Math Basics

www.huawei.com

HUAWEI

Objectives

- After completing this course, you will be able to:
  - Master the basic knowledge and application of Linear Algebra.
  - Master the basic knowledge and application of Probability Theory and Information Theory.
  - Master numerical calculation functions, and the classification and solution of optimization problems.

Contents

1. **Linear Algebra**
   - **Concept and Calculation of Matrices**
   - Special Matrices
   - Eigendecomposition
2. Probability Theory and Information Theory
3. Numeric Calculation

Linear Algebra

- **Linear algebra** is a branch of algebra that mainly deals with linear problems. **Linear relationship** means that the relationship between mathematical objects is expressed in one form. The first problem to be solved by linear algebra is to solve **linear equations**.
- **Determinants** and **matrices** serve as powerful tools for dealing with linear problems and promote the development of linear algebra. The introduction of the concept of **vector** enables the vector space, and linear problems can be solved using the vector space. Vector space and its linear transformation, and related matrix theories, constitute the core of linear algebra.
- Linear algebra is characterized by a large number of variables and complex relationships. Its methods include careful logical reasoning, artful summarization, and complex and skillful numerical calculation.

Case (1)

- To avoid obesity and improve employees' health, the Big Data Department organized a monthly running activity at the beginning of 2018. The rules were as follows: The department set the monthly target for participants at the beginning of the month. The participants who fulfilled the targets would be rewarded while those who failed would be punished. The calculation rule of the reward or penalty amount was as follows:

$$w_i = (s_i - d_i)x_i = h_i x_i,$$

In the preceding equation, w_i is the total reward/penalty amount in the month i, s_i is the total mileage, d_i is the monthly target, h_i is the difference between the actual distance and monthly target, and x_i is the reward/penalty amount of each kilometer every month. This activity received good feedback and was later adopted by the Cloud Department. The following tables listed the difference between the actual distance and monthly target and total reward/penalty amount of some participants in the first quarter:

Month / Name	h_1	h_2	h_3	w
A	10	8	12	20
B	4	4	2	8
C	2	-4	-2	-5

Table 1 Big Data Department

Month / Name	h_1	h_2	h_3	w
D	2	4	5	10
E	4	2	2	6
F	-2	2	2	3

Table 2 Cloud Department

Case (2)

- In the preceding case, what is the reward/penalty amount set by the Big Data Department for each kilometer in each month? The equations are as follows using the given data:

$$\begin{cases} 10x_1 + 8x_2 + 12x_3 = 20 \\ 4x_1 + 4x_2 + 2x_3 = 8 \\ 2x_1 - 4x_2 - 2x_3 = -5 \end{cases} \quad (1.1)$$

In this way, the solutions of the equations are the answer to the question.

Scalar, Vector, and Matrix

- **Vector**

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \mathrm{M} \\ x_n \end{bmatrix}$$

- **Mtrix** The preceding is a M-row N-column matrix, which is denoted as:

$$\begin{matrix} a_{11} & a_{12} & \mathrm{L} & a_{1n} \\ a_{21} & a_{22} & \mathrm{L} & a_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ a_{m1} & a_{m2} & \mathrm{L} & a_{mn} \end{matrix}$$

The simple form is $A = A_{m\times n} = (a_{ij})_{m\times n} = (a_{ij})$. The special matrix whose row quantity and column quantity are both *n* is called n-order matrix.

$$A = \begin{bmatrix} a_{11} & a_{12} & \mathrm{L} & a_{1n} \\ a_{21} & a_{22} & \mathrm{L} & a_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ a_{m1} & a_{m2} & \mathrm{L} & a_{mn} \end{bmatrix}$$

Determinant

- **det(A)**

$$\det(A) = \begin{vmatrix} a_{11} & a_{12} & \mathrm{L} & a_{1n} \\ a_{21} & a_{22} & \mathrm{L} & a_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ a_{m1} & a_{m2} & \mathrm{L} & a_{mn} \end{vmatrix}.$$

- **Significance**
  - The determinant is equal to the product of all the eigenvalues of the matrix.
  - The absolute value of the determinant can be thought of as a measure of how much multiplication by the matrix expands or contracts space. If the determinant is 0, then space is contracted completely along at least one dimension, causing it to lose all its volume. If the determinant is 1, then the transformation preserves volume.

Matrix Operation

- **Matrix addition**:

  Suppose that $A = (a_{ij})_{s\times n}$ and $B = (b_{ij})_{s\times n}$ are s x n matrices, and the sum of the two matrices is $C = A + B = (a_{ij} + b_{ij})_{s\times n}$.

- **Scalar and matrix multiplication**:

  Suppose $A = (a_{ij})_{s\times n}$ and $k \in K$. The product of k and matrix A is $kA = (ka_{ij})_{s\times n}$. The addition of a scalar and matrix follows the same rule.

- **Matrix multiplication**:

  Suppose $A = (a_{ij})_{s\times n}$ and $B = (b_{ij})_{n\times p}$,

  $$C = AB = (c_{ij})_{s\times p},$$

  where $C_{i,j} = \sum_k A_{i,k}B_{k,j}$

Matrix Transposition

- **Transposed matrix:** A^T (also written as A′)

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \Rightarrow A^T=\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix}.$$

- **Nature of a transposed matrix:**
  - $(A^T)^T = A$
  - $(A+B)^T = A^T + B^T$
  - $(\lambda A)^T = \lambda A^T$
  - $(AB)^T = B^T A^T$

Trace Operator

- **Trace operator :**

$$Tr(A) = \sum_i A_{i,i}$$

- **Nature of a trace operator:**
  - $Tr(A) = Tr(A^T)$
  - $Tr(a) = a$
  - $Tr(ABC) = Tr(CAB) = Tr(BCA)$

Case Calculation

- In the preceding case, the calculation is as follows:
  - The running result of the big data department and cloud department in the first quarter can be calculated as follows:

$$A=\begin{bmatrix}10 & 8 & 12\\ 4 & 4 & 2\\ 2 & -4 & -2\end{bmatrix},\quad B=\begin{bmatrix}2 & 4 & 5\\ 4 & 2 & 2\\ -2 & 2 & 2\end{bmatrix}.$$

  - Perform the following operation on the matrices:

$$C_1=A+B=\begin{bmatrix}12 & 12 & 17\\ 8 & 6 & 4\\ 0 & -2 & 0\end{bmatrix},\quad C_2=2A=\begin{bmatrix}20 & 16 & 24\\ 8 & 8 & 4\\ 4 & -8 & -4\end{bmatrix},\quad C_3=AB=\begin{bmatrix}28 & 80 & 90\\ 20 & 28 & 32\\ -8 & -4 & -2\end{bmatrix}.$$

  - According to the matrix multiplication rule, the equations (1.1) can be represented by a matrix as follows:

$$\begin{cases}10x_1+8x_2+12x_3=20\\ 4x_1+4x_2+2x_3=8\\ 2x_1-4x_2-2x_3=-5\end{cases}\Rightarrow\begin{bmatrix}10 & 8 & 12\\ 4 & 4 & 2\\ 2 & -4 & -2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix}=\begin{bmatrix}20\\ 8\\ -5\end{bmatrix}\Rightarrow Ax=C.$$

Contents

1. **Linear Algebra**
   - Concept and Calculation of Matrices
   - **Special Matrices**
   - Eigendecomposition
2. Probability Theory and Information Theory
3. Numeric Calculation

Identity and Inverse Matrices

- **Identity matrix:**

$$I_n = \begin{bmatrix} 1 & 0 & \mathrm{L} & 0 \\ 0 & 1 & \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \mathrm{L} & 1 \end{bmatrix}.$$

- The **matrix inverse** of A is denoted as A^{-1}, and it is defined as the matrix such that $\mathrm{A}^{-1}\mathrm{A} = \mathrm{I_n}$.

Diagonal Matrix

- **Diagonal matrix:**

$$D=\begin{bmatrix} \lambda_1 & 0 & \mathrm{L} & 0 \\ 0 & \lambda_2 & \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \mathrm{M} & \lambda_n \end{bmatrix}.$$

- **Nature of a diagonal matrix:**
  - The sum, difference, product, and square power of the elements on the diagonal matrix are the sum, difference, product, and square power of the elements along the main diagonal.
  - The inverse matrix is as follows:

$$D^{-1}=\begin{bmatrix} \lambda_1^{-1} & 0 & \mathrm{L} & 0 \\ 0 & \lambda_2^{-1} & \mathrm{L} & 0 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & 0 & \mathrm{M} & \lambda_n^{-1} \end{bmatrix}.$$

Symmetric Matrix

- **Symmetric matrix:**

$$A=\begin{bmatrix} a_{11} & a_{12} & \mathrm{L} & a_{1n} \\ a_{12} & a_{22} & \mathrm{L} & a_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ a_{1n} & a_{2n} & \mathrm{L} & a_{nn} \end{bmatrix}.$$

- **Orthogonal matrix:**

$$AA^T = A^T A = I_n \Rightarrow A^{-1} = A^T.$$

Contents

1. **Linear Algebra**
   - Concept and Calculation of Matrices
   - Special Matrices
   - **Eigendecomposition**
2. Probability Theory and Information Theory
3. Numeric Calculation

HUAWEI

Eigendecomposition (1)

- One of the most widely used kinds of **matrix decomposition** is called eigendecomposition, in which we decompose a matrix into a set of **eigenvectors** and **eigenvalues**. We can decompose matrices in ways that show us information about their functional properties that is not obvious from the representation of the matrix as an array of elements.
- Suppose that A is a n-level matrix in the digital domain K. If there is a non-zero column vector α in K^n that meets the following:

$$A\alpha = \lambda\alpha, and\ \lambda \epsilon K,$$

λ is called an **eigenvalue** of A, and α is a **eigenvector** of A and belongs to the eigenvalue λ.

- Example: $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\alpha = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, 由于$A\alpha = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2\alpha.$

Therefore, 2 is an eigenvalue of A, and α is an eigenvector of A and belongs to eigenvalue 2.

Eigendecomposition (2)

- Obtaining the eigenvalues and eigenvectors of matrix A:

$$
\begin{aligned}
& A\alpha = \lambda\alpha \\
\Leftrightarrow \quad & A\alpha - \lambda\alpha = 0 \\
\Leftrightarrow \quad & (A - \lambda I)\alpha = 0 \\
\overset{\alpha \neq 0}{\Leftrightarrow} \quad & |A - \lambda I| = 0 \\
\Leftrightarrow \quad & \begin{vmatrix} a_{11} - \lambda & a_{12} & \mathrm{L} & a_{1n} \\ a_{21} & a_{22} - \lambda & \mathrm{L} & a_{2n} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ a_{n1} & a_{n2} & \mathrm{L} & a_{nn} - \lambda \end{vmatrix} = 0.
\end{aligned}
$$

In the preceding information, $|A - \lambda I| = 0$ is a feature equation of matrix A, λ is a solution (characteristic root) of the feature equation. To obtain the eigenvector α, substitute the characteristic root λ into $A\alpha = \lambda\alpha$.

Eigendecomposition (3)

- Example: Find the eigenvalues and eigenvectors of the matrix $A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$.

Solution: The characteristic polynomial of A is $\begin{vmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1 = (4-\lambda)(2-$

Eigendecomposition (4)

- Suppose that a matrix A has n linearly independent eigenvectors $\{\alpha_1, \ldots, \alpha_n\}$ with corresponding eigenvalues {λ1, . . . , λn}. The eigendecomposition of A is then given by

$$A = Pdiag(\lambda)P^{-1},$$

  where $P = \{\alpha_1, \alpha_2, \cdots, \alpha_n\}$, and $\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$.

- **Matrix accuracy:**
  - A matrix whose eigenvalues are all positive is called positive definite.
  - A matrix whose eigenvalues are all positive or zero valued is called positive semidefinite.
  - If all eigenvalues are negative, the matrix is negative definite.
  - If all eigenvalues are negative or zero valued, it is negative semidefinite.

Singular Value Decomposition

- The matrix is decomposed into singular vectors and singular values. The matrix $A = (a_{ij})_{m \times n}$ can be decomposed into a product of three matrices:

$$A = UDV^T,$$

Among $U = (b_{ij})_{m \times m}$, $D = (c_{ij})_{m \times n}$, and $V^T = (d_{ij})_{n \times n}$, the matrices *U* and *V* are both defined to be orthogonal matrices. The columns of *U* are known as the left-singular vectors. The columns of *V* are known as as the right-singular vectors. The matrix *D* is defined to be a diagonal matrix. Note that *D* is not necessarily square. Elements on the diagonal line of *D* is referred to as a singular value of the matrix.

Moore-Penrose Pseudoinverse

- The **Moore-Penrose pseudoinverse** enables us to make some headway in finding the solution of $Ax = y(A = (a_{ij})_{m\times n}, m \neq n)$. The pseudoinverse of A is defined as a matrix:

$$A^{+} = \lim_{\alpha \to 0} (A^{T}A + \alpha I)^{-1} A^{T},$$

  Practical algorithms for calculating the pseudoinverse are based on the formula:

$$A^{+} = VD^{+}U^{T},$$

- where *U*, *D* and *V* are the singular value decomposition of *A*, and the pseudoinverse D^{+} of a diagonal matrix *D* is obtained by taking the reciprocal of its non-zero elements then taking the transpose of the resulting matrix.

Example: Principal Component Analysis (1)

- **Principal Component Analysis (PCA)**: a statistical method. Through **orthogonal transform**, a group of variables that may have correlation relationships are converted into a set of linear unrelated variables, and the converted variables are called main components.
- **Basic principles:** Assume that there are n objects, and each object is composed of $\{x_1, \ldots, x_p\}$. The following table lists the factor data corresponding to each object.

Factor / Object	x_1	x_2	…	x_j	…	x_p
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
…	…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}

Example: Principal Component Analysis (2)

- The original variables are $x_1, \ldots, x_p$ After the dimension-reduction processing, set their comprehensive indexes. That is, the new variables are $z_1, \ldots, z_m$ $(m \leq p)$. $z_1, \ldots, z_m$ are called the first, the second, ..., the *m*th main component of $x_1, \ldots, x_p$. We have the following expression:

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \ldots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \ldots + l_{2p}x_p \\ \text{L L L L L L L L L L} \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \ldots + l_{mp}x_p \end{cases}.$$

- To obtain m principal components, the steps are as follows:
  - The coefficient l_{ij} meets the following rules: z_i is not related to $z_j (i \neq j; i, j = 1, 2, \ldots, m)$. z_1 has the largest variance among all linear combinations of $x_1, \ldots, x_p$. z_2 has the largest variance among all linear combinations of $x_1, \ldots, x_p$ that is not related to z_1. z_m has the largest variance among all linear combinations of $x_1, \ldots, x_p$ that is not related to $z_1, z_2, \ldots, z_{m-1}$.
  - According to the above rules, l_{ij} is a **eigenvector** of *m* **large eigenvalues** of the **coefficient matrix** corresponding to $x_1, \ldots, x_p$.
  - If the **cumulative contribution rate** of the first *i* main components reaches 85% to 90%, those components are used as the new variables.

Example: Principal Component Analysis (3)

- Correlation coefficient matrix and correlation coefficient:

$$R=\begin{bmatrix} r_{11} & r_{12} & \mathrm{L} & r_{1p} \\ r_{21} & r_{22} & \mathrm{L} & r_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ r_{p1} & r_{p2} & \mathrm{L} & r_{pp} \end{bmatrix}, \qquad r_{ij}=\frac{\sum_{k=1}^{p}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{p}(x_{ki}-\overline{x_i})^2\sum_{k=1}^{p}(x_{kj}-\overline{x_j})^2}}.$$

$||x||_2$

- Contribution rate of the main components and cumulative contribution rate:

$$Q_i=\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}(i=1,2,\mathrm{L}\ ,p), \qquad Q=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}(i=1,2,\mathrm{L}\ ,p).$$

Contents

1. Linear Algebra
2. **Probability Theory and Information Theory**
   - **Basic Concepts of Probability Theory**
   - Random Variables and Their Distribution Functions
   - Numerical Characteristics of Random Variables
   - Information Theory
3. Numeric Calculation

Why Do We Use Probability?

- While probability theory allows us to make uncertain statements and to reason in the presence of uncertainty, information theory enables us to quantify the amount of uncertainty in a probability distribution.
- There are three possible sources of uncertainty:
  - Inherent stochasticity in the system being modeled
  - Incomplete observability
  - Incomplete modeling

Random Test

- The test that meets the following three characteristics is called a **random test**:
  - It can be repeated under the same condition.
  - There may be more than one result of each test, and all possible results of the test can be specified in advance.
  - Before a test, we cannot determine which result will appear.
- Example:
  - E_1: Toss two coins and check the outcome (front or back).
  - E_2: Throw a dice and check the number of points that may appear.

Sample Point, Sample Space, and Random Variables Event

- **Sample point**: each possible result of a random test, which is represented by e.
- **Sample space**: a collection of all possible results of a random test, which is represented by $S = \{e_1, e_2, \dots, e_n\}$.
- **Random variables event**: any subset of the sample space *S*. If a sample point of event A occurs, event A occurs. In particular, a random event containing only one sample point is called a basic event.
- Example:

  Random test: Throw a dice and check the outcome.

  Sample space: $S = \{1, 2, 3, 4, 5, 6\}$

  Sample point: $e_i = 1, 2, 3, 4, 5, 6$

  Random event A_1: "The outcome is 5", that is, $A_1 = \{x | x = 5\}$.

Frequency and Probability

- **Frequency**: Under the same conditions, perform tests for n times. The occurrence of event A is called the frequency of event A. The ratio $\frac{n_A}{n}$, occurrence probability of event A, is recorded as $f_n(A)$.
- **Probability:** Suppose that E is a random test and S is the sample space. Assign a real number $\mathrm{P}(\mathrm{A})$ (event probability) on each event A of E. The set function $\mathrm{P}(*)$ must meet the following conditions:
  - Non-negative: For each event A, $0 \leq P(A) \leq 1$.
  - Standard: For the inevitable event S, $P(S) = 1$.
  - Countable additivity: $\{A_1, \ldots\}$ are events incompatible with each other. That is, if $A_i A_j = \emptyset, i \neq j, i, j = 1, 2, \cdots$, we have $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$.

Random Variable

- The **random variable** indicates a single- and real-valued function that represents a random test of various results.
- Example 1: Random test E_4: Toss two dice and check the sum of the results. The sample space of the test is $S = \{e\} = \{(i,j)|i,j = 1,2,3,4,5,6\}$. i indicates the first outcome and j indicates the second outcome. X is the sum of the two outcomes, which is a random variable.
- $X = X(e) = X(i,j) = i + j, i, j = 1, 2, \cdots, 6.$
- Example 2: Random test E_1: Throw two coins and check the outcome (front side H or back side T). The sample space for the test is $S = \{HH, HT, TH, TT\}$. Y, as the total occurrence of the back side T, is a random variable.

$$Y = Y(e) = \begin{cases} 0, e = HH, \\ 1, e = HT, TH, \\ 2, e = TT. \end{cases}$$

Contents

1. Linear Algebra
2. **Probability Theory and Information Theory**
   - **Basic Concepts of Probability Theory**
   - **Random Variables and Their Distribution Functions**
   - Numerical Characteristics of Random Variables
   - Information Theory
3. Numeric Calculation

Discrete Random Variables and Distribution Law

- **Discrete random variables:** All the values of random variables may be finite or infinite. A typical random variable is the number of vehicles passing through a monitoring gate within one minute.
- **Distribution law:** If all the possible values of discrete random variable X are $x_k(k = 1, 2, \cdots)$, the probability of X getting a possible value $\{X = x_k\}$ is:

$$P\{X = x_k\} = p_k, k = 1, 2, \cdots.$$

As defined for probability, p_k should meet the following conditions:

(1) $p_k \geq 0, k = 1, 2, \cdots.$

(2) $\sum_{k=1}^{\infty} p_k = 1.$

The distribution law can also be expressed in a table:

X	x_1	x_2	$\cdots$	x_n	$\cdots$
p_k	p_1	p_2	$\cdots$	p_n	$\cdots$

Special Distribution - Bernoulli Distribution

- **Bernoulli distribution (0-1 distribution, two-point distribution, a-b distribution):**

  If random variable X can be either 0 or 1, its distribution law is:

  $$P\{X = k\} = p^k(1-p)^{1-k}, k = 0, 1 \ \ (0 < p < 1),$$

  That is, X obeys Bernoulli distribution with the p parameter.

- The distribution law of Bernoulli distribution can also be written as below:

X	0	1
p_k	$1-p$	p

$E(X) = p$, $\mathrm{V}ar(X) = p(1-p)$.

Special Distribution - Binomial Distribution

- **n independent repetitive tests**: The experiment E is repeated n times. If the results of each experiment do not affect each other, the n experiments are said to be independent of each other.
- The experiments that meet the following conditions are called **n Bernoulli experiments**:
  - Each experiment is repeated under the same conditions.
  - There are only two possible results per experiment: A and $\bar{A}$ and $P(A) = p$.
  - The results of each experiment are independent of each other.

  If the times of event A occurring in n Bernoulli experiments are expressed by X, the probability of event A occurring for k times in n experiments is as below:

$$P(X = k) = C_n^k p^k (1-p)^{n-k}, k = 0, 1, 2, \cdots n,$$

  At this time, X obeys **binomial distribution** with n and p parameters. This is expressed as $X \sim B(n,p)$, where $E(X)$ equals np and $Var(x)$ equals $np(1-p)$.

Special Distribution - Poisson Distribution

- **Poisson theorem:** If λ>0 is set as a constant, n is any positive integer, and np equals λ, the following applies to any fixed non-negative integer k:

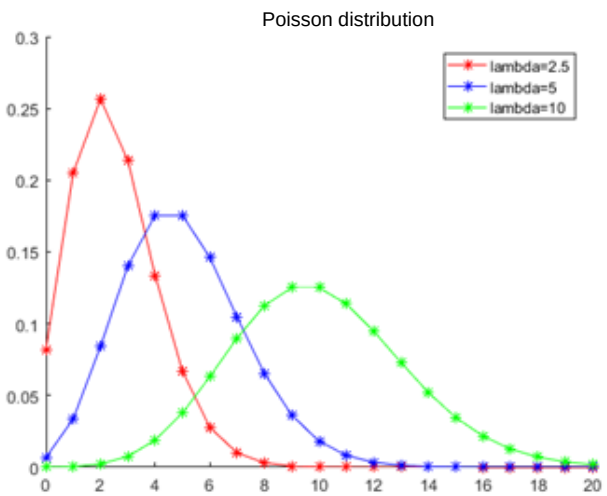
$$\lim_{n\to\infty} C_n^k p^k (1-p)^{n-k} \approx \frac{\lambda^k e^{-\lambda}}{k!}$$

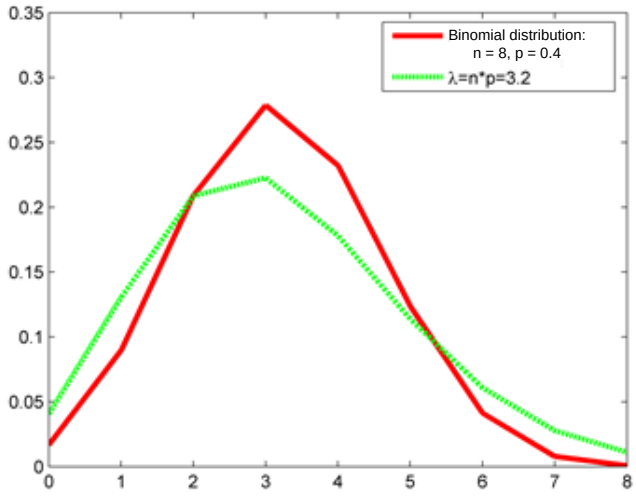
- **Poisson distribution:** If all possible values of random variables are 0, 1, 2, ..., the probability of taking each value is:

$$P\{X=k\} = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \quad \cdots$$

Then, X obeys Poisson distribution with parameter λ. It is expressed as $X \sim P(\lambda)$, where $E(X)$ equals λ, and $D(X)$ equals λ.

Association Between Poisson Distribution and Binomial Distribution





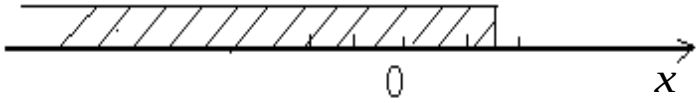
- The mathematical models of Poisson distribution and Binomial distribution are both Bernoulli-type. Poisson distribution has the appropriately equal calculation as binomial distribution when n is very large and p very small.

Distribution Function

- **Distribution function:** If X is a random variable, and x is an arbitrary real number, function $F(x)$ is called the distribution function of X.

$$F(x) = P\{X \le x\}, -\infty < x < \infty$$

  - Distribution function $F(x)$ has the following basic properties:
    - $F(x)$ is a function of no subtraction.
    - $0 \le F(x) \le 1$, and $F(-\infty) = \lim_{x \to -\infty} F(x) = 0, F(\infty) = \lim_{x \to \infty} F(x) = 1$.
    - $F(x+0) = F(x)$, that is, $F(x)$ is of right continuity.
  - Significance of distribution function $F(x)$: If X is regarded as the coordinate of a random point on the number axis, the function value of distribution function $F(x)$ at x indicates the probability that X falls in the interval $(-\infty, x)$.



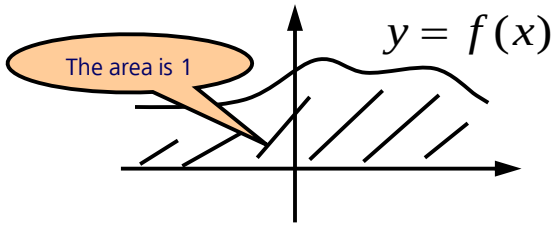
Continuous Random Variables and Probability Density Function

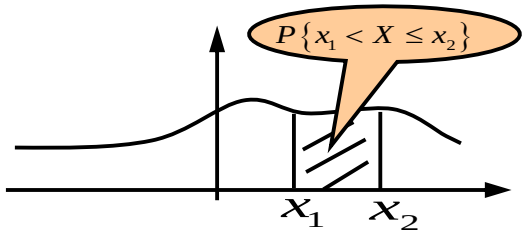
- If distribution function $F(x)$ for random variable X has a non-negative function $f(x)$, and the following applies to arbitrary real number x:

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

  Then, X is called a **continuous random variable**, and function $f(x)$ is called the **probability density function** of X, or probability density.

- Probability density $f(x)$has the following properties:
  - $f(x) \geq 0$.
  - $\int_{-\infty}^{+\infty} f(x)dx = 1$.
  - For arbitrary real number $x_1, x_2 (x_1 < x_2)$, $P\{x_1 < X \leq x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$.
  - If $f(x)$is continuous at x, $F'(x) = f(x)$.
  - The probability value of random variable X taking any real number is 0, that is, $P(X=a) = 0$.



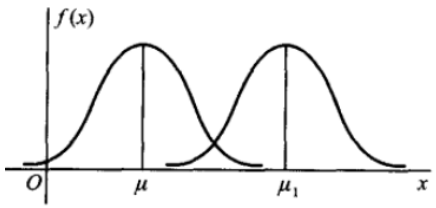


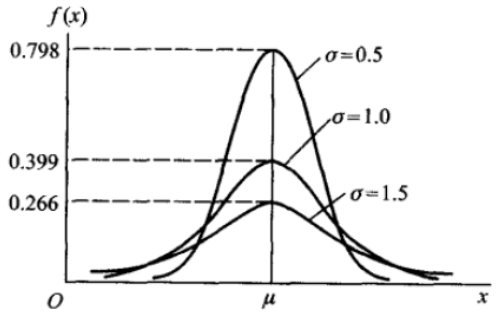
Special Distribution - Normal Distribution

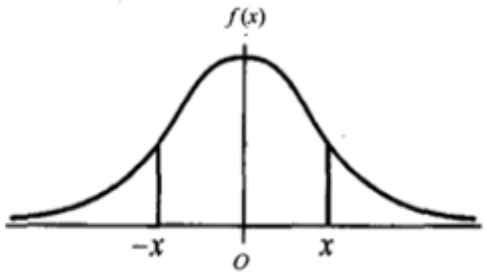
- If the probability density function of continuous random variable X is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty,$$

where μ, σ (σ>0) is constant, X obeys the normal distribution or Gaussian distribution of μ, σ, which is expressed as X~N (μ,σ^2). Especially when $\mu = 0$, $\sigma = 1$, random variable X obeys the standard normal distribution, which is expressed as X~N (0, 1).





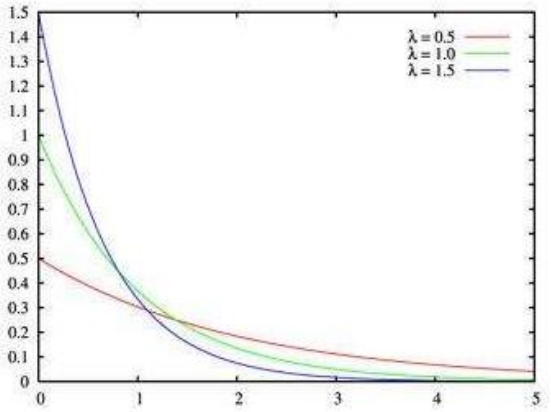


Special Distribution - Exponential Distribution

- If the probability density of continuous random variable X is

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, x>0\\ 0 \quad ,otherwise\end{cases}$$

where λ>0 is a constant, indicating the time when a random event occurs once, X obeys the exponential distribution with parameter λ. This distribution is expressed as $X\sim E(\lambda).E(X)=\frac{1}{\lambda},Var(X)=\frac{1}{\lambda^2}$.

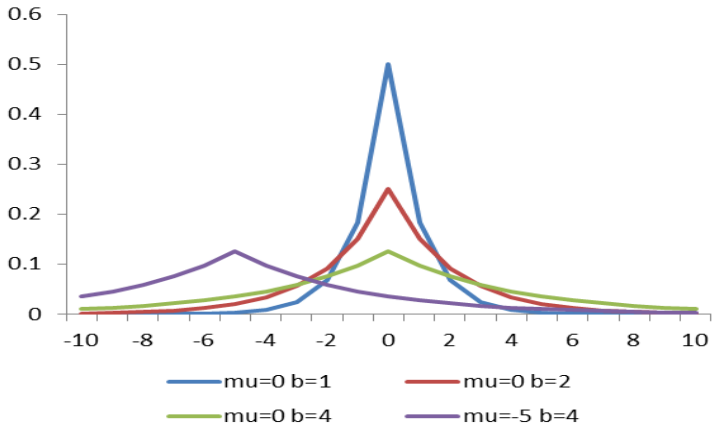


Special Distribution – Laplace Distribution

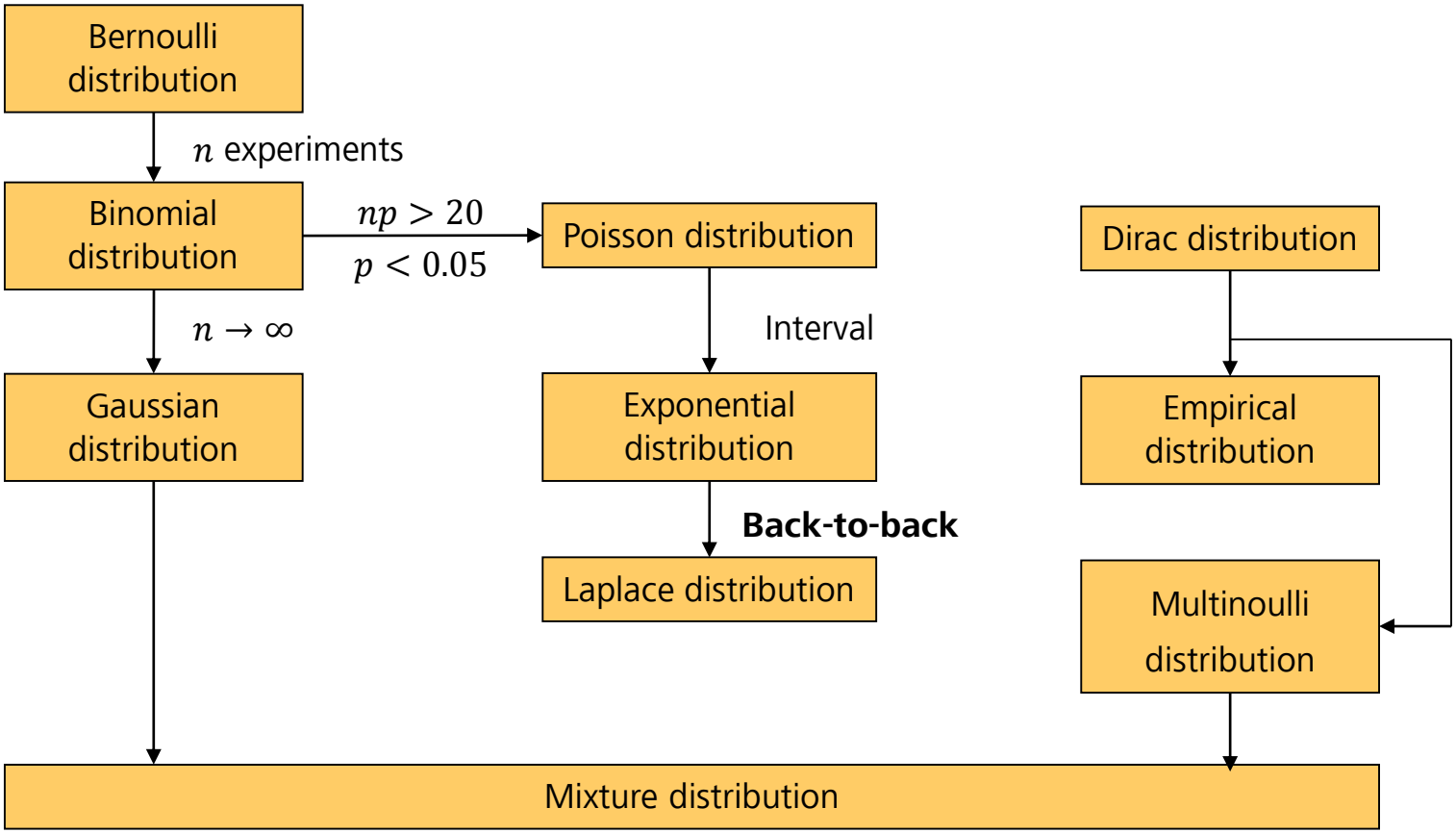
- If the probability density of continuous random variable X is

$$Laplace(x;\mu,b) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}},$$

where μ is the position parameter, and b is the scale parameter, X obeys the Laplace distribution. This distribution is expressed as $X \sim Laplace(x;\mu,b)$. $E(X) = \mu, Var(X) = 2b^2$.



Summary of Probability Distribution



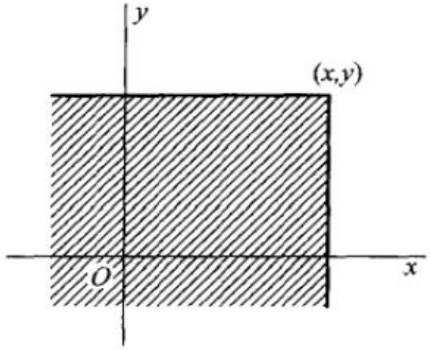
Two-Dimensional Random Variable and Joint Distribution Function

- **Two-dimensional random variable**: E is a random experiment, and its sample space is $S = \{e\}$. If $X = X(e)$ and $Y = Y(e)$ are defined as random variables on S, they make a vector (X,Y), called two-dimensional random variable.
- **Distribution function of two-dimensional random variable**: If (X,Y) is a two-dimensional random variable, for any real numbers x,y, the binary function applies:

$$F(x,y) = P\{(X \le x) \cap (Y \le y)\} = P\{X \le x, Y \le y\}$$

It is called a distribution function for a two-dimensional random variable (X,Y), or a joint distribution function for random variables X and Y.

- **Significance of the joint distribution function**: If (X,Y) is considered as the coordinate of a random point on the plane, distribution function $F(x,y)$ at (x,y) is the probability of random point (X,Y) falling in the infinite rectangular field at the point (x,y) vertex and at the lower left of the point.



Two-Dimensional Discrete Random Variable and Joint Distribution Law

- Two-dimensional discrete random variable: All possible values of discrete random variable (X,Y) can be finite or infinite pairs.
- Joint distribution law of X and Y:

Y \ X	x_1	x_2	…	x_i	…
y_1	p_{11}	p_{12}	…	p_{1i}	…
y_2	p_{21}	p_{22}	…	p_{2i}	…
⋮	⋮	⋮		⋮	
y_j	p_{j1}	p_{j2}	…	p_{ji}	…
⋮	⋮	⋮		⋮	

Two-Dimensional Continuous Variable and Joint Probability Density

- If distribution function $F(x, y)$ of two-dimensional random variable (X,Y) has a non-negative function $f\ (x,y)$ that makes the following apply to arbitrary x, y

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv,$$

(X, Y) is a continuous two-dimensional random variable and function is the joint probability density for two-dimensional random variable X.

Marginal Distribution

- **Marginal distribution function:** Two-dimensional random variable (X,Y) as a whole has distribution function $F(x,y)$. X and Y are random variables, and they also have their distribution functions, which are expressed as $F\_X(x)$ and $F\_Y(x)$ and are called as the marginal distribution functions of two-dimensional random variable (X,Y) about X and Y, respectively. $F_X(x) = P\{X \le x\} = P\{X \le x, Y \le \infty\} = F(x,\infty)$
  - For discrete random variable:
    - Marginal distribution function: $F_X(x) = \sum_{x_i \le x} \sum_{j=1}^{\infty} p_{ij}$.
    - Marginal density function: $p_{i\cdot} = \sum_{j=1}^{\infty} p_{ij}, j = 1,2,\cdots$.
  - For continuous random variable:
    - Marginal distribution function: $F_X(x) = F(x,\infty) = \int_{-\infty}^{x}[\int_{-\infty}^{+\infty} f(x,y)dy]dx$.
    - Marginal density function: $f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$.

Conditional Probability and Bayes Formula

- In many cases, we are interested in the probability that an event occurs when a given event is ongoing. This probability is called **conditional probability**.

$$P(Y|X) = \frac{P(YX)}{P(X)}$$

- We often need to compute $P(X|Y)$ when $P(Y|X)$ is specified, and if we know $P(X)$, we can use the **Bayes formula** to compute:

$$P(X|Y) = \frac{P(XY)}{P(Y)} = \frac{P(Y|X)P(X)}{P(Y)}$$

- Assuming that X is a probabilistic space $\{X\_1, X\_2, \ldots, X\_n\}$ composed of independent events, $P(Y)$ can be expanded with a **full probability formula**: $P(Y) = P(Y \mid X\_1)\, P(X\_1) + P(Y \mid X\_2)\, P(X\_2) + \ldots + P(Y \mid X\_n)\, P(X\_n)$. Then, the **Bayes formula** can be expressed as:

$$P(X_i|Y) = \frac{P(Y|X_i)P(X_i)}{\sum_{i=1}^{n} P(Y|X_i)P(X_i)}$$

- The chain rule of conditional probability:

$$P(X_1, X_2, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i|X_1, \ldots, X_{i-1})$$

Independence and Conditional Independence

- Two random variables X and Y, if for all x, y, the following applies

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

Random variables X and Y are of **mutual independence**, which is expressed as $X \perp Y$.

- If for each value of Z for the conditional probability about X and Y, the following applies

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z),$$

Random variables X are of **conditional independence** at given random variable Z, which is expressed as $X \perp Y|Z$.

Examples of Bayesian rules

- Wang went to hospital for a blood test, and got a positive result, indicating that he may have been attacked by the X disease. According to data on the Internet, 1% of the people who were sick of this disease were false positive, and 99% were true positive. In those who did not get sick of this disease, 1% of the people were false negative, and 99% were true negative. As a result, Wang thought, with only 1% false positive rate, and 99% true positive rate, the probability of Wang getting infected with the X disease should be 99%. However, the doctor told him that the probability of his infection was only about 0.09.

  $X = 1$ (infected), $X = 0$ (not infected), $y = 1$ (tested as positive), $y = 0$ (tested as negative)

$$P(X = 1|y = 1) = \frac{P(X = 1)P(y = 1|X = 1)}{P(y = 1|X = 1)P(X = 1) + P(y = 1|X = 0)P(X = 0)}$$

$$= \frac{P(X = 1) \times 0.99}{0.99 \times P(X = 1) + 0.01 \times (1 - P(X = 1))}$$

If $P(X = 1) = 0.001,\ P(X = 1, y = 1) = 0.09.$

Contents

1. Linear Algebra
2. **Probability Theory and Information Theory**
   - Basic Concepts of Probability Theory
   - Random Variables and Their Distribution Functions
   - **Numerical Characteristics of Random Variables**
   - Information Theory
3. Numeric Calculation

Expectation and Variance

- **Mathematical expectation (or mean, also referred to as expectation):** If the probability of each possible result in the experiments is multiplied by the sum of its results, you get one of the most basic mathematical characteristics. It reflects the mean value of random variables.
  - For discrete random variable: $E(X) = \sum_{k=1}^{\infty} x_k \, p_k, k = 1,2, \cdots.$
  - For continuous random variable: $E(X) = \int_{-\infty}^{\infty} xf(x)dx.$
- **Variance:** A measure of the degree of dispersion in which the probability theory and statistical variance measure random variables or a set of data. According to the probability theory, variance measures the deviation between the random variable and its mathematical expectation.

$$D(X) = Var(X) = E\{[X - E(X)]^2\}$$

In addition, $\sqrt{D(X)}$,$expressed\ as\ \sigma(X)$,is called standard variance or mean variance.$X^* = \frac{X - E(X)}{\sigma(X)}$, is called standard variable for X.

Covariance, Correlation Coefficients, and Covariance Matrices

- **Covariance:** In a sense, it indicates the strength of linear correlation of two variables and the scale of these variables.

$$Cov(X,Y) = E\big(X - E(X)\big)E\big(Y - E(Y)\big).$$

- The **correlation coefficient** is also called the linear correlation coefficient, which measures the linear relationship between two variables.

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

- **Covariance matrices** for random variable (X_1, X_2):

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

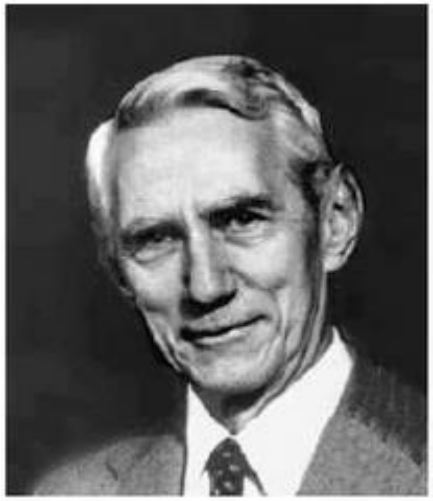
where $c_{ij} = Cov\big(X_i, X_j\big) = E\big\{[X_i - E(X_i)]\big[X_j - E\big(X_j\big)\big]\big\}, i, j = 1,\ 2,\ \cdots, n.$

Contents

1. Linear Algebra
2. **Probability Theory and Information Theory**
   - Basic Concepts of Probability Theory
   - Random Variables and Their Distribution Functions
   - Numerical Characteristics of Random Variables
   - **Information Theory**
3. Numeric Calculation

Information Theory

As a branch of applied mathematics, **information theory** mainly studies how to measure information contained in a signal. The sign of information theory was the publication of Shannon's paper, "A Mathematical Theory of Communication" in 1948. In this paper, Shannon creatively used probability theory to study communication problems, gave a scientific and quantitative description of information, and for the first time proposed the concept of **information entropy**.



Information Quantity

- The basic idea of information theory is that, when an unlikely event happens, it provides more information than a very likely event. If a message says "The sun rose this morning", there is so little information that it is unnecessary to send it; if a message says, "There's an eclipse this morning", the message is informative. The following conditions should be met to define **self-information** $I(x)$ for event $X = x$:
  - $f(p)$ should be a strictly monotonic decreasing function of probability, that is, $p_1 > p_2$, $f(p_1) < f(p_2)$.
  - When $p = 1, f(p) = 0$.
  - When $p = 0$, $f(p) = \infty$.
  - The joint information content of two independent events should be equal to the sum of their respective information quantity.

  Therefore, if the probability of a message is p, the **information quantity** contained in this message is:

$$I(x) = -\log_2 p$$

  Example: If you throw a coin, the information quantity about the coin showing the front or opposite is $I(\text{front}) = I(opposite) = 1bit$.

Information Entropy

- The information contained in the source is the average uncertainty of all possible messages transmitted by the source. Shannon, the founder of Information theory, refers to the amount of content that the source contains as **information entropy**, which is the **statistical average** of the amount of content in data partition D. The information entropy for the classification of m tuples in D is calculated as follows:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i).$$

where p_i is a none-zero probability that any tuple in D belongs to class C_i, $p_i = \frac{|C_{i,D}|}{|D|}$.

- For example, what is the entropy of throwing a coin?

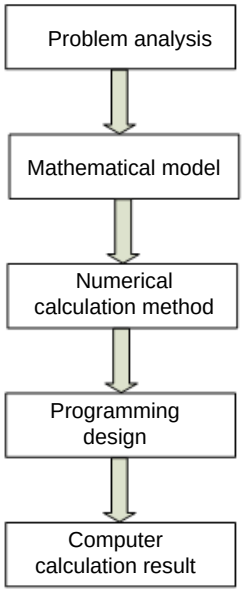
$$Info(D) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1bit.$$

Contents

1. Linear Algebra
2. Probability Theory and Information Theory
3. **Numeric Calculation**
   - **Basic Concepts**
   - Classification of and Solutions to the Optimization Problem

Numerical Calculation

- **Numerical calculation:** Refers to the method and process of effectively using a digital computer to solve approximate solutions of mathematical problems, and the disciplines formed by related theories. The process of solving practical problems with computers is as follows:



Overflow and Underflow

- **Underflow:** An underflow occurs when a number approximate to 0 is rounded to zero. Many functions show a qualitative difference when their arguments are zero rather than a small positive number.
- **Overflow:** Overflow occurs when a large number is approximated to ∞ or −∞. Further operations usually cause these infinite values to become non-numeric.
- **The large number " swallows" the small number:** When $a \gg b$, $a + b = a$, a numerical abnormality occurs.
- The ***Softmax*** function can **numerically stabilize** overflow and underflow:

$$softmax(x)_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}.$$

Number of Ill-Conditions

- **Ill-condition number:** Refers to the speed for a function to change with small changes of input.
- Considering function $f\ (x) = A\text{^}\ (-1)\ x$, when $A \in \mathbb{R}^{n\times n}$ has feature decomposition, the number of conditions is:

$$\max_{i,j} \left| \frac{\lambda_i}{\lambda_j} \right|,$$

This is the modulus ratio of the maximum and minimum eigenvalues. When this ratio is large, matrix inversion is particularly sensitive to input errors.

This sensitivity is the intrinsic characteristics of the matrix itself, not the result of the rounding error in the matrix inversion period. Even if we multiply the exact inverse of the matrix, the matrix of ill-conditions will magnify the pre-existing error. In practice, the error will be further compounded with the numerical error of the inversion process itself.

Contents

1. Linear Algebra
2. Probability Theory and Information Theory
3. **Numeric Calculation**
   - Basic Concepts
   - **Classification of and Solutions to the Optimization Problem**

Optimization Problem

- **Optimization problem:** Refers to the task of changing x to minimize or maximize function $f(x)$. It can be expressed as

  $min(.max)\, f(x)$

  $s.t.\quad g_i(x) \geq 0, i = 1,2,\ldots,m$, inequality constraints

  $h_j(x) = 0, j = 1,2,\ldots,p$, equality constraints

  where $x = (x_1, x_2, \ldots, x_n)^T \in R^n$. We refer to $f(x)$ as the objective function or guideline, or as a **cost function**, **loss function**, or **error function** when minimizing it.

Classification of Optimization Problems (1)

- **Constraint optimization**: a branch of optimization problems. Sometimes, the maximized or minimized $f(x)$ function under all possible values is not what we desire. Instead, we might want to find the maximum or minimum value of $f(x)$ when x is in a certain collection s. The points within the collection s are called **feasible points**.
- With **no constraints**, it can be expressed as:

$$minf(x)$$

The common method is Fermat theorem. If $f'(x) = 0$, the critical point is obtained. Then, verify that the extreme value can be obtained at the critical point.

- With **equality constraints**, it can be expressed as:

$$minf(x)$$
$$s.t. \quad h_i(x) = 0, i = 1,2,\ldots,n.$$

The common method is Lagrange multiplier method, that is, introducing n Lagrange multipliers λ to construct Lagrange function $L(x,\lambda) = f(x) + \sum_{i=1}^{n} \lambda_i h_i(x)$ and then seeking the partial derivative of each variable to be zero. Then, we can get the collection of candidate values, and get the optimal value through verification.

Classification of Optimization Problems (2)

- With **inequality constraints**, it can be expressed as:

$$minf(x)$$
$$s.t.\quad h_i(x) = 0, i = 1,2,\ldots,n,$$
$$g_j(x) \le 0, j = 1,2,\ldots,m.$$

A common method is to introduce new variables λ_i and α_j, to **Generalized Lagrangian functions** based on all equality, inequality constraints and $f(x)$.

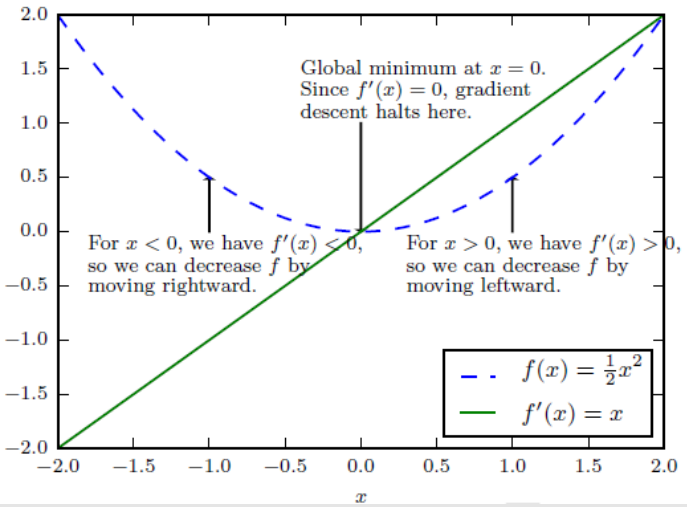
$$L(x,\lambda,\alpha) = f(x) + \sum_i \lambda_i h_i(x) \sum \alpha_j g_j(x),$$

We can use a set of simple properties to describe the most advantageous properties of constrained optimization problems, which are called **KKT (kuhn-kuhn-tucker) conditions**.

- The gradient of the generalized Lagrangian is 0.
- All constraints on x and KKKT multiplier are met.
- Inequality constraints show "complementary slackness type": $\alpha \odot h(x) = 0$.

Gradient Based Optimization Method (1)

- **Gradient descent**: The derivative indicates how to change x to slightly improve y. For example, we know that $f(x - \Delta x sign(f'(x)))$ is smaller than $f(x)$ for Δx that is small enough. So we can move x in the opposite direction of the derivative by a small step to reduce $f(x)$. This technique is called gradient descent.
- The extremum problem of a one-dimensional function:
  - The local extremum point of the function means that $f(x)$ cannot be reduced or increased by moving x.
  - The point where $f^{\wedge\prime}(x)$ =0 is called a critical point or a stationary point.
  - The extremum point of a function must be a stationary point, but a stationary point may not be the extremum point.

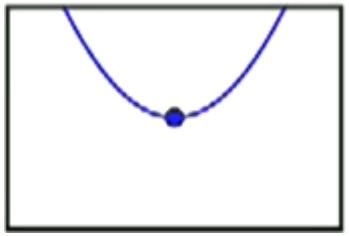


Gradient Based Optimization Method (2)

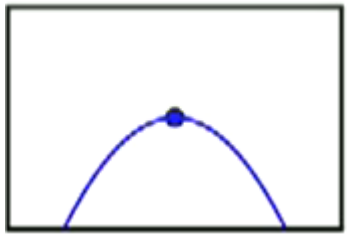
- Convex function: For$\lambda \in (0,1)$, given arbitrary $x_1, x_2 \in R$, the following applies:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

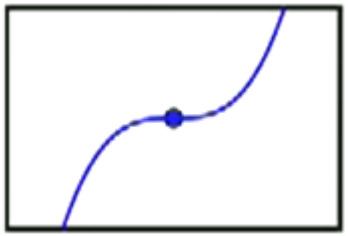
Then, $f(x)$ is called a convex function. The extremum point of the convex function is present at the stationary point.



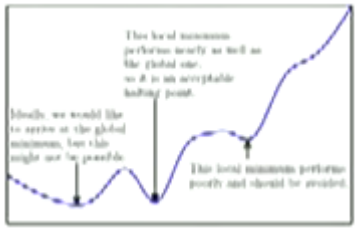
(a)



(b)



(c)



(d)

Gradient Based Optimization Method (3)

- To the case of multidimensional functions, the partial derivative is used to describe the degree of variation of the function relative to the respective variable.
- **Gradient**: It is a derivative relative to vector X, and is expressed as $\nabla\_xf\ (x)$. The derivative of $f\ (x)$ in the direction of u (unit vector) is $u^T\nabla_x f(x)$.
- For a task to minimize $f\ (x)$, we want to find the direction with the fastest downward change, where θ is the angle between u and gradient $\nabla\_xf\ (x)$.

$$\min_{u,u^Tu=1} u^T\nabla_x f(x)$$
$$= \min_{u,u^Tu=1} \|u\|_2 \|\nabla_x f(x)\|_2 \cos\theta$$

You can see that the direction in which $f\ (x)$ value decreases the maximum is the negative direction of the gradient.

Gradient Based Optimization Method (4)

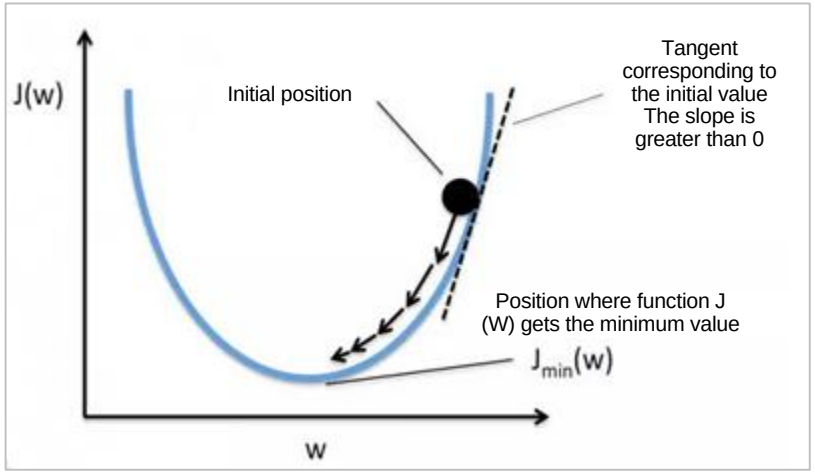
- A positive gradient vector points uphill, and a negative gradient vector points downhill. A move in the negative gradient direction can reduce $f(x)$, which is called **method of steepest descent** or **gradient descent**.
- Under the gradient descent method, the update point is proposed as:

$$x' = x - \varepsilon \nabla_x f(x)$$

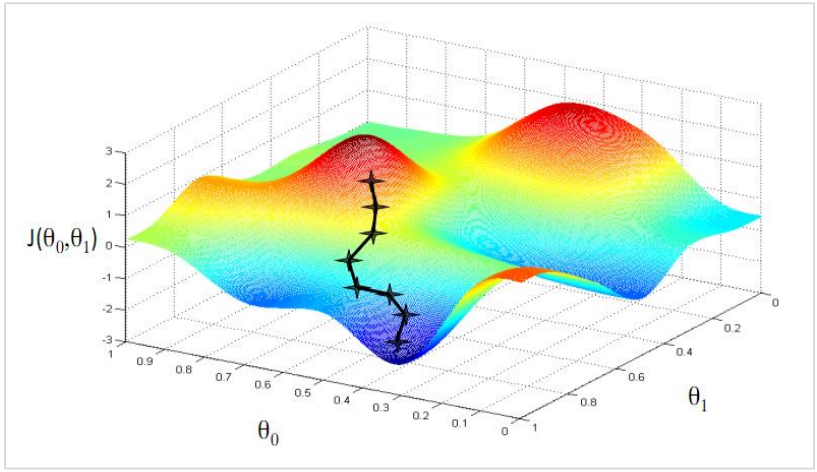
where ε is the learning rate, which is a positive scalar with a fixed step length.

- Iteration converges when the gradient is zero or approaching zero.

Gradient Based Optimization Method (5)



Two-dimensional space



Multi-dimensional space

Quiz

1. What are the relations and differences between a distribution function, distribution law and density function of a random variable?

Quiz

1. (Single-Choice) Matrix A has 3 rows and 2 columns. Matrix B has 2 rows and 3 columns. Matrix C has 3 rows and 3 columns. Which of the following operations makes sense? ( )

   A. AC

   B. BC

   C. A + B

   D. AB - BC

2. (True or False) Principal component analysis (PCA) is a statistical method. By means of orthogonal transformation, a group of variables that may have correlations are converted to a group of linearly related variables, and the converted group of variables is called principal component. ( )

   A. True

   B. False

Quiz

3. (Single-Choice) X and Y are random variables, and C is a constant. Which of the following descriptions of the properties of mathematical expectations is incorrect? ( )

   A. E (C) = C

   B. E (X + Y) = E (X) + E(Y)

   C. E (CX) = CE (X)

   D. E (XY) = E (X) E (Y)

4. (True or False) The correlation coefficient, also called the linear correlation coefficient, is used to measure the linear relationship between two variables, and the value is a real number greater than 0. ( )

   A. True

   B. False

Summary

- This chapter mainly describes the basics of deep learning, covering linear algebra, probability and information theory, and numerical calculation, and builds a foundation for further learning.

Summary

- Huawei Learning website:
  - http://support.huawei.com/learning/Index!toTrainIndex
- Huawei support knowledge base:
  - http://support.huawei.com/enterprise/servicecenter?lang=zh

Thanks

www.huawei.com