f(x)\,dx - \int g(x)\,dx$$

5. **Integration by Parts:**

$$\int u\,dv = uv - \int v\,du$$

6. **Substitution Rule:**

$$\int f(g(x))g'(x)\,dx = \int f(u)\,du \quad (\text{where } u = g(x))$$

---

3. Probability and Statistics

3.1 Descriptive Statistics

Descriptive statistics summarize and describe the main features of a dataset. They provide simple summaries about the sample and the measures. Descriptive statistics are broken down into measures of central tendency, measures of dispersion, and other statistical metrics.

3.1.1 Measures of Central Tendency

Measures of central tendency describe the centre of a dataset. The three main measures are:

- **Mean:** The arithmetic average of a dataset.

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where n is the number of observations, and x_i is each individual observation.

- **Median:** The middle value of a dataset when it is ordered from least to greatest. If the dataset has an even number of observations, the median is the average of the two middle numbers.

- **Mode:** The value that appears most frequently in a dataset. A dataset may have one mode, more than one mode, or no mode at all.

3.1.2 Measures of Dispersion

Measures of dispersion describe the spread of data points in a dataset. The main measures are:

- **Variance:** The average of the squared differences from the mean.

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \text{Mean})^2$$

- **Standard Deviation:** The square root of the variance, providing a measure of dispersion in the same units as the original data.

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

- **Range:** The difference between the highest and lowest values in a dataset.

$$\text{Range} = \text{Max} - \text{Min}$$

3.1.3 Skewness and Kurtosis

- **Skewness:** Measures the asymmetry of the data distribution. A skewness value can be positive or negative, indicating the direction of the skew.
  - **Positive Skew:** Right tail is longer; mean > median.
  - **Negative Skew:** Left tail is longer; mean < median.
- **Kurtosis:** Measures the "tailedness" of the data distribution. It describes the shape of the distribution's tails in relation to a normal distribution.
  - **Leptokurtic (positive kurtosis):** Distribution has heavier tails and a sharper peak.
  - **Platykurtic (negative kurtosis):** Distribution has lighter tails and a flatter peak.
  - **Mesokurtic:** Distribution is similar to a normal distribution.

3.1.4 Visualization Techniques

- **Histograms:** Graphical representations of the distribution of numerical data using bars of different heights.
- **Box Plots:** Summarize the distribution of a dataset using a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

3.2 Probability Theory

Probability theory is the branch of mathematics that deals with the analysis of random events. It provides a framework for quantifying the likelihood of various outcomes.

3.2.1 Basic Concepts

- **Sample Space (S):** The set of all possible outcomes of a random experiment. For example, when flipping a coin, the sample space is $S = \{\text{Heads}, \text{Tails}\}$.
- **Event (E):** A subset of the sample space. An event can consist of one or more outcomes. For example, getting a head when flipping a coin is an event.

3.2.2 Probability Axioms and Rules

- **Axiom 1:** The probability of any event E is a non-negative number: $P(E) \geq 0$.
- **Axiom 2:** The probability of the entire sample space is 1: $P(S)=1$.
- **Axiom 3:** If $E1, E2, \ldots, En$ are mutually exclusive events, then the probability of their union is the sum of their probabilities:

$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$$

Probability Rules:

- **Complement Rule:** The probability of the complement of an event E (i.e., the event that E does not occur) is:

$$P(E^c) = 1 - P(E)$$

- **Addition Rule:** For any two events E and F:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

- **Multiplication Rule:** For any two events E and F:

$$P(E \cap F) = P(E) \cdot P(F|E)$$

3.2.3 Conditional Probability and Bayes' Theorem

- **Conditional Probability:** The probability of event E given that event F has occurred is:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}, \quad \text{provided } P(F) > 0$$

- **Bayes' Theorem:** Relates the conditional and marginal probabilities of random events. It is expressed as:

$$P(E|F) = \frac{P(F|E) \cdot P(E)}{P(F)}, \quad \text{provided } P(F) > 0$$

3.2.4 Independence and Dependence of Events

- **Independence:** Two events E and F are independent if the occurrence of one does not affect the probability of the other. Mathematically, E and F are independent if:

$$P(E \cap F) = P(E) \cdot P(F)$$

- **Dependence:** If the occurrence of one event affects the probability of the other, the events are dependent. Mathematically, E and F are dependent if:

$$P(E \cap F) \neq P(E) \cdot P(F)$$

3.3 Probability Distributions

Probability distributions describe how the probabilities of different outcomes are distributed for a random variable.

3.3.1 Discrete Distributions

- **Bernoulli Distribution:**
  - Represents the outcome of a single binary experiment (e.g., coin flip).
  - The random variable X can take value 1 (success) with probability p and value 0 (failure) with probability 1 − p.
  - Probability mass function (PMF): $P(X = x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$. .
- **Binomial Distribution:**
  - Represents the number of successes in n independent Bernoulli trials.
  - Parameters: number of trials n and probability of success p.
  - PMF: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$.
- **Poisson Distribution:**
  - Represents the number of events occurring in a fixed interval of time or space.
  - Parameter: λ (average rate of occurrence).
  - PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0, 1, 2, \ldots$.

3.3.2 Continuous Distributions

- **Uniform Distribution:**
  - All outcomes in a continuous range are equally likely.
  - Parameters: lower bound a and upper bound b.
- **Normal (Gaussian) Distribution:**
  - Represents the distribution of many natural phenomena.
  - Parameters: mean μ and standard deviation σ.
- **Exponential Distribution:**
  - Represents the time between events in a Poisson process.
  - Parameter: rate λ.

3.3.3 Properties and Moments

- Mean (Expected Value):
  - Discrete: $E(X) = \sum_x xP(X = x)$
  - Continuous: $E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$
- Variance:
  - Discrete: $\mathrm{Var}(X) = \sum_x (x - E(X))^2 P(X = x)$
  - Continuous: $\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,dx$
- Skewness:
  - Measure of asymmetry of the distribution.
  - $\text{Skewness} = \frac{E[(X-\mu)^3]}{\sigma^3}$
- Kurtosis:
  - Measure of the "tailedness" of the distribution.
  - $\text{Kurtosis} = \frac{E[(X-\mu)^4]}{\sigma^4}$

3.3.4 Applications in ML

- **Gaussian Mixture Models (GMM):**
  - A probabilistic model representing a mixture of multiple Gaussian distributions.
  - Useful for clustering and density estimation.
  - Parameters are estimated using the Expectation-Maximization (EM) algorithm.
- **Naive Bayes:**
  - A classification technique based on Bayes' theorem.
  - Assumes independence between features.
  - Commonly used with categorical data and text classification.

3.4 Inferential Statistics

Inferential statistics allows us to make predictions or inferences about a population based on a sample of data drawn from that population. It encompasses various techniques to analyse sample data and make generalizations about a larger population.

3.4.1 Sampling and Sampling Distributions

- **Sampling:** The process of selecting a subset of individuals from a population to estimate characteristics of the whole population.
  - **Simple Random Sampling:** Each member of the population has an equal chance of being selected.
  - **Stratified Sampling:** The population is divided into strata, and random samples are taken from each stratum.
  - **Cluster Sampling:** The population is divided into clusters, some clusters are randomly selected, and all members of selected clusters are sampled.
- **Sampling Distribution:** The probability distribution of a given statistic based on a random sample.
  - **Central Limit Theorem:** For sufficiently large sample sizes, the sampling distribution of the sample mean is approximately normally distributed, regardless of the distribution of the population.

3.4.2 Point Estimation and Interval Estimation

- **Point Estimation:** The use of sample data to calculate a single value (known as a point estimate) which serves as a "best guess" for an unknown population parameter.
- **Interval Estimation:** Provides a range of values (an interval) that is likely to contain the population parameter.
  - **Confidence Interval (CI):** A range of values, derived from the sample statistics, that is likely to contain the population parameter.

3.4.3 Hypothesis Testing

- **Hypothesis Testing:** A method of making decisions or inferences about population parameters based on sample data.
  - **Null Hypothesis (H_0):** The hypothesis that there is no effect or no difference; it is the default or starting assumption.
  - **Alternative Hypothesis (H_1):** The hypothesis that there is an effect or a difference.
- **p-value:** The probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis.
  - **Interpretation:** A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- **Test Statistics:** A standardized value used to determine the likelihood of the sample data given the null hypothesis. Examples include z-scores and t-scores.

3.4.4 Confidence Intervals and Significance Levels

- **Confidence Intervals (CI):**
  - A range of values used to estimate the true value of a population parameter.
  - A 95% confidence interval means that if the same population is sampled multiple times, approximately 95% of the calculated confidence intervals will contain the true population parameter.
- **Significance Levels (α\alphaα):**
  - The threshold for rejecting the null hypothesis, commonly set at 0.05.
  - If the p-value is less than α, the result is considered statistically significant, and the null hypothesis is rejected.