[TOC]

Chapter 1 Mathematical Foundation

1.1 The relationship between scalars, vectors, matrices, and tensors

Scalar A scalar represents a single number that is different from most other objects studied in linear algebra (usually an array of multiple numbers). We use italics to represent scalars. Scalars are usually given a lowercase variable name.

Vector A vector represents a set of ordered numbers. By indexing in the order, we can determine each individual number. Usually we give the lowercase variable name of the vector bold, such as xx. Elements in a vector can be represented in italics with a footer. The first element of the vector X is X_1, the second element is X_2, and so on. We will also indicate the type of element (real, imaginary, etc.) stored in the vector.

Matrix A matrix is a collection of objects with the same features and latitudes, represented as a two-dimensional data table. The meaning is that an object is represented as a row in a matrix, and a feature is represented as a column in a matrix, and each feature has a numerical value. The name of an uppercase variable that is usually given to the matrix bold, such as A.

Tensor In some cases, we will discuss arrays with coordinates over two dimensions. In general, the elements in an array are distributed in a regular grid of several dimensional coordinates, which we call a tensor. Use A to represent the tensor "A". The element with a coordinate of (i, j, k) in the tensor A is denoted as $A_{(i,j,k)}$.

Relationship between the four

> The scalar is a 0th order tensor and the vector is a first order tensor. Example: The scalar is the length of the stick, but you won't know where the stick is pointing. Vector is not only knowing the length of the stick, but also knowing whether the stick points to the front or the back. The tensor is not only knowing the length of the stick, but also knowing whether the stick points to the front or the back, and how much the stick is deflected up/down and left/right.

1.2 What is the difference between tensor and matrix?

- From an algebra perspective, a matrix is a generalization of vectors. The vector can be seen as a one-dimensional "table" (that is, the components are arranged in a row in order), the matrix is a two-dimensional "table" (components are arranged in the vertical and horizontal positions), then the n

order tensor is the so-called n dimension "Form". The strict definition of tensors is described using linear mapping.

- Geometrically, a matrix is a true geometric quantity, that is, it is something that does not change with the coordinate transformation of the frame of reference. Vectors also have this property.
- The tensor can be expressed in a 3×3 matrix form.
- A three-dimensional array representing the number of scalars and the representation vector can also be regarded as a matrix of 1 × 1, 1 × 3, respectively.

1.3 Matrix and vector multiplication results

A matrix of m rows of n columns is multiplied by a n row vector, and finally a vector of m rows is obtained. The algorithm is that each row of data in the matrix is treated as a row vector and multiplied by the vector.

1.4 Vector and matrix norm induction

Vector norm Define a vector as: $\vec{a} = [-5, 6, 8, -10]$. Any set of vectors is set to $\vec{x} = (x_1, x_2, ..., x_N)$. The different norms are solved as follows:

- 1 norm of the vector: the sum of the absolute values of the elements of the vector. The 1 norm result of the above vector $\vec{a}$ is: 29.

$$\|\vec{x}\|_1 = \sum_{i=1}^{N} |x_i|$$

- The 2 norm of the vector: the sum of the squares of each element of the vector and the square root. The result of the 2 norm of $\vec{a}$ above is: 15.

$$\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{N} |x_i|^2}$$

- Negative infinite norm of the vector: the smallest of the absolute values of all elements of the vector: the negative infinite norm of the above vector $\vec{a}$ is: 5.

$$\|\vec{x}\|_{-\infty} = \min |x_i|$$

- The positive infinite norm of the vector: the largest of the absolute values of all elements of the vector: the positive infinite norm of the above vector $\vec{a}$ is: 10.

$$\|\vec{x}\|_{+\infty}=\max|x_i|$$

- p-norm of vector:

$$L_p=\|\vec{x}\|_p=\sqrt[p]{\sum_{i=1}^{N}|x_i|^p}$$

Matrix of the matrix

Define a matrix $A=[-1,2,-3;4,-6,6]$. The arbitrary matrix is defined as: $A_{m\times n}$ with elements of a_{ij}.

The norm of the matrix is defined as

$$\|A\|_p:=\sup_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}$$

When the vectors take different norms, different matrix norms are obtained accordingly.

- **1 norm of the matrix (column norm)**: The absolute values of the elements on each column of the matrix are first summed, and then the largest one is taken, (column and maximum), the 1 matrix of the above matrix A The number first gets $[5,8,9]$, and the biggest final result is: 9.

$$\|A\|_1=\max_{1\leq j\leq}\sum_{i=1}^{m}|a_{ij}|$$

- **2 norm of matrix**: The square root of the largest eigenvalue of the matrix A^TA, the final result of the 2 norm of the above matrix A is: 10.0623.

$$\|A\|_2=\sqrt{\lambda_{max}(A^TA)}$$

Where $\lambda_{max}(A^TA)$ is the maximum value of the absolute value of the eigenvalue of A^TA.

- **Infinite norm of the matrix (row norm)**: The absolute values of the elements on each line of the matrix are first summed, and then the largest one (row and maximum) is taken, and the above matrix of A is 1 The number first gets $[6;16]$, and the biggest final result is: 16.

$$\|A\|_\infty=\max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}|$$

- **Matrix kernel norm**: the sum of the singular values of the matrix (decomposed of the matrix svd), this norm can be used for low rank representation (because the minimization of the kernel norm is equivalent to minimizing the rank of the matrix - Low rank), the final result of matrix A above is: 10.9287.
- **Matrix L0 norm**: the number of non-zero elements of the matrix, usually used to represent sparse, the smaller the L0 norm, the more elements, the more sparse, the final result of the above matrix A is :6.
- **Matrix L1 norm**: the sum of the absolute values of each element in the matrix, which is the optimal convex approximation of the L0 norm, so it can also represent sparseness, the final result of the above matrix A is: 22 .
- **F norm of matrix **: the sum of the squares of the elements of the matrix and the square root of the square. It is also commonly called the L2 norm of the matrix. Its advantage is that it is a convex function, which can be solved and easy to calculate. The final result of the above matrix A is: 10.0995.

$$\|A\|_F = \sqrt{(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2)}$$

- **Matrix L21 norm**: matrix first in each column, find the F norm of each column (can also be considered as the vector's 2 norm), and then the result obtained L1 norm (also It can be thought of as the 1 norm of the vector. It is easy to see that it is a norm between L1 and L2. The final result of the above matrix A is: 17.1559.
- **p-norm of the matrix**

$$\|A\|_p = \sqrt[p]{(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^p)}$$

1.5 How to judge a matrix as positive?

- the order master subtype is all greater than 0;
- There is a reversible matrix C such that C^TC is equal to the matrix;
- Positive inertia index is equal to n;
- Contract in unit matrix E (ie: canonical form is E)
- the main diagonal elements in the standard form are all positive;
- the eigenvalues are all positive;
- is a measure matrix of a base.

1.6 Derivative Bias Calculation

Derivative definition:

The derivative represents the ratio of the change in the value of the function to the change in the independent variable when the change in the independent variable tends to infinity. Geometric meaning is the tangent to this point. The physical meaning is the (instantaneous) rate of change at that moment.

Note: In a one-way function, only one independent variable changes, that is, there is only one direction of change rate, which is why the unary function has no partial derivative. There is an average speed and instantaneous speed in physics. Average speed

$$v = \frac{s}{t}$$

Where v represents the average speed, s represents the distance, and t represents the time. This formula can be rewritten as

$$\bar{v} = \frac{\Delta s}{\Delta t} = \frac{s(t_0 + \Delta t) - s(t_0)}{\Delta t}$$

Where Δs represents the distance between two points, and Δt represents the time it takes to walk through this distance. When Δt tends to 0 ($\Delta t \to 0$), that is, when the time becomes very short, the average speed becomes the instantaneous speed at time t_0, expressed as follows :

$$v(t_0) = \lim_{\Delta t \to 0} \bar{v} = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = \lim_{\Delta t \to 0} \frac{s(t_0 + \Delta t) - s(t_0)}{\Delta t}$$

In fact, the above expression represents the derivative of the function s on time t at $t = t_0$. In general, the derivative is defined such that if the limit of the average rate of change exists, there is

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

This limit is called the derivative of the function $y = f(x)$ at point x_0. Remember as $f'(x_0)$ or $y'|_{x=x_0}$ or $\frac{dy}{dx}|_{x=x_0}$ or $\frac{df(x)}{Dx}|_{x=x_0}$.

In layman's terms, the derivative is the slope of the curve at a certain point.

Partial derivative:

Since we talk about partial derivatives, there are at least two independent variables involved. Taking two independent variables as an example, z=f(x,y), from the derivative to the partial derivative, that is, from the curve to the surface. At one point on the curve, there is only one tangent. But at one point on the surface, there are countless lines of tangent. The partial derivative is the rate of change of the multivariate function along the coordinate axis.

Note: Intuitively speaking, the partial derivative is the rate of change of the function along the positive direction of the coordinate axis at a certain point.

Let the function $z = f(x, y)$ be defined in the field of the point (x_0, y_0). When $y = y_0$, z can be regarded as a unary function $f on x (x,y_0)$, if the unary function is derivable at $x = x_0$, there is

$$\lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x} = A$$

The limit of the function A exists. Then say A is the partial derivative of the argument $x = f(x, y)$ at the point (x_0, y_0) about the argument x, denoted as $f_x(x_0, y_0)$ or $Frac\partial z\partial x|_{y=y_0}^{x=x_0}$ or $\frac{\partial f}{\partial x}|_{y=y_0}^{x=X_0}$ or $z_x|_{y=y_0}^{x=x_0}$.

When the partial derivative is solved, another variable can be regarded as a constant and solved by ordinary derivation. For example, the partial derivative of $z = 3x^2 + xy$ for x is $z_x = 6x + y$, this When y is equivalent to the coefficient of x.

The geometric meaning of the partial derivative at a point (x_0, y_0) is the intersection of the surface $z = f(x, y)$ with the face $x = x_0$ or the face $y = y_0$ at $y = y_0$ Or the slope of the tangent at $x = x_0$.

1.7 What is the difference between the derivative and the partial derivative?

There is no essential difference between the derivative and the partial derivative. If the limit exists, it is the limit of the ratio of the change of the function value to the change of the independent variable when the variation of the independent variable tends to zero.

> - Unary function, a y corresponds to a x, and the derivative has only one.
> - A binary function, a z corresponding to a x and a y, has two derivatives: one is the derivative of z to x, and the other is the derivative of z to y, Call it a partial guide.
> - Be careful when seeking partial derivatives. If you refer to one variable, then the other variable is constant. Only the amount of change is derived, and the solution of the partial derivative is transformed into the derivation of the unary function.

1.8 Eigenvalue decomposition and eigenvectors

- eigenvalue decomposition can obtain eigenvalues and eigenvectors;
- The eigenvalue indicates how important this feature is, and the eigenvector indicates what this feature is.

  If a vector $\vec{v}$ is a feature vector of the square matrix A, it will definitely be expressed in the following form:

$$A\nu = \lambda\nu$$

λ is the eigenvalue corresponding to the feature vector $\vec{v}$. Eigenvalue decomposition is the decomposition of a matrix into the following form:

$$A = Q\sum Q^{-1}$$

Where Q is the matrix of the eigenvectors of the matrix A, $\sum$ is a diagonal matrix, and each diagonal element is a eigenvalue, and the eigenvalues are arranged from large to small. The eigenvectors corresponding to these eigenvalues describe the direction of the matrix change (from the primary change to the secondary change arrangement). That is to say, the information of the matrix A can be represented by its eigenvalues and eigenvectors.

1.9 What is the relationship between singular values and eigenvalues?

So how do singular values and eigenvalues correspond? We multiply the transpose of a matrix A by A and the eigenvalues of AA^T, which have the following form:

$$(A^TA)V = \lambda V$$

Here V is the right singular vector above, in addition to:

$$\sigma_i = \sqrt{\lambda_i}, u_i = \frac{1}{\sigma_i}A\mu_i$$

Here σ is the singular value, and u is the left singular vector mentioned above. [Prove that the buddy did not give] The singular value σ is similar to the eigenvalues, and is also ranked from large to small in the matrix $\sum$, and the reduction of σ is particularly fast, in many cases, the first 10% or even the 1% singularity. The sum of the values accounts for more than 99% of the sum of all the singular values. In other words, we can also approximate the description matrix with the singular value of the previous r(r is much smaller than m, n), that is, the partial singular value decomposition:

$$A_{m\times n} \approx U_{m\times r}\sum_{r\times r}V^T_{r\times n}$$

The result of multiplying the three matrices on the right will be a matrix close to A. Here, the closer r is to n, the closer the multiplication will be to A.

1.10 Why should machine use probability?

The probability of an event is a measure of the likelihood that the event will occur. Although the occurrence of an event in a randomized trial is accidental, randomized trials that can be repeated in large numbers under the same conditions tend to exhibit significant quantitative patterns. In addition to dealing with uncertainties, machine learning also needs to deal with random quantities. Uncertainty and randomness may come from multiple sources, using probability theory to quantify uncertainty. Probability theory plays a central role in machine learning because the design of machine learning algorithms often relies on probability assumptions about the data.

> For example, in the course of machine learning (Andrew Ng), there is a naive Bayesian hypothesis that is an example of conditional independence. The learning algorithm makes assumptions about the content to determine if the email is spam. Assume that the probability condition that the word x appears in the message is independent of the word y, regardless of whether the message is spam or not. Obviously this assumption is not without loss of generality, because some words almost always appear at the same time. However, the end result is that this simple assumption has little effect on the results, and in any case allows us to quickly identify spam.

1.11 What is the difference between a variable and a random variable?

Random variable

A real-valued function (all possible sample points) for various outcomes in a random phenomenon (a phenomenon that does not always appear the same result under certain conditions). For example, the number of passengers waiting at a bus stop at a certain time, the number of calls received by the telephone exchange at a certain time, etc., are all examples of random variables. The essential difference between the uncertainty of random variables and fuzzy variables is that the latter results are still uncertain, that is, ambiguity.

**The difference between a variable and a random variable: ** When the probability of the value of the variable is not 1, the variable becomes a random variable; when the probability of the random variable is 1, the random variable becomes a variable.

> For example: When the probability of a variable x value of 100 is 1, then $x = 100$ is determined and will not change unless there is further operation. When the probability of the variable x is 100, the probability of 50 is 0.5, and the probability of 100 is 0.5. Then the variable will change with different conditions. It is a random variable. The probability of 50 or 100 is 0.5, which is 50%.

1.12 The relationship between random variables and probability distribution?

A random variable simply represents a state that may be achieved, and a probability distribution associated with it must be given to establish the probability of each state. The method used to describe the probability of the probability of each possible state of a random variable or a cluster of random variables is the **probability distribution**.

Random variables can be divided into discrete random variables and continuous random variables.

The corresponding function describing its probability distribution is

Probability Mass Function (PMF): Describes the probability distribution of discrete random variables, usually expressed in uppercase letters P.

Probability Density Function (PDF): A probability distribution describing a continuous random variable, usually expressed in lowercase letters p.

1.12.1 Discrete random variables and probability mass functions

PMF maps each state that a random variable can take to a random variable to obtain the probability of that state.

- In general, $P(x)$ represents the probability of $X = x$.
- Sometimes to avoid confusion, explicitly write the name of the random variable $P($x=x) $
- Sometimes you need to define a random variable and then formulate the probability distribution x it follows. Obey $P($x$) $

PMF can act on multiple random variables simultaneously, ie joint probability distribution $P(X = x, Y = y)$* means $X = x$ and the same as $Y = y$ Probability can also be abbreviated as $P(x, y)$.

If a function P is a PMF of the random variable X, then it must satisfy the following three conditions:

- P's domain must be a collection of all possible states
- $∀x∈ $x, $0 \leq P(x) \leq 1$.
- $\sum_{x \in X} P(x) = 1$. We call this property normalized

1.12.2 Continuous Random Variables and Probability Density Functions

If a function p is a PDF of x, then it must satisfy the following conditions

- The domain of p must be a collection of all possible states of xx.
- $\forall x \in X, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$ because $p(x)$ is not the specific probability of representing this state, and Is a relative size (density) of probability. The specific probability requires integration to find.

- $\int p(x)dx = 1$, the score is down, the sum is still 1, and the sum of the probabilities is still 1.

Note: PDF$p(x) $does not directly give a probability to a particular state, giving a density. In contrast, it gives a probability that the area falling within a small area of δx$ is $p(x)\delta x$. Thus, we can't find the probability of a particular state. What we can find is that the probability that a state x falls within a certain interval $[a, b]$ is $\int_a^b p(x)dx$.

1.13 Common probability distribution

1.13.1 Bernoulli Distribution

Bernoulli distribution is a single binary random variable distribution, single parameter ϕ∈[0,1] control, ϕ gives the probability that the random variable is equal to 1. The main properties are:

$$
\begin{aligned}
P(x = 1) &= \phi \\
P(x = 0) &= 1 - \phi \\
P(x = x) &= \phi^x(1 - \phi)^{1-x}
\end{aligned}
$$

Its expectations and variances are:

$$
\begin{aligned}
E_x[x] &= \phi \\
Var_x(x) &= \phi(1 - \phi)
\end{aligned}
$$

Multinoulli distribution is also called **category distribution**, which is a random distribution of individual *k*k values, often used to represent the distribution of **object classifications**. where k is a finite value. Multinoulli distribution consists of Vector $\vec{p}\in[0,1]^{k-1} $parameterized, each component p_i represents the probability of the i state, and $p_k = 1 - 1^T p$.

Scope of application: **Bernoulli distribution** is suitable for modeling **discrete **random variables.

1.13.2 Gaussian distribution

Gauss is also called Normal Distribution. The probability function is as follows:

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

Where μ and σ are mean and variance, respectively. The center peak x coordinate is given by μ, the width of the peak is controlled by σ, and the maximum point is $x=\ Obtainedatmu$, the inflection point is $x=\mu\pm\sigma$

In the normal distribution, the probability of ±1σ, ±2σ, and ±3σ are 68.3%, 95.5%, and 99.73%, respectively. These three numbers are best remembered.

In addition, let $\mu=0,\sigma=1$ Gaussian distribution be reduced to the standard normal distribution:

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

Efficiently evaluate the probability density function:

$$N(x;\mu,\beta^{-1})=\sqrt{\frac{\beta}{2\pi}}exp\left(-\frac{1}{2}\beta(x-\ Mu)^2\right)$$

Among them, $\beta=\frac{1}{\sigma^2}$ controls the distribution precision by the parameter $\beta\in(0,\infty)$.

1.13.3 When is a normal distribution?

Q: When is a normal distribution? Answer: There is no prior knowledge distributed on real numbers. When I don't know which form to choose, the default choice of normal distribution is always wrong. The reasons are as follows:

1. The central limit theorem tells us that many independent random variables approximate a normal distribution. In reality, many complex systems can be modeled as normally distributed noise, even if the system can be structurally decomposed.
2. Normal distribution is the distribution with the greatest uncertainty among all probability distributions with the same variance. In other words, the normal distribution is the distribution with the least knowledge added to the model.

Generalization of normal distribution: The normal distribution can be generalized to the R^n space, which is called the **multiple normal distribution**, whose parameter is a positive definite symmetric matrix $\sum$:

$$N(x;\vec{\mu},\sum)=\sqrt{\frac{1}{2\pi^n det(\sum)}}exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\bar{\sum}1(\vec{x}-\vec{\mu})\right)$$

Efficiently evaluate the probability density of mostly normal distributions:

$$N(x;\vec{\mu},\vec{\beta}^{-1})=\sqrt{det(\vec{\beta})}(2\pi)^n exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\beta(\vec{x}-\vec{\mu})\right)$$

Here, $\vec{\beta}$ is a precision matrix.

1.13.4 Exponential distribution

In deep learning, the exponential distribution is used to describe the distribution of the boundary points at $x=0$. The exponential distribution is defined as follows:

$$p(x;\lambda)=\lambda 1_{x\geq 0}exp(-\lambda x)$$

The exponential distribution uses the indication function $I_{x>=0}$ to make the probability of a negative value of x zero.

1.13.5 Laplace Distribution

A closely related probability distribution is the Laplace distribution, which allows us to set the peak of the probability mass at any point of μ

$$Laplace(x;\mu;\gamma)=\frac{1}{2\gamma}exp\left(-\frac{|x-\mu|}{\gamma}\right)$$

1.13.6

Dirac distribution and empirical distribution

The Dirac distribution ensures that all the masses in the probability distribution are concentrated at one point. The Diract-distributed Dirac δ function (also known as the **unit pulse function**) is defined as follows:

$$p(x)=\delta(x-\mu),x\neq\mu$$

$$\int_a^b\delta(x-\mu)dx=1,a<\mu<b$$

Dirac distribution often appears as an integral part of the empirical distribution

$$\hat{p}(\vec{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\vec{x} - \vec{x}^{(i)})$$

, where m points $x^1, \ldots, x^m$ is the given data set, **experience distribution** will have probability density $\frac{1}{m}$ Assigned to these points.

When we train the model on the training set, we can assume that the empirical distribution obtained from this training set indicates the source of the sample**.

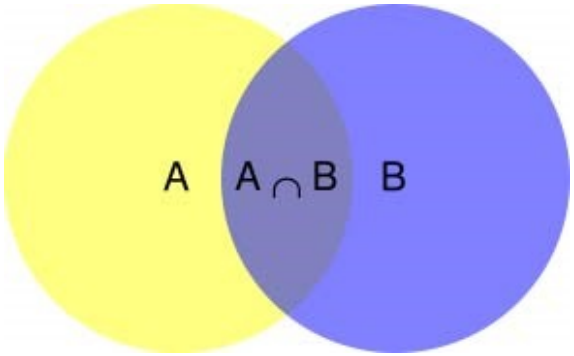
** Scope of application**: The Dirac δ function is suitable for the empirical distribution of **continuous ** random variables.

1.14 Example Understanding Conditional Probability

The conditional probability formula is as follows:

$$P(A/B) = P(A \cap B)/P(B)$$

Description: The event or subset A and B in the same sample space Ω, if an element randomly selected from Ω belongs to B, then the next randomly selected element The probability of belonging to A is defined as the conditional probability of A on the premise of B.



According to the Venn diagram, it can be clearly seen that in the event of event B, the probability of event A occurring is $P(A \bigcap B)$ divided by $P(B)$. Example: A couple has two children. What is the probability that one of them is a girl and the other is a girl? (I have encountered interviews and written tests) **Exhaustive law**: Knowing that one of them is a girl, then the sample space is for men, women, women, and men, and the probability that another is still a girl is 1/3. **Conditional probability method**: $P(female|female) = P(female)/P(female)$, couple has two children, then its sample space is female, male, female, male, male Male, $P(female)$ is 1/4, $P(female) = 1 - P(malemale) = 3/4$, so the last $1/3$. Everyone here may misunderstand that men, women and women are in the same situation, but in fact they are different situations like brothers and sisters.

1.15 What is the difference between joint probability and edge probability?

The difference: Joint Probability: Joint Probability refers to a probability that, like $P(X = a, Y = b)$, contains multiple conditions, and all conditions are true at the same time. Joint probability refers to the probability that multiple random variables satisfy their respective conditions in a multivariate probability distribution. Edge Probability: An edge probability is the probability that an event will occur, regardless of other events. The edge probability refers to a probability similar to $P(X = a)$, $P(Y = b)$, which is only related to a single random variable.

**Contact: ** The joint distribution can find the edge distribution, but if only the edge distribution is known, the joint distribution cannot be obtained.

1.16 The chain rule of conditional probability

From the definition of conditional probability, the following multiplication formula can be directly derived: Multiplication formula Let A, B be two events, and $P(A) > 0$, then

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

Promotion

$$P(ABC) = P(C|AB)P(B|A)P(A)$$

In general, the induction method can be used to prove that if $P(A_1A_2...A_n) > 0$, then there is

$$P(A_1A_2...A_n) = P(A_n|A_1A_2...A_{n-1})P(A_{n-1}|A_1A_2...A_{n-2})...P(A_2|A_1)P(A_1) = P(A_1)\prod_{i=2}^{n} P(A_i|A_1A_2...A_{i-1})$$

Any multi-dimensional random variable joint probability distribution can be decomposed into a conditional probability multiplication form with only one variable.

1.17 Independence and conditional independence

Independence The two random variables x and y, the probability distribution is expressed as a product of two factors, one factor containing only x and the other factor containing only y, and the two random variables are independent. Conditions sometimes bring independence between events that are not independent, and sometimes they lose their independence because of the existence of this condition. Example: $P(XY) = P(X)P(Y)$, event X is independent of event Y. Given Z at this time,

$$P(X,Y|Z) \neq P(X|Z)P(Y|Z)$$

When the event is independent, the joint probability is equal to the product of the probability. This is a very good mathematical nature, but unfortunately, unconditional independence is very rare, because in most cases, events interact with each other.

Conditional independence Given Z, X and Y are conditional, if and only if

$$X \perp Y|Z \iff P(X,Y|Z) = P(X|Z)P(Y|Z)$$

The relationship between X and Y depends on Z, not directly.

> **Example** defines the following events: X: It will rain tomorrow; Y: Today's ground is wet; Z: Is it raining today? The establishment of the >Z event has an impact on both X and Y. However, given the establishment of the Z event, today's ground conditions have no effect on whether it will rain tomorrow.

1.18 Summary of Expectation, Variance, Covariance, Correlation Coefficient

Expectation In probability theory and statistics, the mathematical expectation (or mean, also referred to as expectation) is the sum of the probability of each possible outcome in the trial multiplied by the result. It reflects the average value of random variables.

- Linear operation: $E(ax+by+c) = aE(x) + bE(y) + c$
- Promotion form: $E(\sum_{k=1}^{n} a_i x_i + c) = \sum_{k=1}^{n} a_i E(x_i) + c$
- Function expectation: Let $f(x)$ be a function of x, then the expectation of $f(x)$ is
    - Discrete function: $E(f(x)) = \sum_{k=1}^{n} f(x_k)P(x_k)$
    - Continuous function: $E(f(x)) = \int_{-\infty}^{+\infty} f(x)p(x)dx$

> Note:
>
> - The expectation of the function is not equal to the expected function, ie $E(f(x)) = f(E(x))$
> - In general, the expectation of the product is not equal to the expected product.
> - If X and Y are independent of each other, $E(xy) = E(x)E(y)$.

Variance

The variance in probability theory is used to measure the degree of deviation between a random variable and its mathematical expectation (ie, mean). Variance is a special expectation. defined as:

$$Var(x) = E((x - E(x))^2)$$

> Variance nature:
>
> 1)$Var(x) = E(x^2) - E(x)^2$ 2) The variance of the constant is 0; 3) The variance does not satisfy the linear nature; 4) If X and Y are independent of each other, $Var(ax + by) = a^2Var(x) + b^2Var(y)$

Covariance Covariance is a measure of the linear correlation strength and variable scale of two variables. The covariance of two random variables is defined as:

$$Cov(x, y) = E((x - E(x))(y - E(y)))$$

Variance is a special covariance. When $X = Y$, $Cov(x, y) = Var(x) = Var(y)$.

> Covariance nature:
>
> 1. The covariance of the independent variable is 0.
> 2. Covariance calculation formula:

$$Cov(\sum_{i=1}^{m} a_i x_i, \sum_{j=1}^{m} b_j y_j) = \sum_{i=1}^{m} \sum_{j=1}^{m} a_i b_j Cov(x_i y_i)$$

> 3. Special circumstances:

$$Cov(a + bx, c + dy) = bdCov(x, y)$$

Correlation coefficient The correlation coefficient is the amount by which the linear correlation between the variables is studied. The correlation coefficient of two random variables is defined as:

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}}$$

> The nature of the correlation coefficient:
>
> 1. Bordered. The range of correlation coefficients is , which can be regarded as a dimensionless covariance.
> 2. The closer the value is to 1, the stronger the positive correlation (linearity) of the two variables. The closer to -1, the stronger the negative correlation, and when 0, the two variables have no correlation.