Machine Learning

Mathematical Basis

Teaching Assistant：Shuwei Yan

Today's Topics

- Probability
- Calculus
- Vector and Matrix

Today’s Topics

- ***Probability***
- Calculus
- Vector and Matrix

Random Experiment

- Definition:
  - Repeatable
  - Known Range of Outcomes
  - Unknown Specific Outcome

- Example:

E1

E2

Sample Space

- Definition: The set of all possible outcomes of E, denoted by Ω.

- In E1, Ω = {Head Side, Tail Side}

- In E2, Ω = {1, 2, 3, 4, 5, 6}

Random Event

- Definition: An outcome that may or may not occur in E

- Example：

In E1:

A: Head Side

B: Tail Side

……

In E2:

A: 2

B: Even Number

……

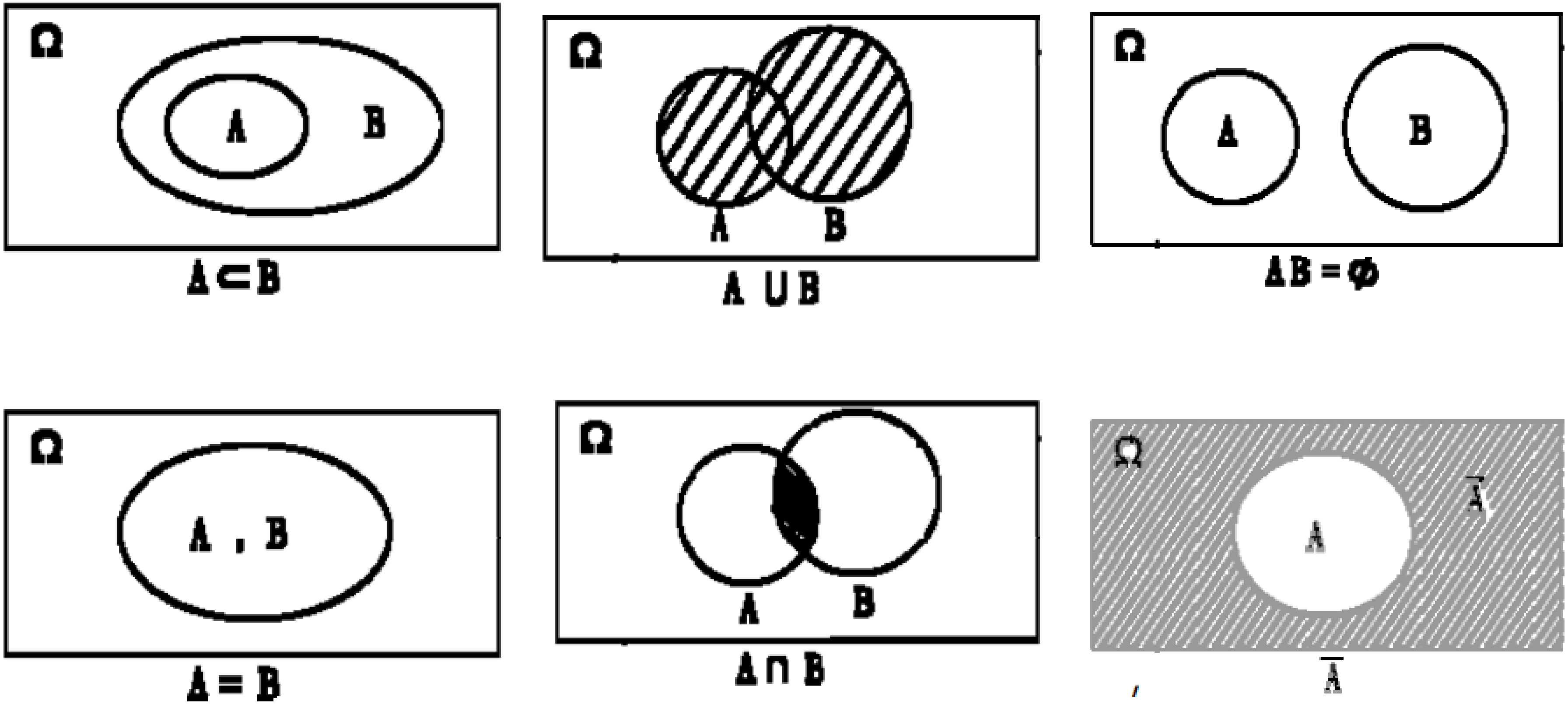
Certain Event and Impossible Event

- Certain Event: An event that is certain to occur in every trial
- Impossible Event: An event that cannot occur in any trial

- Certain Event $\longrightarrow$ P=1
- Impossible Event P=0

- P=1 $\not\longrightarrow$ Certain Event
- P=0 Impossible Event

Relationships Between Random Events



Venn Diagram

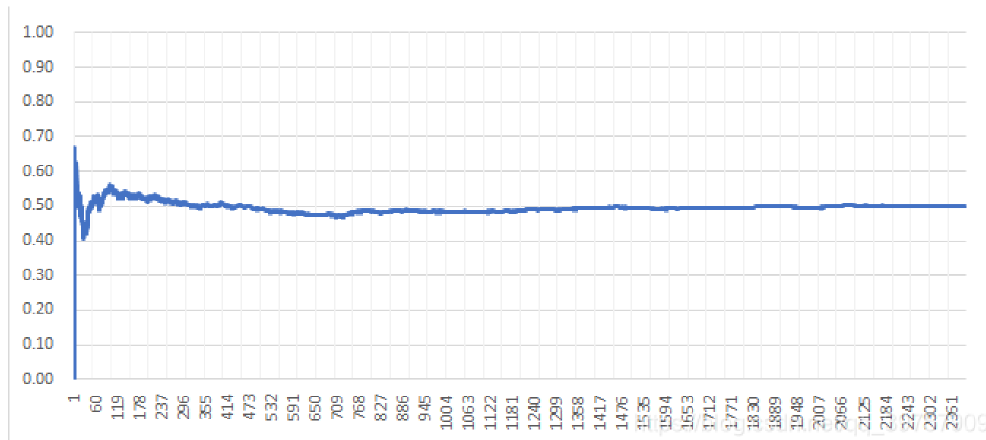
Relational Operations

Venn Diagram

	∪	∩	—
∪			
∩			
—			

Frequency and Probability

- Frequency: In n repeated trials, if event A occurs n_A times, then $\frac{n_A}{n}$ is called the frequency of event A in n trials, denoted as $f_n(A)$.



- Probability: As the number of repeated trials n increases, the frequency $f_n(A)$ of event A stabilizes around a constant p, then the constant p (where 0≤p≤1) is called the probability of event A, denoted as P(A)=p.

Property

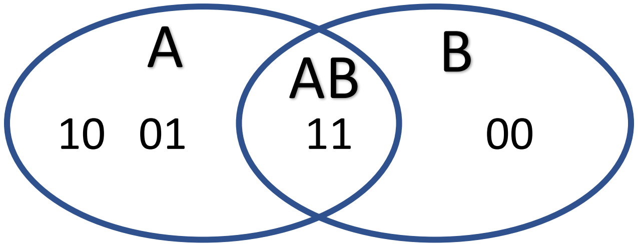
- $0 \le P(A) \le 1$
- $P(\Omega)=1$
- If $A \subseteq B$, $P(A) \leq P(B)$
- If $A \subseteq B$, $P(B - A) = P(B) - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(AB)$
- ……

Conditional Probability

- Definition: The probability of event A occurring given that event B has occurred, denoted as $P(A|B)$.

Conditional Probability

- Example:
- Flip a coin twice:
- Let event A represent "at least one head appears," and event B represent "both flips result in the same face."
- Calculate: P(B|A)



Let 1 represent heads and 0 represent tails.
The sample space $\Omega=\{11,10,01,00\}$.
Event A={11,10,01} and event B={11,00}.

- When event A occurs, the sample space is reduced to {11,10,01}, and only 11 represents event B.
- $P(B|A) = \frac{1}{3}$

Conditional Probability

- Definition: The probability of event A occurring given that event B has occurred, denoted as P(A|B).

$$P(B|A) = \frac{N(AB)}{N(A)} = \frac{\frac{N(AB)}{N(\Omega)}}{\frac{N(A)}{N(\Omega)}} = \frac{P(AB)}{P(A)}$$

Multiplication Rule

- $P(B|A) = \frac{P(AB)}{P(A)}$ ➡ $P(AB) = P(A)P(B|A) \quad P(A) > 0$

- $P(A|B) = \frac{P(AB)}{P(B)}$ ➡ $P(AB) = P(B)P(A|B) \quad P(B) > 0$

Expansion:

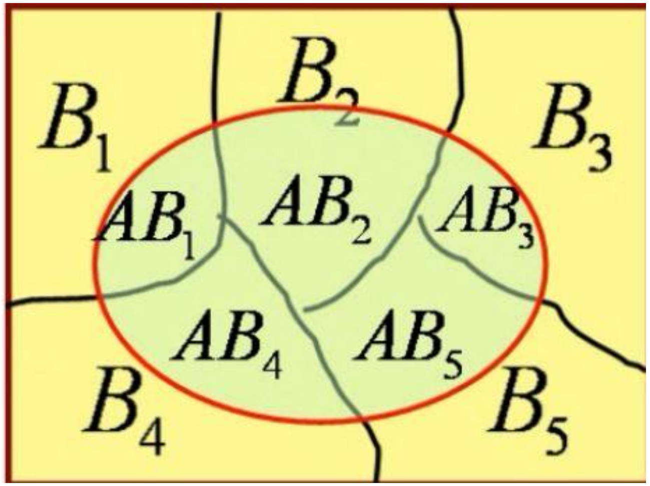
$$P(ABC) = P(AB)P(C|AB) = P(A)P(B|A)P(C|AB) \qquad P(AB) > 0$$

$$P(A_1A_2A_3 \dots A_n) = P(A_1)P(A_2|A_2)P(A_3|A_1A_2) \dots P(A_n|A_1A_2A_3 \dots A_{n-1})$$
$$P(A_1A_2A_3 \dots A_{n-1}) > 0$$

Total Probability Formula

- Let the sample space of the experiment E be Ω, and let (B1,B2,…,Bn) be a partition of Ω, with P(Bi)>0 for i=1,2,…,n. Let A be an event of E.

$$P(A) = P(AB_1) + P(AB_2) + \cdots + P(AB_n)$$



Multiplication Rule

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_n)P(A|B_n) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$$

Total Probability Formula

- Example： A store sells televisions produced by two factories. Televisions from Factory 1 account for 70%, and those from Factory 2 account for 30%. The qualification rate from Factory 1 is 95%, and that from Factory 2 is 80%. Find the qualification rate of the televisions sold by the store.
- A： a television is qualified
- B： televisions from Factory 1
- C： televisions from Factory 2.

$$P(B) = 0.7, P(A|B) = 0.95$$
$$P(C) = 0.3, P(A|C) = 0.8$$

$$P(A) = P(B)P(A|B) + P(C)P(A|C) = 0.7 \times 0.95 + 0.3 \times 0.8 = 0.905$$

Bayes' Theorem

$$P(AB) = P(A)P(B|A) \quad P(A) > 0$$

$$P(AB) = P(B)P(A|B) \quad P(B) > 0$$

$$\Downarrow$$

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

- let (B1,B2,…,Bn) be a partition of the sample space **Ω**

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

$$(i = 1,2, \dots, n)$$

Bayes' Theorem

- Example: When the machine is well-adjusted, the qualification rate of the products is 95%, and when the machine has a certain fault, the qualification rate is 50%. The probability that the machine is well-adjusted is 90%. Given that a product is qualified, find the probability that the machine is well-adjusted.
- Let A represent a qualified product, and B represent a well-adjusted machine. Then:

$$P(B) = 0.9$$
$$P(\bar{B}) = 1 - P(B) = 0.1$$
$$P(A|B) = 0.95$$
$$P(A|\bar{B}) = 0.5$$

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

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

$$= 0.945$$

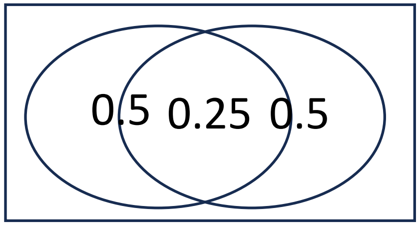
Prior and Posterior Probability

$$P(B) = 0.9 \qquad P(B|A) = 0.945$$

- The probability that the machine is well-adjusted, P(B)=0.9, is derived from past data analysis and is called the prior probability.
- The conditional probability is the probability that is revised after obtaining the information that the product is qualified, and is called the posterior probability.

Independence of Events

- For any two events A and B, if P(AB)=P(A)P(B), then events A and B are said to be independent.
- Events A and B are independent



$\Rightarrow$

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

A and $\bar{B}$， B and $\bar{A}$， $\bar{A}$and $\bar{B}$ are independent

- For events $A_1, A_2, \dots, A_n$:
- Pairwise Independence $$P(A_iA_j) = P(A_i)P(A_j) \quad 1 \le i < j \le n$$
- Mutual Independence $$P(A_{i_1}, A_{i_2}, \dots, A_{i_k}) = P(A_{i_1})P(A_{i_2})\dots P(A_{i_k})$$

Random Variable

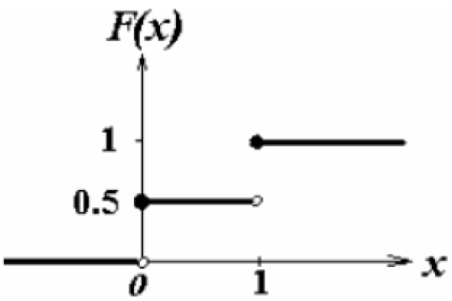
$X = X(e)$ Events $\longrightarrow$ Real Value

- Example：In E1：

$$X = \begin{cases} 0, & \text{Head Side} \\ 1, & \text{Tail Side} \end{cases}$$

- Distribution Function：

$$F(x) = P\{X \le x\} \Rightarrow F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$



Discrete Random Variable

- The random variable X can take a finite number of values or a countably infinite number of values.

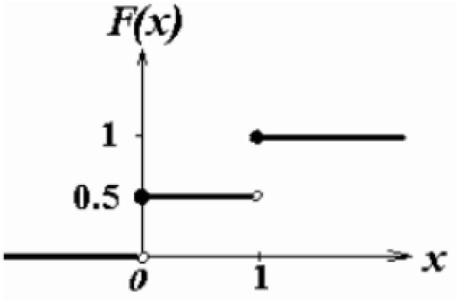
$$p_k = P\{X = x_k\}, \qquad k = 1,2,3,\ldots$$

- Probability distribution

X	0	1
p	$\frac{1}{2}$	$\frac{1}{2}$

- The distribution function is step-shaped

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$



Discrete Random Variable

- 0-1 distribution

$$P\{X=1\}=p, P\{X=0\}=1-p$$

- Binomial Distribution $X \sim B(n,p)$

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

- Poisson Distribution $X \sim Possion\ (\lambda)$

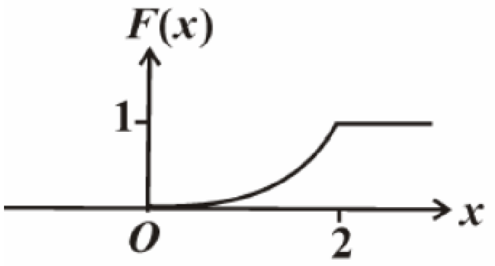
$$P\{X=k\}=e^{-\lambda}\frac{\lambda^k}{k!}, k=0,1,2\cdots, \lambda>0$$

$$\sum p = 1$$

Continuous Random Variable

- The distribution function is the integral of the Probability Density Function

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



- The probability that a X falls within interval I is the integral of its Probability Density Function over I

$$P\{X \in I\} = \int_{a}^{b} f(x)dx \qquad I = (a, b]$$

Continuous Random Variable

- Uniform Distribution

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{else}\end{cases}$$

- Exponential Distribution

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, x\ge 0\\ 0, \quad \text{else}\end{cases}$$

- Gaussian Distribution (Normal Distribution)

$$\mathbf{X}\sim N(\mu,\sigma^2)$$

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

$$\left.\right\} \int \mathrm{f}(X)dx = 1$$

Joint Probability Distribution

- Definition:
  - The probability distribution of two or more random variables.

- Example:
- Hitting Coordinates (x, y) in Target Shooting:

Joint Probability Distribution

- Distribution function

$$F(x, y) = P(X \leq x, Y \leq y)$$

- Two-Dimensional Discrete Joint Probability Distribution

$$F(x_0, y_0) = \sum_{x_i \leq x_0} \sum_{y_j \leq y_0} p_{ij}$$

- Two-Dimensional Continuous Joint Probability Distribution

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

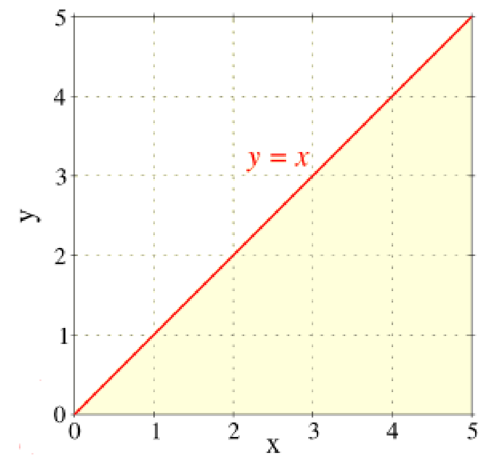
Marginal Distribution

- Definition: The distribution of a single variable within a joint probability distribution.

$$F_X(x) = P(X \leq x) = P(X \leq x, Y < \infty) = F(x, \infty)$$

$$f_X(x) = \int_{\Omega_Y} f(x, y)dy$$

Ω_Y represents the range of values
that Y can take when X = x, for example:



Mathematical Expectation

- Discrete Random Variable

$$P\{X = x_i\} = p_i, \quad i = 1,2,\ldots \quad \Rightarrow \quad E(X) = \sum_{i=1}^{+\infty} x_i p_i$$

- Continuous Random Variable

$$f(x) \quad \Rightarrow \quad E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

Mathematical Expectation

- Example 1

X	1	2	3	4
p	1/8	1/4	1/2	1/8

- $E(x)$
- $= 1 \times \frac{1}{8} + 2 \times \frac{1}{4} + 3 \times \frac{1}{2} + 4 \times \frac{1}{8}$
- $= \frac{1}{8} + \frac{1}{2} + \frac{3}{2} + \frac{1}{2}$
- $= \frac{21}{8}$

- Example 2

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{else} \end{cases}$$

- $E(x)$
- $= \int_{-\infty}^{+\infty} x f(x) dx$
- $= \int_{-\infty}^{a} 0 dx + \int_{a}^{b} x \frac{1}{b-a} dx + \int_{b}^{+\infty} 0 dx$
- $= 0 + \frac{x^2}{2(b-a)} |_a^b + 0$
- $= \frac{b^2-a^2}{2(b-a)}$
- $= \frac{a+b}{2}$

Mathematical Expectation

- Mathematical Expectation of the Function of a Random Variable $Y = g(X)$
- Discrete Random Variable

$$E(Y) = E[g(x)] = \sum_{i=1}^{+\infty} g(x_i) p_i$$

- Continuous Random Variable

$$E(Y) = E[g(x)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Mathematical Expectation

- Example 1

X	1	2	3	4
p	1/8	1/4	1/2	1/8

- $E(x+2)^2$
- $= 9 \times \frac{1}{8} + 16 \times \frac{1}{4} + 25 \times \frac{1}{2} + 36 \times \frac{1}{8}$
- $= \frac{9}{8} + 4 + \frac{25}{2} + \frac{9}{2}$
- $= \frac{177}{8}$

- Example 2

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{else} \end{cases}$$

- $E(x+2)^2$
- $= \int_{-\infty}^{+\infty} x f(x) dx$
- $= \int_{-\infty}^{+\infty} (x+2)^2 f(x) dx$
- $= \int_a^b (x+2)^2 \frac{1}{b-a} dx$
- $= \frac{(x+2)^3}{3(b-a)} \Big|_a^b$
- $= \frac{(b+2)^3 - (a+2)^3}{3((b+2)-(a+2))}$
- $= \frac{(b+2)^2 + (b+2)(a+2) + (a+2)^2}{3}$

Mathematical Expectation

- Property:

- $E(c) = c$

- $E(c_1X_1 + c_2X_2 + \cdots + c_nX_n) = c_1E(X_1) + c_2E(X_2) + \cdots + c_nE(X_n)$

- When Random Variables are Independent

$$E(X_1X_2 \dots X_n) = E(X_1)E(X_2)E(X_n)$$

Variance

- Discrete Random Variable

$$D(X) = E[X - E(X)]^2 = \sum_i [x_i - E(X)]^2 p_i$$

- Continuous Random Variable

$$D(X) = E[X - E(X)]^2 = \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x) dx$$

- Property:
  - $D(c) = 0$
  - $D(cX) = c^2 D(X)$
  - $D(X) = E(x^2) - [E(x)]^2$

Exercise

- A batch of products is produced by 4 factories, with production quantities of 3000, 2000, 2500, and 2500 units, respectively. The defect rates of these factories are 5%, 8%, 15%, and 10%, respectively.
- (1) Calculate the defect rate of the entire batch of products.
- (2) If a product is randomly selected and found to be defective, calculate the probability that it was produced by Factory 1

Exercise

- X follows a uniform distribution on the interval [0, π]. Find：

$$E(sinX), E(X^2), E(X - E(X))^2$$

- Find the variance of the uniform distribution on [a, b]

Today's Topics

- Probability
- ***Calculus***
- Vector and Matrix

Derivative

- Geometric Meaning: The Slope of the Tangent Line to the Function Curve at a Point

$y = C$	$y' = 0$
$y = a^x$	$y' = a^x \ln a$
$y = e^x$	$y' = e^x$
$y = x^n$	$y' = nx^{n-1}$
$y = \log_a x$	$y' = \frac{1}{x \ln a}$
$y = \ln x$	$y' = \frac{1}{x}$
$y = \sin x$	$y' = \cos x$
$y = \cos x$	$y' = -\sin x$

Derivative

- Four Basic Operations

$$(u \pm v)' = u' \pm v' \text{ ....} ①$$

$$(uv)' = u'v + uv' \text{ ....} ②$$

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2} \text{ ....} ③$$

- Chain Rule

h(x) = f(g(x))   h’(x) = f’(g(x)) * g’(x)

Partial Derivative

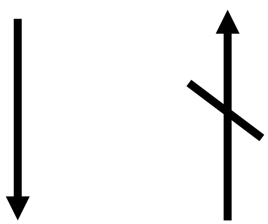
- Definition: Hold other variables constant and differentiate with respect to one variable

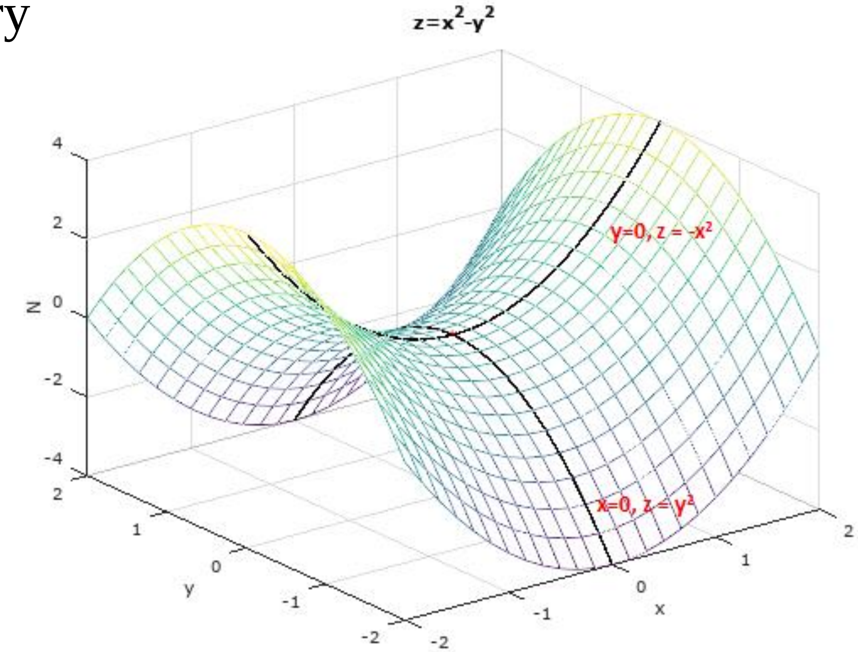
- Example”
  - 1. Find the partial derivatives of z=x2+3xy+y2 at the point (1, 2)
  - 8 and 7
  - 2. $z = x^3y^2 - 3xy^3 - xy + 1$，find $\frac{\partial^2 z}{\partial x \partial y}$ and $\frac{\partial^2 z}{\partial y \partial x}$
  - $6x^2y - 9y^2 - 1$
  - $6x^2y - 9y^2 - 1$

Maximum and Minimum Value

- Possible scenarios:
- 1. Extreme values on the boundary
- 2. Stationary points

- Find stationary points:





- Derivative or all partial derivatives are zero

Hessian matrix

- $H(x_0, y_0) = \begin{bmatrix} f_{xx}(x_0, y_0) & f_{xy}(x_0, y_0) \\ f_{yx}(x_0, y_0) & f_{yy}(x_0, y_0) \end{bmatrix} = \begin{bmatrix} A & B \\ B & C \end{bmatrix}$

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Lagrange multipliers

Steps for Using Lagrange Multipliers to Find Conditional Extrema:

1.Identify the objective function and constraint:

1. Let f(x) be the function to optimize.
2. Let g(x)=0 be the constraint.

2.Form the Lagrangian function:

L(x,λ)=f(x)+λ·g(x)

3.Compute partial derivatives:

Take the partial derivatives of L with respect to each variable xi and λ.

4.Set partial derivatives to zero:

$\frac{\partial L}{\partial \mathrm{x}_i} = 0$ for all x_i, $\frac{\partial L}{\partial \lambda} = 0$

5.Solve the system of equations to find critical points (x,λ).

6.Determine the nature of critical points (maxima, minima, or saddle points) using appropriate tests.

Lagrange multipliers

- **Example:**

  Find the extrema of $f(x,y) = x^2 + y^2$ subject to the constraint $x + y = 1$.

- **Solution:**

1. **Constraint:** $g(x,y) = x + y - 1 = 0$
2. **Lagrangian:** $L = x^2 + y^2 + \lambda(x + y - 1)$
3. **Partial derivatives:**

   $\partial L/\partial x = 2x + \lambda = 0$

   $\partial L/\partial y = 2y + \lambda = 0$

   $\partial L/\partial \lambda = x + y - 1 = 0$
4. **Critical point:** $(x, y) = (1/2, 1/2)$, $\lambda = -1$
5. **Evaluate f:** $f(1/2, 1/2) = (1/2)^2 + (1/2)^2 = 1/4 + 1/4 = 1/2$
6. **Conclusion:** The minimum value is 1/2 at (1/2, 1/2).
   (Since the function is convex, this is the global minimum.)

Exercise

- A company produces two products, with the profit function $P(x, y) = xy$, where x and y represent the production quantities of the two products, respectively. During the production process, raw materials are consumed. Producing each unit of product A requires 2 kilograms of raw materials, and producing each unit of product B requires 1 kilogram of raw materials. The company has a total of 100 kilograms of raw materials, which must be completely used. How should production be arranged to maximize profit? Please solve using the method of Lagrange multipliers.

Today's Topics

- Probability
- Calculus
- ***Vector and Matrix***

vector

Column vector

$$\begin{bmatrix} 1 \\ 0 \\ 5 \\ 6 \\ 2 \end{bmatrix}$$

Row vector

$$\begin{bmatrix} 1 & 0 & 5 & 6 & 2 \end{bmatrix}$$

Vector Norm

$$x = [x_0, x_1, \dots, x_m]^T$$

1-Norm: $\|x\|_1 = \sum_{i=1}^{m} |x_i|$

2-Norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{m} x_i^2}$

p-Norm: $\|x\|_p = (\sum_{i=1}^{m} |x_i|^p)^{\frac{1}{p}}$

∞-Norm: $\|x\|_\infty = \max_i |x_i|$

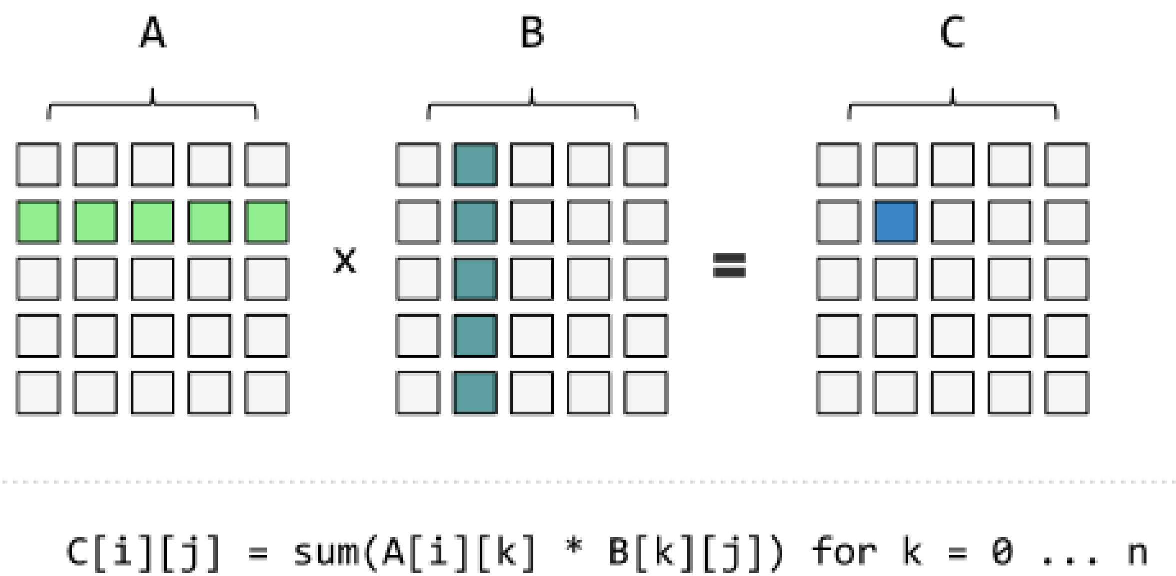
Matrix and Transpose

$$\underset{m\times n}{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \Rightarrow \underset{n\times m}{A^T} = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & & & \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \Rightarrow \mathbf{A}^T = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \Rightarrow \mathbf{B}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Matrix Multiplication



The number of columns in matrix A equals the number of rows in matrix B.

Matrix Multiplication

- Example

$$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 0 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 11 & 10 \\ 9 & 14 \end{bmatrix}$$

$$11 = 1 \times 1 + 3 \times 0 + 2 \times 5$$
$$10 = 1 \times 3 + 3 \times 1 + 2 \times 2$$
$$9 = 4 \times 1 + 0 \times 0 + 1 \times 5$$
$$14 = 4 \times 3 + 0 \times 1 + 1 \times 2$$

Eigenvalues and Eigenvectors (of a square matrix)

$$\lambda x = Ax, x \neq 0 \qquad\qquad (A - \lambda I)x = 0, x \neq 0$$

- Eigenvalues： $\lambda \; when \; |A - \lambda I| = 0$
- **Example**： Eigenvalues and Eigenvectors of $\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$
- $\begin{bmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{bmatrix} = (3-\lambda)^2 - 1 = 0$
- λ=2， 4
- $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} x = 0, x = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} x = 0, x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Trace of a Matrix

- tr(A) is the sum of the diagonal elements of the matrix.
- $tr(A) =$ the sum of the eigenvalues.
- $tr(ABCD) = tr(BCDA) = tr(CDAB) = tr(DABC)$

Exercise

- $A = \begin{bmatrix} 7 & 2 & -3 \\ 0 & 6 & 1 \\ -1 & 1 & 8 \end{bmatrix}, B = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix}$

- $AB = ?$

- Eigenvalues of B