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**: A vector is an array of numbers. The numbers are arranged in order. We can identify each individual number by its index in that ordering.

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

- Matrix: a numerical table of m rows and n columns, which consists of m x n numbers and can be marked as $m \times n$. $a_{ij}(i = 1, 2, \cdots, m;\ j = 1, 2, \cdots, n)$:

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

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

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

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.

Determinant

- The **determinant** of a square matrix, denoted det(A), is a function that maps matrices to real scalars.

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

- The significance of the determinant is as follows:
  - 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 $C = A + B = (a_{ij} + b_{ij})_{s\times n}$.
  Note: The two matrices can be added only when the matrices have the same row quantity and column quantity.
- **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 $C_{i,j} = \sum_k A_{i,k}B_{k,j}$

Note: In order for AB to be defined, A must have the same number of columns as B has rows.

- The addition and multiplication of the matrix is called the linear operation of the matrix and meets the arithmetic law.
- Matrix multiplication meets the combination law and distribution law, but does not meet the switching law. To be specific:
  - $A(B + C) = AB + AC$
  - $A(BC) = (AB)C$
  - $AB \neq BA$

Matrix Transposition

- **Transposed matrix**: The transpose of a matrix is an operator which flips a matrix over its diagonal, that is, it switches the row and column indices of the matrix by producing another matrix denoted as A^T (also written as A′).
- Example

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

- 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

- The **trace operator** gives the sum of all the diagonal entries of a matrix:

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

- 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

Identity and Inverse Matrices

- **Identity matrix**: All the entries along the main diagonal are 1, while all the other entries are 0. An identity matrix does not change any vector when we multiply that vector by that 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 $A^{-1}A = I_n$.

Diagonal Matrix

- **Diagonal matrix**: consists mostly of zeros and have non-zero entries only along the main diagonal. It is often written as $diag(\lambda_1, \lambda_2, \cdots, \lambda_n)$

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

- 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:

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

- If a square matrix A is similar to a diagonal matrix (that is, if there is a reversible matrix P that makes $P^{-1}AP$ a diagonal matrix), A is diagonalized.

Symmetric Matrix

- **Symmetric matrix**: If $A^T = A$ ($a_{ij} = a_{ji}$) in square matrix $A = (a_{ij})_{n\times n}$, A is a symmetric matrix.

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

- **Orthogonal matrix**: If $AA^T = A^T A = I_n$ in the square matrix $A = (a_{ij})_{n\times n}$, A is an orthogonal matrix. That is, $A^{-1} = A^T$.

Contents

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:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad A\alpha = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2\alpha.$$

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

- If 2 is an eigenvalue of A and α is an eigenvector of A and belongs to eigenvalue 2, $k\alpha$ is also 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:

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

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 $\begin{vmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1 = (4-\lambda)(2-\lambda)$. Therefore, the eigenvalues of A are $\lambda_1 = 2$ and $\lambda_2 = 4$.

Taking $\lambda_1 = 2$, the corresponding eigenvector satisfies $\begin{bmatrix} 3-2 & -1 \\ -1 & 3-2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, we find that $x_1 = x_2$. Therefore, the corresponding eigenvector is $p_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. When $\lambda_1 = 2$, the eigenvector is $kp_1 (k \neq 0)$.

Taking $\lambda_2 = 4$, we find that $x_1 = -x_2$. The eigenvector is $p_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. When $\lambda_2 = 4$, the eigenvector is $kp_2 (k \neq 0)$.

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.

- The SVD is more generally applicable. Every real matrix has a singular value decomposition, but the same is not true of the eigenvalue decomposition. For example, if a matrix is not square, the eigendecomposition is not defined.

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.

- When A has more columns than rows, then solving a linear equation using the pseudoinverse provides one of the many possible solutions. Specifically, it provides the solution x = A+y with minimal Euclidean norm $||x||_2$ among all possible solutions.
- When A has more rows than columns, it is possible for there to be no solution. In this case, using the pseudoinverse gives us the x for which Ax is as close as possible to y in terms of Euclidean norm $||\boldsymbol{A}x - y||_2$.

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}

- A research object is often a complex system with multiple elements. Too many variables increase the difficulty and complexity of problem analysis. Use the relationship between original variables to replace the original variables with limited new variables, and make these few variables retain the variable information as much as possible. In this way, the problem is simplified.
- Principal component analysis is a statistical method to divide the original variables into a few comprehensive indexes. It is, in fact, a dimension reduction processing technology.

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:

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

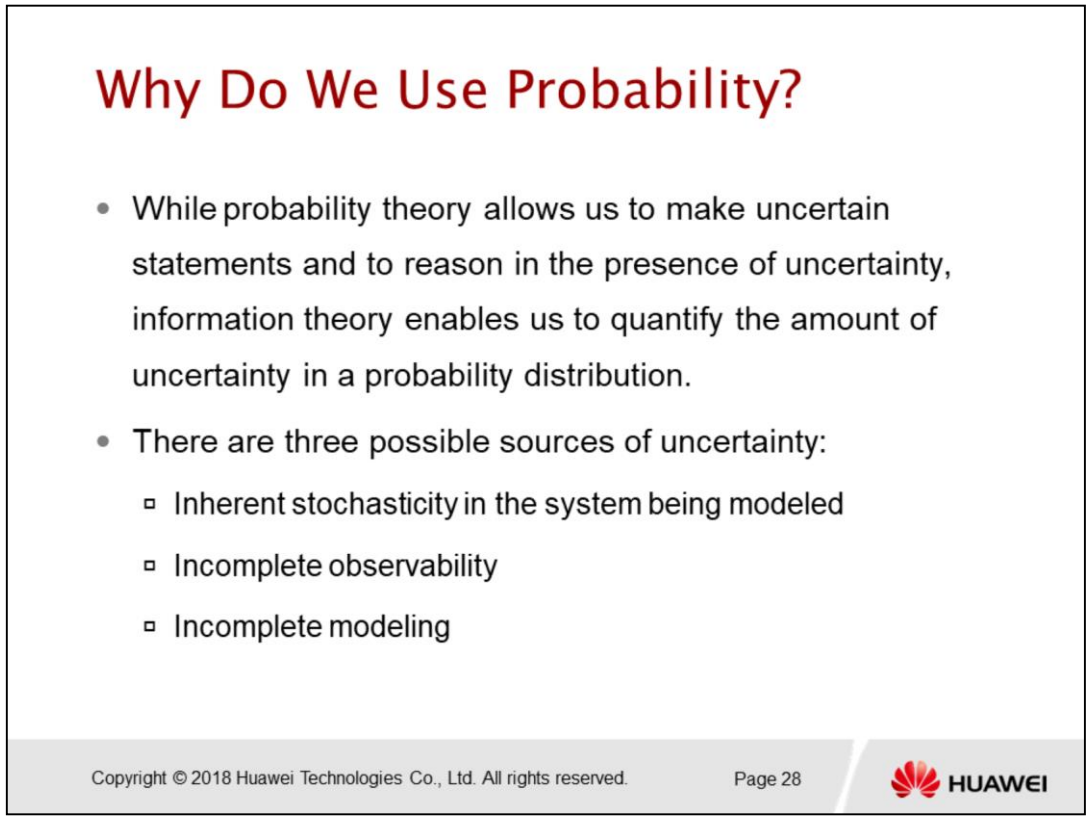
$\|x\|_2$

- 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



- Probability comes from gambling.
- There are three possible sources of uncertainty:
  - Inherent stochasticity in the system being modeled. For example, a hypothetical card game where we assume that the cards are truly shuffled into a random order.
  - Incomplete observability. Simply put, if you play cards with a casino master who remembers the order of cards, the next card is a probability event for you but a determined event for the casino master. The two of you judge the matter based on the different information that you have obtained respectively.
  - Incomplete modeling. When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions.

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.

- Random tests, also known as tests, are often represented by letter E.

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, $A_1 = \{x | x = 5\}$.

- The sample space S contains all sample points and is a subset of its own. It always occurs in each test, which is called an inevitable event. An empty set ∅ does not contain any sample point and is also a subset of the sample space. However, an empty set does not occur in each test and is known as an impossible event.

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 $P(A)$ (event probability) on each event A of E. The set function $P(*)$ must meet the following conditions:
  - Non-negative: For each event A, $0 \le P(A) \le 1$.
  - Standard: For the inevitable event S, $P(S) = 1$.
  - Countable additivity: $\{A_1, \dots\}$ are events incompatible with each other. That is, if $A_i A_j = \emptyset, i \ne j, i, j = 1, 2, \cdots$, we have $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$.

- A large number of experiments show that as the number of tests increases gradually, the frequency will stabilize to a constant. That is, when $n \to \infty$, $f_n(A)$ is close to $P(A)$ in a certain sense. Based on this fact, we can use $P(A)$ to indicate the possibility of the event occurred in a test.

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.

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

- Not all random tests can have results expressed in numbers. When the elements of the sample space are not numbers (for example, *S* = {Front side, Back side}, it is difficult to describe and study S. We can map each result (sample point *e*) of the random test with a real number *x* and then we have the concept of random variables. The purpose of introducing random variables is as follows: use the value range of random variables to indicate random events; leverage higher mathematics to study the random phenomenon.
- Example 1: Each result ($e = (i,j)\epsilon S$) of the experiment corresponds to a specified value ($i + j$). X is a single- and real-valued function defined in the sample space S. The definition field is the sample space. The value field is a set of real numbers: $\{2, 3, \ldots, 11, 12\}$.
- Example 2: When the elements of the sample space are not numbers, it is difficult to describe and study *S*. The random variable *Y* makes the sample point correspond to a real number.

Contents

HUAWEI

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	…	x_n	…
p_k	p_1	p_2	…	p_n	…

- The distribution law can be expressed by a formula or a table.
- The probability of notching a goal by a basketball player is 0.9. Then, what is the distribution law of the number X of notching in two times of independent shooting?

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

$E(X) = p$, $Var(X) = p(1-p)$.

- For example, the gender of a newborn baby and shoot targeting probability obey 0-1 distribution.

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 $np(1-p)$.

- When n of binomial distribution is very large, the probability is very difficult to calculate.
- The probability of notching a goal by a basketball player is 0.9. Then, what is the distribution law of two times of notching in four times of independent shooting?

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:

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

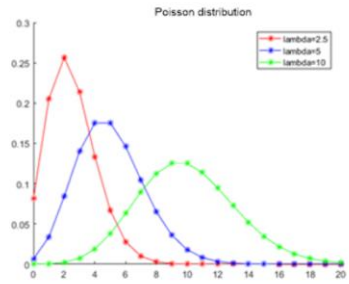
- **Poisson distribution:** If all possible values of random variables are 0, 1, 2, ..., the probability of taking each value is:

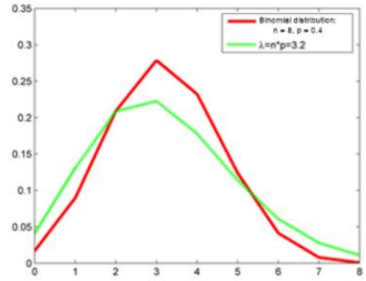
$$P\{X=k\} = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \quad \cdots$$

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.

- In 2012, 28 people were killed in a school shooting in Connecticut State, USA. Data shows that from 1982 to 2012, the United States had a total of 62 (mass) shootings. Of these, 7 occurred in 2012, the most frequent year. Was it a coincidence that there were so many shootings in 2012, or was it a worsening of American security? Do the shooting cases follow Poisson distribution?
  - A shooting case is a small probability event.
  - The shootings are independent and do not affect each other.
  - Is the probability of a shooting case stable?
- Poisson process properties:
  - Independent increment; steady increment; counting process.
  - Within $[t,\Delta t]$ time, the probability of one occurrence is $\lambda\Delta t+o(\Delta t)$, and that of two or more occurrences is $o(\Delta t)$.

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.

HUAWEI

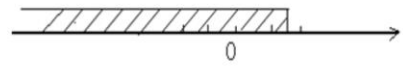
l Poisson distribution means the frequency of occurrence in a period, for instance, a day or an hour. For example, the number of patients in a day in a hospital, or the number of people waiting a bus at a bus station in an hour.

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 $(-\infty, x)$.



- Example: If you randomly throw a dice, what is the distribution function of the dice?

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, $P(X=a) = 0$.

HUAWEI

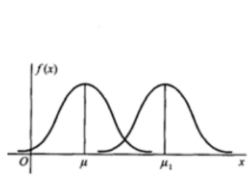
- The probability value of continuous random variable X taking any real number is 0: The interval endpoint does not need to be considered when the probability value of a continuous random variable falling in an interval is calculated. That is, $P(a < X < b) =$

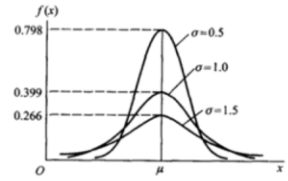
Special Distribution – Normal Distribution

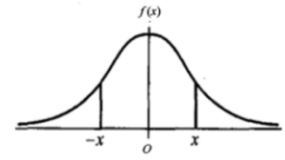
- If the probability density function of continuous random variable X is

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

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 μ = 0, σ = 1, random variable X obeys the standard normal distribution, which is expressed as X~N (0, 1).







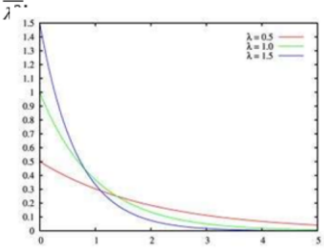
- A large number of random variables obey or approximately obey normal distribution in the nature and society.

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 $X\sim E(\lambda).E(X)=\frac{1}{\lambda},\mathrm{Var}(X)=\frac{1}{\lambda^2}$.



HUAWEI

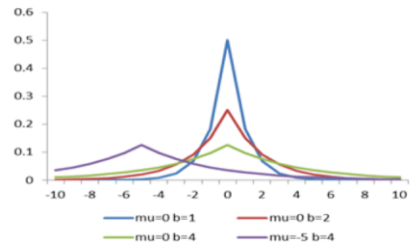
- No memory: $P\ (X>s+t \mid X>s) = P\ (X>t)$. If X is the life of a component, and the component has been used for s hours, the conditional probability of using it for at least $s+t$ hours is equal to the probability of using it for at least t hours from the start of use. This means that the component has been used for s hours without memory. Exponential distribution is widely used in reliability theory and queuing theory. Some people jokingly say that random variables following exponential distribution are "young forever". A 60-year-old person and a newborn baby have the same probability of living 10 more years. Do you believe it? If a person's life expectancy does follow an exponential distribution, the answer is yes.
- Poisson distribution refers to the number of times that an event occurs and exponential distribution refers to the time interval of independent random event occurrence.

Special Distribution – Laplace Distribution

- If the probability density of continuous random variable X is

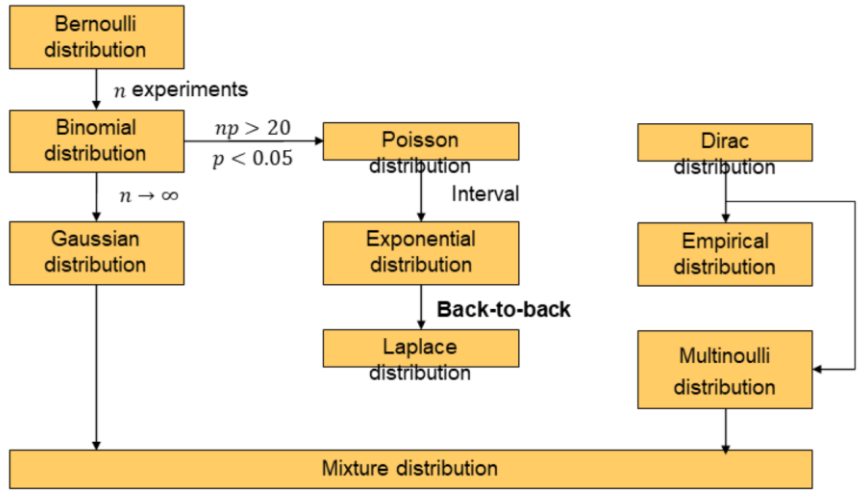
$$Laplace(x;\mu,b)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}},$$

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



- In the probability theory and statistics, Laplace distribution is a continuous probability distribution named with the name of Pierre-Simon Laplace. It can also be called double exponential distribution because it can be viewed as the back-to-back splicing of two exponential distributions at different positions. The difference between two independent random variables with the same probability exponential distribution is the random time Brownian motion according to exponential distribution, and therefore it follows Laplace distribution. It plays an important role in voice recognition and JPEG image compression.

Summary of Probability Distribution



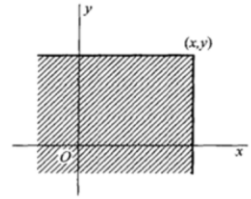
Two-Dimensional Random Variable and Joint Distribution Function

- **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.



HUAWEI

- The properties of two-dimensional random variables (X,Y) are not only related to X and Y, but also dependent on the interrelation of the two random variables. Therefore, it is not enough to study the properties of X or Y, and it is necessary to study (X,Y) as a whole.
- The properties of the joint distribution function:
  - $F(x,y)$ is a no-subtraction function for variables x and y. That is, for arbitrarily fixed y, $x_2 > x_1$, $(x_2,y) \ge F(x_1,y)$; for arbitrarily fixed x, $y_2 > y_1$, $F(x,y_2) \ge F(x,y_1)$
  - $0 \le F(x,y) \le 1$,and
  - $F(x+0,y) = F(x,y), F(x,y+0) = F(x,y)$, namely, $F(x,y)$ is about x right continuous, about y also right continuous.
  - For any $(x_1,y_1),(x_2,y_2), x_1 < x_2, y_1 < y_2$,the following inequations apply:
    - For any fixed y,$F(-\infty,y) = 0$,
    - For any fixed x,$F(x,-\infty) = 0$,
    - $F(-\infty,-\infty) = 0, F(\infty,\infty) = 1$
  - $F(x_2,y_2) - F(x_2,y_1) + F(x_1,y_1) - F(x_1,y_2) \ge 0$

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 basketball players play. The probability of player A notching a goal is 0.8, while that of player B is 0.9. Players A and B notch 5 goals in total, player A notches X goals and player B Y goals. What is joint distribution law of X and Y?

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.

- n-dimensional random variable: Set E as a random experiment, whose sample space is $S = \{e\}$,设$X_1 = X_1(e)$,$X_2 = X_2(e)$, ..., $X_n = X_2(e)$ is a random variable defined on the S. An n dimension vector consisting of them is called n-dimensional random variable.
- The joint distribution function of n-dimensional random variables: For arbitrary n real numbers $x_1, x_2, \dots, x_n$, n function $F(x_1, x_2, \dots, x_n) = P\{X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n\}$ is called the joint distribution function of n dimensional random variable $X_1, X_2, \dots, X_n$.

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,...,X_n} composed of independent events, $P(Y)$ can be expanded with a **full probability formula**: $P(Y) = P(Y|X\_1)\,P(X\_1) + P(Y|X\_2)\,P(X\_2) + \ldots + P(Y|X\_n)\,P(X\_n)$. 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 $X \perp Y|Z$.

- X: being late or not, Y: staying up late or not, Z: getting up late or not. X and Y under the condition of Z are mutually independent.
- X: future values, Y: past values, Z: current value. If this condition is met, future values are relevant only to the present and have nothing to do with the past.

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

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.

- Several important properties of mathematical expectation:
  - If C is a constant, E (C) = C.
  - If X is a random variable, and C is a constant, E (CX) = CE (x).
  - If X,Y are two random variables, E (X+Y) =E (x) +E (Y).
  - If X,Y are independent random variables, E (XY) =E (x) E (y).
- Several important properties of variance:
  - If C is a constant, D (C) =0.
  - If X is a random variable, and C is a constant, D (CX) =$C^{2}D$ (x).
  - If X,Y are two random variables, D (X+Y) =D (x) +D (Y).
  - The important condition of D (X) =0 is X with the probability of 1 constant E (X), that is P{X=E (X)}=1.

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

- The covariance has the following properties:
  - $Cov(aX, bY) = abCov(X,Y), a, b$ are constants.
  - $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$.
  - If two random variables are independent of each other, the covariance is 0. The converse condition does not stand.
  - If the covariance of two random variables is not 0, two random variables are associated.
- The properties of the correlation coefficient:
  - $|\rho_{XY}| \leq 1$.
  - The necessary and sufficient condition for $|\rho_{XY}| = 1$ is that there are constants a, b that make $P\ (Y{=}a{+}bX) = 1$.

Contents

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?

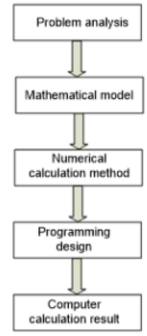
$$Info(D) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1bit.$$

Contents

1. Linear Algebra
2. Probability Theory and Information Theory
3. **Numerical 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:



HUAWEI

- The numerical calculation mainly studies how to use a computer to solve various mathematical problems, including the discretization of a continuous system and the solution to a discrete equation, and consider the error, convergence and stability.
- Machine learning algorithms usually require a large number of numerical calculations. It is necessary to solve the mathematical problem by updating the estimated value of the iterative process, instead of deriving the formula from the analytic process to provide the correct solution. Common operations include optimization and the solution to linear equations. For digital computers, real numbers cannot be accurately represented in finite memory, so it is difficult to compute only the functions involving real numbers.

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:

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

- In computer programming, sometimes for the same calculation problem, rounding errors in different algorithms cause different impact on calculation results. For the algorithm where a rounding error has little impact on calculation precision, it has a good numerical stability, but on the contrary condition the numerical stability of the algorithm is poor. The rounding error of the designed algorithm should be controllable under certain conditions. Otherwise, like the Butterfly Effect, the sunny Americas will be stormy after a few months. In the same way: 1.01^100≈2.7048138294,0.99^100≈0.3660323412732.
- In order to improve the stability of numerical values, we need to follow the following principles when designing algorithms:
  - Minimize operations.
  - For an additive operation, avoid adding a large number to a small number.
  - Avoid subtracting two approximate numbers.
  - Avoid small numbers as divisors or large numbers as multipliers.

Number of Ill-Conditions

- **Ill-condition number:** Refers to the speed for a function to change with small changes of input.
- Considering function $f(x) = A^{\wedge}(-1)\,x$, 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.

- If a function is quickly changed by a slight perturbation in input, this function is not good for scientific calculations, because rounding errors in the input can result in significant changes in output.

Contents

1. Linear Algebra
2. Probability Theory and Information Theory
3. **Numerical 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.

- It is to find the best solution under given conditions.

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:

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

  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:

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

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

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

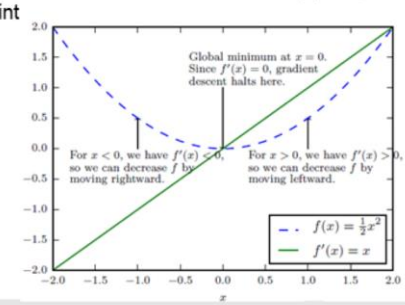
We can use a set of simple properties to describe the most advantageous properties of constrained optimization problems, which are called **KKT (kuhn-kuhn-tucker) conditions**.

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

- The KKT conditions transform the solution to the original function into the solution to the dual problem. The dual problem and the original optimization problem have the same optimal value and optimal solution, and the constrained optimization problem is transformed into an unconstrained optimization problem.

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



HUAWEI

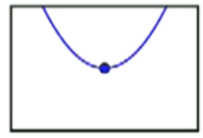
- Derivative: When independent variable x of function $y=f(x)$ produces an increment Δx on point x_0, the ratio of increment Δy of function output value to Δx of the increment of the independent variable has an extreme value a at Δx close to zero, a is the derivative at x_0, which is recorded as $f'(x_0)$ or $\frac{df(x_0)}{dx}$.
- Significance of derivative: Derivative $f'(x)$ represents the slope of $f(x)$ at point x. That is, it indicates how the input changes can be scaled to obtain the corresponding change in the output: $f(x + \Delta x) = f(x) + \Delta x f'(x)$.

Gradient Based Optimization Method (2)

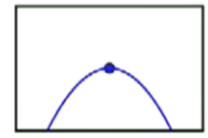
- Convex function: For$\lambda \in (0,1)$, given arbitrary $x_1, x_2 \in R$, the following applies:

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

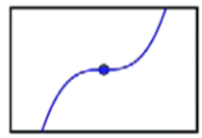
Then, $f(x)$ is called a convex function. The extremum point of the convex function is present at the stationary point.



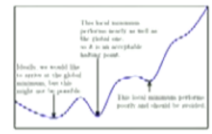
(a)



(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)

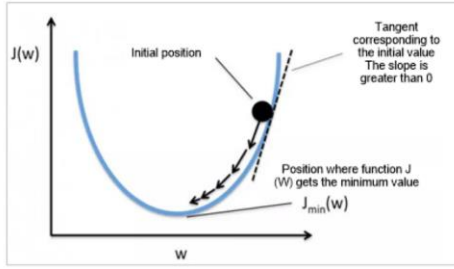
- A positive gradient vector points uphill, and a negative gradient vector points downhill. A move in the negative gradient direction can reduce $f(x)$, which is called **method of steepest descent** or **gradient descent**.
- Under the gradient descent method, the update point is proposed as:

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

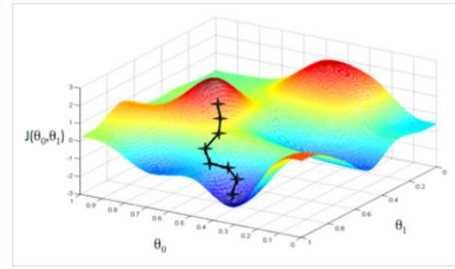
where ε is the learning rate, which is a positive scalar with a fixed step length.

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

Gradient Based Optimization Method (5)

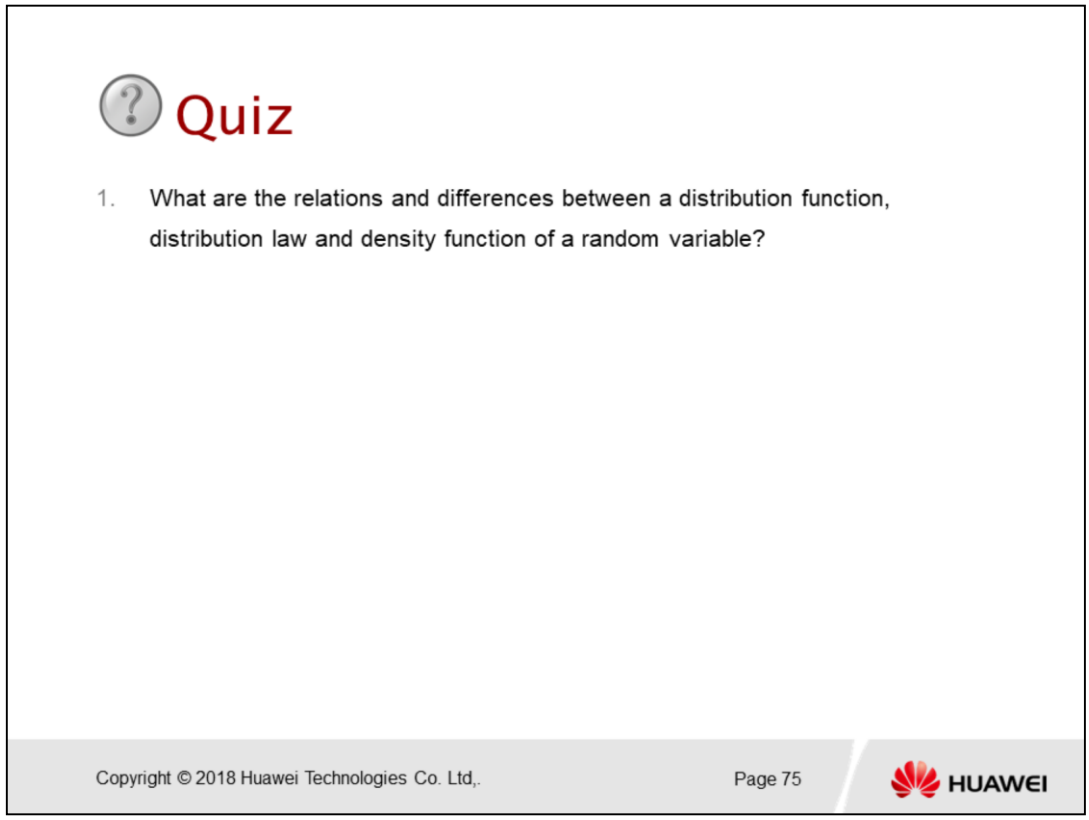


Two-dimensional space

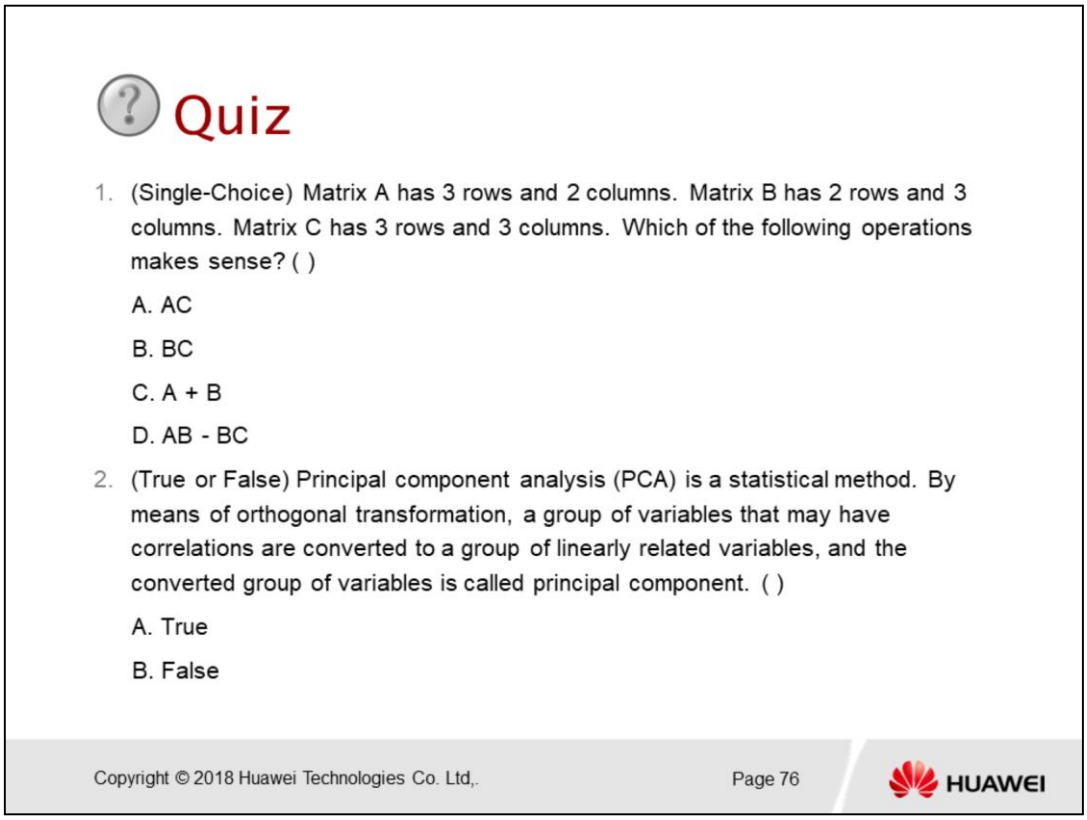


Multi-dimensional space

- We can choose ε in several different ways.
  - The popular way is to select a small constant. Sometimes we choose the step length that the directional derivative disappears by calculating.
  - The second way is to compute $f(x - \varepsilon\nabla_x f(x))$ based on several ε, and select ε that generates the minimum objective function value, which is called linear search.
- Although gradient descent is limited to optimization problems in a continuous space, the general concept of moving one small step into a better situation can be extended to a discrete space.



A distribution function describes the value law of a random variable, which can be discrete or continuous. A distribution law describes only the rule of a discrete random variable. A density function describes only the value rule of a continuous random variable.



- Answers: 1. B 2. B

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

- Answers: 3. D 4. B

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