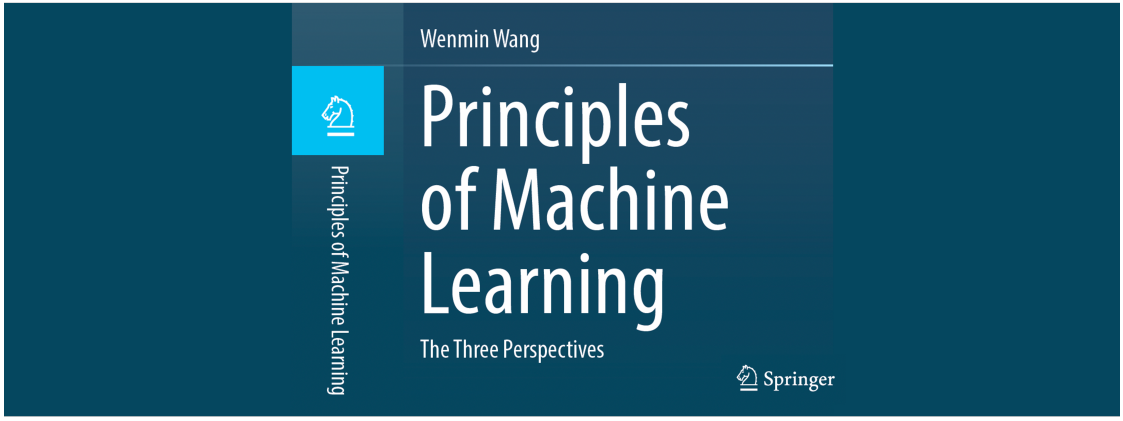

Wenmin Wang
Principles of Machine Learning
The Three Perspectives
Principles of Machine Learning
Springer

Part II Frameworks

4 Statistical Framework

Contents

About the Statistical Framework

- Statistics and probability theory complement each other, so do statistical framework and probabilistic framework.
- The theoretical foundation of the framework is statistics and mathematical statistics.
- The core of the framework is statistical learning theory.

Statistics and Mathematical statistics

Statistics

Statistics is a discipline that collects, organizes, represents, analyzes, interprets, and makes predictions based on data. The main methods include descriptive statistics and inferential statistics.

Mathematical statistics

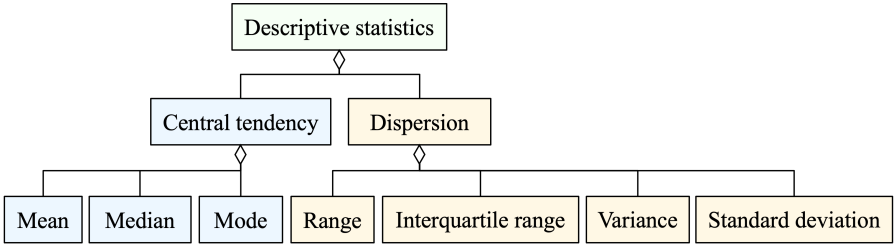
Mathematical statistics, also known as modern statistics, is the application of mathematics in statistics. Its mathematical foundation is probability theory, and it also includes mathematical analysis, linear algebra, stochastic analysis, and differential equations. The main content includes probability distributions, statistical inference, regression analysis, and nonparametric statistics.

Statistical Learning

- Statistical learning is a theoretical analysis framework of machine learning, originating from statistics and functional analysis. The most direct result of statistical learning theory is the support vector machine (SVM) algorithm in machine learning.
- Parametric models and nonparametric models are important models in statistics, especially in statistical learning.
- Kernel methods also play an important role in statistical learning.
- Statistical learning theory is a framework of machine learning, but not equivalent to machine learning, and the view that machine learning is modern statistics is somewhat biased.

About Descriptive Statistics

Descriptive statistics is the summary statistics that quantitatively calculates and describes data through numerical calculations or visualization.



Its numerical calculations mainly include measures of central tendency and measures of dispersion.

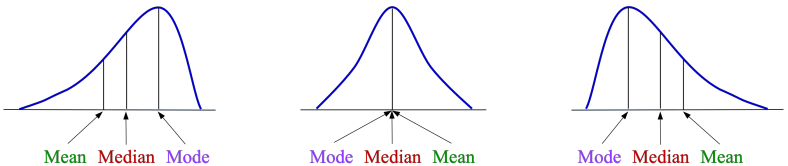
Central Tendency

Central tendency refers to the central value or typical value of a probability distribution, mainly includes the mean, median, and mode.

Let $x_1, x_2, \ldots, x_n$ is a dataset, $\overline{x}_n$ denote mean, and m denote median:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad m = \frac{1}{2}\left(x_{\lfloor (n+1)/2 \rfloor} + x_{\lceil (n+1)/2 \rceil}\right) .$$

And, mode is the number that appears most frequently in the dataset.

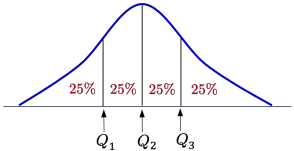


Dispersion

Dispersion measures the change in the distribution of a dataset, consists of range, interquartile range, variance, and standard deviation.

Range is the difference between largest and smallest values.

Interquartile range (IQR) is the difference between upper and lower quartiles.



$$\text{IQR} = Q_3 - Q_1.$$

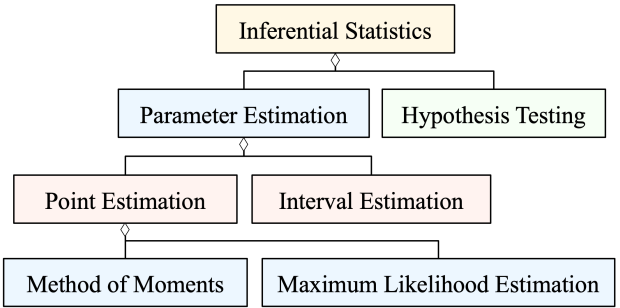
Let σ^2 denote variance, and σ denote standard deviation:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2 \ , \qquad \sigma = \sqrt{\sigma^2}.$$

About Inferential Statistics

Inferential statistics is the process to draw a sample from population, infer the distribution characteristics of the sample through statistical analysis, and infer, predict, and estimate the properties of population.

It consists of parameter estimation and hypothesis testing mainly.

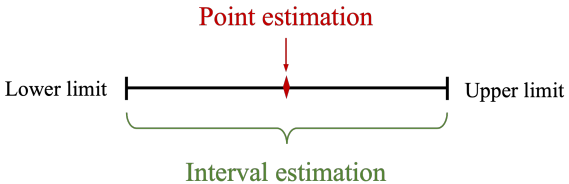


Parameter Estimation

Parameter estimation is to calculate some statistics based on sample data, which is used as an approximation to population parameter.

Two subtypes:

- Point estimation: estimating the parameters of population distribution based on the characteristic value of sample.
- Interval estimation: taking an appropriate interval from the sample to estimate the parameters of population distribution.



Point Estimate vs Interval Estimate

Point Estimation	Interval Estimation
• Method of Moments • Maximum Likelihood Estimation • Bayes Estimation • Expectation–maximization Algorithm	• Inverting A Test Statistic • Pivotal Quantities • Pivoting the Cumulative Distribution Function • Bayesian Intervals

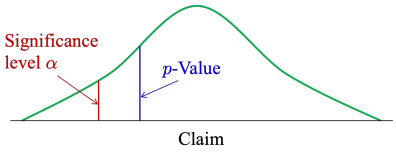
Commonly used methods of point estimate and interval estimate.

Hypothesis Testing

Hypothesis testing is the process of making a statistical hypothesis about the population parameters, and then testing whether this statistical hypothesis holds based on sample data.

Example: Vitamin C and the Flu

- Null hypothesis (H_0): Taking vitamin C cannot prevent the flu.
- Alternative hypothesis (H_1): Taking vitamin C can prevent the flu.
- Significance level (α): Setting to 0.04.
- P-value (p): Calculated to be 0.20.



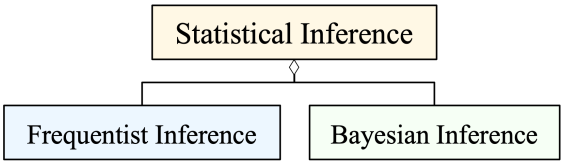
Since $p = 0.18 \gg \alpha = 0.04$, so can not reject the null hypothesis H_0.

About Statistical Inference

Statistical inference is the process of using probability theory methods to infer the probability distribution characteristics of random variables.

It is different from the inferential statistics that is the process of inferring the properties of the population by statistics on the sample.

Two methods: frequentist inference and Bayesian inference.



In machine learning, “inference” is to make predictions with a trained model.

Frequentist Inference

Frequentist inference is to obtain the properties of population by analyzing the frequency or proportion in the sample data, originated from the frequentist interpretation of probability.

It can be represented as $P(D|H)$, i.e. the probability of data given hypothesis:

- Data as random: if same experiment repeated, data changes randomly.
- Hypotheses as fixed: the probability of the hypothesis is $P(H) \rightarrow \{1, 0\}$.

Coin flipping: a classic frequentist experiment, its Bernoulli distribution:

$$f(x; p) = p^x(1 - p)^{1-x}.$$

- If x is equal to 1, then $f(1; p) = p$;
- Else x is equal to 0, $f(0; p) = 1 - p$.

Frequentist Inference

Maximum likelihood estimation (MLE) is often used to determine the parameters of model based on sample data.

Let $D = x_1, x_2, \ldots, x_n$ be the sample, and θ be the parameters of model. Using the likelihood maximization of parameter to obtain its likelihood function,

$$L(D;\theta) = P(D|\theta) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Based on the above likelihood function, we can obtain its MLE, and find the parameter values that maximize the likelihood function:

$$\hat{\theta} = \arg\max_{\theta} L(D;\theta).$$

Bayesian Inference

Bayesian inference is to infer the maximum posterior probability of parameters based on Bayes theorem.

It can be represented as $P(H|D)$, i.e. the probability of hypothesis given data:

- Data as fixed: refers to the data we have.
- Hypotheses as random: the probability of hypothesis is $0 \leq P(H) \leq 1$.

Let D be data, and θ be parameters. Then posterior probability θ given data D:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} \propto P(\theta)P(D|\theta).$$

Consequently, the maximum parameter value of the posterior probability is:

$$\hat{\theta} = \arg\max_{\theta} P(\theta|D).$$

Frequentist Inference vs. Bayesian Inference

Theoretical foundations: Frequentist vs. Bayesian
Conditional probability: $P(\text{data}|\text{hypothesis})$ vs. $P(\text{hypothesis}|\text{data})$

Category	Frequentist Inference	Bayesian Inference
Probability	Long-run frequency	Degree of belief
Hypothesis	Fixed	Random
Data	Random	Fixed
Caculation	$P(H\|D)$	$P(H\|D)$
Math machinery	Repeated experiments	Bayes' theorem

Statistical Model

A statistical model (SM) can be represented as a 3-tuple $\mathrm{SM} = \langle \mathcal{X}, \mathcal{Y}, P \rangle$, where: $\mathcal{X}$ is input space, $\mathcal{Y}$ is output space, and P is probability function.

Using a random data generator, $P(\boldsymbol{x})$, generates a set of observation data D:

$$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}, \text{ where } \boldsymbol{x} \in \mathcal{X}.$$

And using a annotator, $P(y|\boldsymbol{x})$, makes a set of training sample S:

$$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \ldots, (\boldsymbol{x}_n, y_n)\}, \text{ where } y \in \mathcal{Y}.$$

The $P(\boldsymbol{x})$ and $P(y|\boldsymbol{x})$ are two "true but unknown" probability distributions:

- "true": $P(\boldsymbol{x})$ and $P(y|\boldsymbol{x})$ are objectively existing;
- "unknown": $P(\boldsymbol{x})$ and $P(y|\boldsymbol{x})$ cannot be calculated accurately.

Aphorism about Model

A model is a simplification or approximation of reality, but it is not a 100% accurate representation of reality.

There is a common aphorism about model in the field of statistics:

"All models are wrong, but some are useful."

This aphorism is said to come from the British statistician George E. P. Box (1919-2013). But some people believe that this saying is not original to Box, and similar sayings have existed for a long time. But at least it can be said that Box made it a famous aphorism.

This aphorism is considered not only applicable to statistical models, but also often applicable to other scientific models.

Statistical Learning Model

A statistical learning model (SLM) can be seen as a 4-tuple $\text{SLM} = \langle \mathcal{X}, \mathcal{Y}, P, H \rangle$, where $\langle \mathcal{X}, \mathcal{Y}, P \rangle$ is the same as the statistical model, and H is hypothesis set.

Let $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, a training sample set S is generated by:

- Generator $P(\boldsymbol{x})$: generate a set of observation data $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$.
- Annotator $P(y|\boldsymbol{x})$: label each $\boldsymbol{x}_i \in D$ with its corresponding y_i.
- According to $P(\boldsymbol{x}, y) = P(\boldsymbol{x})P(y|\boldsymbol{x})$, we get $S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$.

Using S to train the hypothesis set H, and obtain an optimal hypothesis function $h \in H$, so that $\hat{y} = h(\boldsymbol{x})$ has a smallest error with the annotated output y, that is the loss function $\mathcal{L}(\hat{y}, y)$ is minimized:

$$\arg\min_{\hat{y}=h(\boldsymbol{x})} \mathcal{L}(\hat{y}, y) = \arg\min_{h \in H} \mathcal{L}(h(\boldsymbol{x}), y).$$

Growth Function

Growth function, also known as shatter coefficient or shattering number, is used to measure the complexity of hypothesis set H.

Let $H : \mathcal{X} \to \{0, 1\}$ be hypothesis set, $D \subseteq \mathcal{X}$ be dataset.

Constraint set: H_D is defined as the set of H acting on D:

$$H_D = \{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_n)) \mid \boldsymbol{x}_i \in D, h(\boldsymbol{x}_i) \in H\}.$$

Growth function: $G_H : \mathbb{N} \to \mathbb{N}$ is defined as:

$$G_H(n) = \max_{n=|D|} |H_D|$$

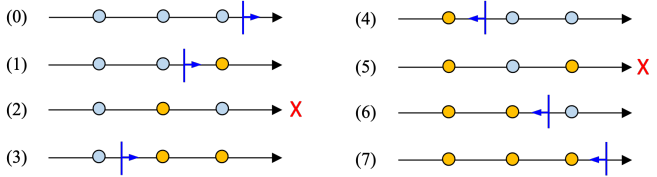
Shattering: D can be shattered by H, if there exists $|H_D| = 2^n$.

Growth Function

Given $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, $\forall \boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^1$, and $h(\boldsymbol{x}_i) \to \{0, 1\}$, $\forall h \in H$.

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$h(x_1)$	0	0	0	0	1	1	1	1
$h(x_2)$	0	0	1	1	0	0	1	1
$h(x_3)$	0	1	0	1	0	1	0	1

Where the combination (2) and (5) cannot be shattered, so $G_H(3) = 8 - 2 = 6$.



VC Dimension

VC dimension is proposed by V. Vapnik and A. Chervonenkis, used to measure the complexity of hypothesis set but easier to calculate than growth function.

VC Dimension: denoted as $\mathrm{VCdim}(H)$, be defined as:

$$\mathrm{VCdim}(H) = \max\{n \mid G_H(n) = 2^n\}$$

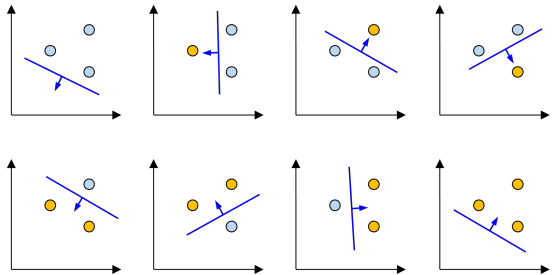
Theorem

Consider some set of n data points in $\mathbb{R}^m$. Choose any one of the data points as the origin, then these n data points can be shattered by oriented hyperplanes, if and only if the position vectors of the remaining data points are linearly independent.

VC Dimension

Corollary

The VC dimension of the set of oriented hyperplanes in $\mathbb{R}^m$ is $m + 1$.



Shattering three data in two-dimensional space.

Parametric Models

For a parameterized statistical model, $\mathrm{PSM} = \langle \mathcal{X}, \mathcal{Y}, P, \Theta \rangle$, it is called a parametric model, if Θ be a set of parameters with a finite number of parameters.

In other words, PSM is called a parametric model, if under $\theta \in \Theta$, the probability over $\boldsymbol{x} \in D \subseteq \mathcal{X}$ is independent of D, i.e.

$$P(\boldsymbol{x}|\theta, D) = P(\boldsymbol{x}|\theta).$$

So the complexity of PSM is bounded even if the amount of D is unbounded.

Parametric Models

Gaussian mixed model (GMM) is a typical parametric model.

$$
\begin{aligned}
P\left(\boldsymbol{x}|\theta\right) &= \sum_{i=1}^{k} \phi_i \mathcal{N}\left(\boldsymbol{x}|\mu_i, \sigma_i\right), \\
\mathcal{N}\left(\boldsymbol{x}|\mu_i, \sigma_i\right) &= \frac{1}{\sqrt{\left(2\pi\right)^k \left|\sigma_i\right|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\mu_i\right)^{\mathsf{T}} \sigma^{-1}\left(\boldsymbol{x}-\mu\right)\right).
\end{aligned}
$$

Where the parameters are μ_i and σ_i, that is, $\theta = (\mu, \sigma)$.

Parametric Models

Examples of other parametric models.

Model Name	Parameter	Expression
Poisson distribution	$\theta = \lambda$	$P(j\|\theta) = \frac{\lambda^j}{j!}\exp(-\lambda)$
Gaussian distribution	$\theta = (\mu, \sigma)$	$p_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
Weibull distribution	$\theta = (\lambda, \beta, \mu)$	$p_\theta(x) = \frac{\beta}{\lambda}\left(\frac{x-\mu}{\lambda}\right)^{\beta-1}\exp\left(-\left(\frac{x-\mu}{\lambda}\right)^{\beta}\right)\mathbb{I}(x > \mu)$
Binomial distribution	$\theta = (n, p)$	$P(k\|\theta) = \frac{n!}{k!(n-k!)}p^k(1-p)^{n-k}$

Nonparametric Models

For a parameterized statistical model $\mathrm{PSM} = \langle \mathcal{X}, \mathcal{Y}, P, \Theta \rangle$, it is called a nonparametric model if Θ is a parameter set without fixed number of parameters.

In other words, PSM is called nonparametric model, if the number of parameters in Θ is not fixed, but increases or decreases with the amount of data in D.

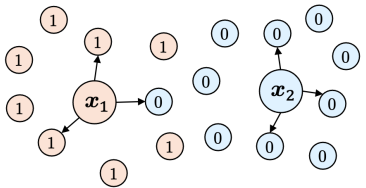
Nonparametric models are not parameter-free, but are parameterized without fixed dimensions, and are therefore also known as variable parametric models.

Nonparametric Models

k-nearest neighbors is a typical algorithm based on nonparametric model.

$$p\left(y=c \mid \boldsymbol{x}, \boldsymbol{D}, k\right)=\frac{1}{k} \sum_{i \in N_k(\boldsymbol{x}, \boldsymbol{D})} \mathbb{I}\left(y_i=c\right),$$

where: $N_k\left(\boldsymbol{x}, \boldsymbol{D}\right)$ is the index of k-nearest neighbors of data point $\boldsymbol{x}$ in $\boldsymbol{D}$; and $\mathbb{I}\left(\omega\right)$ is the indicator function, which equals 1 if ω is true, and 0 otherwise.



A schematic diagram of classifying dataset $\boldsymbol{D}$.

$$\begin{aligned} p\left(y=1 \mid \boldsymbol{x}_1, \boldsymbol{D}, k=3\right) &= 2/3, \\ p\left(y=1 \mid \boldsymbol{x}_2, \boldsymbol{D}, k=3\right) &= 0/3. \end{aligned}$$

Parametric Models vs. Nonparametric Models

Differences between parametric models and nonparametric models.

Parametric Models	Nonparametric Models
Use a fixed number of parameters	Use a non-fixed number of parameters
Parameter analysis is to test means	Nonparametric analysis is to test medians
Only applicable to variables	Applicable to both variables and attributes
Always make strict assumptions about the data	Usually make fewer assumptions about the data
Require fewer data	Require more data
Assume a certain probability distribution	No assumption of the probability distribution
Handle intervals data or ratio data	Handle raw data
Faster computation speed	Slower computation speed

Parametric Models vs. Nonparametric Models

Typical algorithms based on parametric models and nonparametric models.

Algorithms based on Parametric Model	Algorithms based on Nonparametric Model
• Logistic regression • Linear discriminant analysis • *k*-Means • Perceptron • Naïve Bayes	• Decision trees • Gaussian process regression • *k*-Nearest Neighbors • Kernel density estimation • Kernel Support Vector Machine

About Kernel Methods

Kernel methods can calculate the inner product of all data pairs in a high-dimensional feature space, without calculating their coordinates of the data in that space.

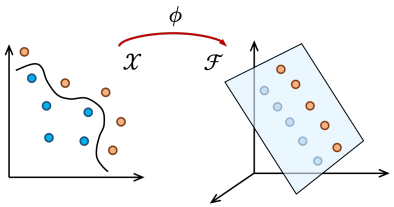
Since inner product operations are much easier than coordinate calculations, this is known as the kernel trick.

Many machine learning algorithms use kernel methods, such as kernel perceptron, kernel SVM, Gaussian processes, principal component analysis (PCA), canonical correlation analysis (CCA), ridge regression, spectral clustering, and linear adaptive filters.

Kernel Trick

In machine learning, we often encounter non-linear problems, that is, the sample data is linearly inseparable. For this, we can consider the following two treatment methods.

(i) based on the sample dataset S, to get a non-linear hypothesis $h \in H$ closest to the target probability function $h \sim P(y|\boldsymbol{x})$, such that $h : \mathcal{X} \to \mathcal{Y}$.



(ii) to map input space to a linearly separable high-dimensional feature space $\mathcal{F}$, i.e. to find a feature mapping ϕ, satisfying: $\phi : \mathcal{X} \to \mathcal{F}$, where $\dim(\mathcal{F}) \gg \dim(\mathcal{X})$.

And find a $h \in H$, such that $h : \mathcal{F} \to \mathcal{Y}$.

Hilbert Space

Given $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^m$, their inner product $\langle \cdot, \cdot \rangle$ can be calculated by:

$$\langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \sum_{i=1}^{m} [x]_i \, [x]'_i.$$

A Hilbert space $\mathcal{F}$ is inner product space with separability and completeness.

Completeness: every Cauchy sequence $\varphi_n \, (n \geq 1)$ converges to φ, satisfies:

$$\sup_{m>n} \parallel \varphi_n - \varphi_m \parallel \to 0, \quad \text{when } \varphi \to \infty.$$

Separability: if $\forall \epsilon > 0$, there exists $\varphi_i \in \mathcal{F}$, such that:

$$\min_i \parallel \varphi_i - \varphi \parallel < \epsilon, \quad \forall \varphi \in F.$$

Kernel Functions

Let $\mathcal{X}$ be a non-empty input space, the symmetric function,

$$\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \, ,$$

is known as the kernel function, also simply called the kernel.

The symmetric function refers to the property as below:

$$\kappa \left(\boldsymbol{x}, \boldsymbol{x}' \right) = \kappa \left(\boldsymbol{x}', \boldsymbol{x} \right) , \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}.$$

The kernel function is to perform feature mapping on all data pairs in the input space to obtain feature vectors located in Hilbert space $\varphi \left(\boldsymbol{x} \right)$ and $\varphi \left(\boldsymbol{x}' \right)$, and then calculate their inner product, i.e.

$$\kappa \left(\boldsymbol{x}, \boldsymbol{x}' \right) = \left\langle \varphi \left(\boldsymbol{x} \right) , \varphi \left(\boldsymbol{x}' \right) \right\rangle .$$

Kernel Functions

Commonly used kernel functions.

Name	Kernel Function
Polynomial kernel	$\kappa(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^d$
Hyperbolic tangent kernel	$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\kappa \langle \boldsymbol{x}, \boldsymbol{x}' \rangle + \Theta)$
Gaussian kernel	$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right)$
Laplacian kernel	$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|x-x'\|}{\sigma}\right)$
Power kernel	$\kappa(x, x') = \| x - x' \|^d$

Thank You