1
Introduction

In this book, we study a system which perceives the real world. Such a system has to estimate an information source by observation. If the information source is a probability distribution, then the estimation process is called statistical learning, and the system is said to be a statistical model or a learning machine.

A lot of statistical models have hierarchical layers, hidden variables, a collection of modules, or grammatical structures. Such models are nonidentifiable and contain singularities in their parameter spaces. In fact, the map from a parameter to a statistical model is not one-to-one, and the Fisher information matrix is not positive definite. Such statistical models are called singular. It has been difficult to examine the learning process of singular models, because there has been no mathematical theory for such models.

In this book, we establish a mathematical foundation which enables us to understand the learning process of singular models. This chapter gives an overview of the book before a rigorous mathematical foundation is developed.

1.1 Basic concepts in statistical learning

To describe what statistical learning is, we need some basic concepts in probability theory. For the reader who is unfamiliar with probability theory, Section 1.6 summarizes the key results.

1.1.1 Random samples

Let N be a natural number and $\mathbb{R}^N$ be the N-dimensional real Euclidean space. We study a case when information data are represented by vectors in $\mathbb{R}^N$.

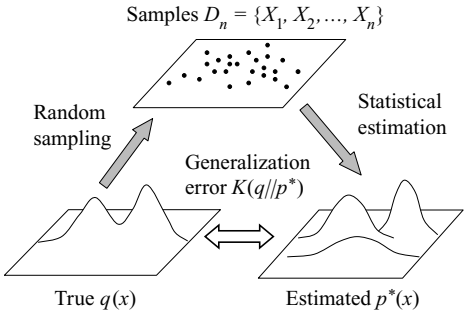

Fig. 1.1. Statistical learning

Firstly, using Figure 1.1, let us explain what statistical learning is. Let $(\Omega, \mathcal{B}, P)$ be a probability space and $X : \Omega \to \mathbb{R}^N$ be a random variable which is subject to a probability distribution $q(x)dx$. Here $q(x)$ is a probability density function and dx is the Lebesgue measure on $\mathbb{R}^N$.

We assume that random variables $X_1, X_2, \ldots, X_n$ are independently subject to the same probability distribution as X, where n is a natural number. In statistical learning theory, $q(x)$ is called a true probability density function, and random variables $X_1, X_2, \ldots, X_n$ are random samples, examples, random data, or training samples. The probability density function of the set of independent random samples is given by

$$q(x_1)q(x_2)\cdots q(x_n).$$

In practical applications, we obtain their realizations by observations. The natural number n is said to be the number of random samples. The set of random samples or random data is denoted by

$$D_n = \{X_1, X_2, \ldots, X_n\}.$$

The main purpose of statistical learning is to construct a method to estimate the true probability density function $q(x)$ from the set D_n.

In this book, we study a method which employs a parametric probability density function. A conditional probability density function $p(x|w)$ of $x \in \mathbb{R}^N$ for a given parameter $w \in W$ is called a learning machine or a statistical model, where W is the set of all parameters. Sometimes the notation $p(x|w) = p_w(x)$ is used. Then $w \mapsto p_w$ gives a map from the parameter to the probability density function. We mainly study the case when W is a subset of the d-dimensional real Euclidean space $\mathbb{R}^d$ or a d-dimensional real analytic manifold. An *a priori* probability density function $\varphi(w)$ is defined on W. We assume that, for any $w \in W$, the support of the probability distribution $p(x|w)$ is equal to that of

$q(x)$ and does not depend on w. That is to say, for any $w \in W$,

$$\overline{\{x \in \mathbb{R}^N; p(x|w) > 0\}} = \overline{\{x \in \mathbb{R}^N; q(x) > 0\}},$$

where $\overline{S}$ is the closure of a set S in $\mathbb{R}^N$.

In statistical learning or statistical inference, the main study concerns a method to produce a probability density function $p^*(x)$ on $\mathbb{R}^N$ based on the set of random samples D_n, using the parametric model $p(x|w)$. Such a function

$$D_n \mapsto p^*(x)$$

is called a statistical estimation method or a learning algorithm. Note that there are a lot of statistical estimation methods and learning algorithms. The probability density function $p^*(x)$, which depends on the set of random variables D_n, is referred to as the estimated or trained probability density function. Generally, it is expected that the estimated probability density function $p^*(x)$ is a good approximation of the true density function $q(x)$, and that it becomes better as the number of random samples increases.

1.1.2 Kullback–Leibler distance

In order to compare two probability density functions, we need a quantitative value which shows the difference between two probability density functions.

Definition 1.1 (Kullback–Leibler distance) For given probability density functions $q(x)$, $p(x) > 0$ on an open set $A \subset \mathbb{R}^N$, the Kullback–Leibler distance or relative entropy is defined by

$$K(q\|p) = \int_A q(x) \log \frac{q(x)}{p(x)} dx.$$

If the integral is not finite, $K(q\|p)$ is defined as $K(q\|p) = \infty$.

Theorem 1.1 *Assume that $q(x)$, $p(x) > 0$ are continuous probability density functions on an open set A. Then the following hold.*
(1) For arbitrary $q(x)$, $p(x)$, $K(q\|p) \geq 0$.
(2) $K(q\|p) = 0$ if and only if $q(x) = p(x)$ for any $x \in A$.

Proof of Theorem 1.1 Let us introduce a real function

$$S(t) = -\log t + t - 1 \quad (0 < t < \infty).$$

Then $S(t) \geq 0$, and $S(t) = 0$ if and only if $t = 1$. Since $\int q(x)dx = \int p(x)dx = 1$,

$$K(q\|p) = \int_A q(x)\, S\Big(\frac{p(x)}{q(x)}\Big)\, dx,$$

which shows (1). Assume $K(q\|p) = 0$. Since $S(p(x)/q(x))$ is a nonnegative and continuous function of x, $S(p(x)/q(x)) = 0$ for any $x \in A$, which is equivalent to $p(x) = q(x)$. □

Remark 1.1 The Kullback–Leibler distance is called the relative entropy in physics. In information theory and statistics, the Kullback–Leibler distance $K(q\|p)$ represents the loss of the system $p(x)$ for the information source $q(x)$. The fact that $K(q\|p)$ is not symmetric for $q(x)$ and $p(x)$ may originate from the difference of their roles. Historically, relative entropy was first defined by Boltzmann and Gibbs in statistical physics in the nineteenth century. In the twentieth century it was found that relative entropy plays a central role in information theory and statistical estimation.

We can measure the difference between the true density function $q(x)$ and the estimated one $p^*(x)$ by the Kullback–Leibler distance:

$$K(q\|p^*) = \int q(x) \log \frac{q(x)}{p^*(x)} dx.$$

In statistical learning theory, $K(q\|p^*)$ is called the generalization error of the method of statistical estimation $D_n \mapsto p^*$. In general, $K(q\|p^*)$ is a measurable function of the set of random samples D_n, hence it is also a real-valued random variable. The training error is defined by

$$K_n(q\|p^*) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p^*(X_i)},$$

which is also a random variable. One of the main purposes of statistical learning theory is to clarify the probability distributions of the generalization and training errors for a given method of statistical estimation. The expectation values $E[K(q\|p^*)]$ and $E[K_n(q\|p^*)]$ are respectively called the mean generalization error and the training error. If the mean generalization error is smaller, the statistical estimation method is more appropriate. The other purpose of statistical learning theory is to establish a mathematical relation between the generalization error and the training error. If the generalization error can be estimated from the training error, we can select the suitable model or hyperparameter among several statistical possible models.

Definition 1.2 (Likelihood function) For a given set of random samples D_n and a statistical model $p(x|w)$, the likelihood function $L_n(w)$ of $w \in W \subset \mathbb{R}^d$ is defined by

$$L_n(w) = \prod_{i=1}^{n} p(X_i|w).$$

If $p(x|w) = q(x)$, then $L_n(w)$ is equal to the probability density function of D_n.

Definition 1.3 (Log likelihood ratio function) For a given true distribution $q(x)$ and a parametric model $p(x|w)$, the log density ratio function $f(x, w)$, the Kullback–Leibler distance $K(w)$, and the log likelihood ratio function $K_n(w)$ are respectively defined by

$$f(x, w) = \log \frac{q(x)}{p(x|w)}, \tag{1.1}$$

$$K(w) = \int q(x) f(x, w) dx, \tag{1.2}$$

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, w), \tag{1.3}$$

where $K_n(w)$ is sometimes referred to as an empirical Kullback–Leibler distance.

From the definition,

$$E[f(X, w)] = E[K_n(w)] = K(w).$$

By using the empirical entropy

$$S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i), \tag{1.4}$$

the likelihood function satisfies

$$-\frac{1}{n} \log L_n(w) = K_n(w) + S_n. \tag{1.5}$$

The empirical entropy S_n does not depend on the parameter w, hence maximization of the likelihood function $L_n(w)$ is equivalent to minimization of $K_n(w)$.

Remark 1.2 If a function $S(t)$ satisfies $S''(t) > 0$ and $S(1) = 0$, then

$$\int_A q(x)\, S\Big(\frac{p(x)}{q(x)}\Big)\, dx$$

has the same property as the Kullback–Leibler distance in Theorem 1.1. For example, using $S(t) = (1 - t^a)/a$ for a given a, $0 < a < 1$, a generalized distance is defined by

$$K^{(a)}(q\|p) = \int q(x)\Big(\frac{1 - (p(x)/q(x))^a}{a}\Big)\, dx.$$

For example, if $a = 1/2$, Hellinger's distance is derived,

$$K^{(1/2)}(q\|p) = \int (\sqrt{q(x)} - \sqrt{p(x)})^2 dx.$$

In general Jensen's inequality claims that, for any measurable function $F(x)$,

$$\int q(x)S(F(x))dx \geq S\Big(\int q(x)F(x)dx\Big),$$

where the equality holds if and only if $F(x)$ is a constant function on $q(x) > 0$. Hence $K^{(a)}(q\|p) \geq 0$ and $K^{(a)}(q\|p) = 0$ if and only if $q(x) = p(x)$ for all x. Hence $K^{(a)}(q\|p)$ indicates a difference of $p(x)$ from $q(x)$. The Kullback–Leibler distance is formally obtained by $a \to +0$. For arbitrary probability density functions $q(x)$, $p(x)$, the Kullback–Leibler distance satisfies $K(q\|p) \geq K^{(a)}(q\|p)$, because

$$K(q\|p) - K^{(a)}(q\|p) = \int q(x)\Big(\frac{af(x,w) + e^{-af(x,w)} - 1}{a}\Big)dx \geq 0.$$

Moreover, if $K(q\|p) \neq 0$ then

$$\lim_{a\to+0} \frac{K_a(q\|p)}{K(q\|p)} = 1.$$

Therefore, from the learning theory of $K(q\|p)$, we can construct a learning theory of $K^{(a)}(q\|p)$.

Remark 1.3 If $E[K(w)] < \infty$ then, by the law of large numbers, the convergence in probability

$$K_n(w) \to K(w)$$

holds for each $w \in W$. Furthermore, if $E[K(w)^2] < \infty$ then, by the central limit theorem,

$$\sqrt{n}(K_n(w) - K(w))$$

converges in law to the normal distribution, for each $w \in W$. Therefore, for each w, the convergence in probability

$$-\frac{1}{n}\log L_n(w) \to K(w) - S$$

holds, where S is the entropy of the true distribution $q(x)$,

$$S = -\int q(x) \log q(x) dx.$$

It might seem that minimization of $K_n(w)$ is equivalent to minimization of $K(w)$. If these two minimization problems were equivalent, then maximization of $L_n(w)$ would be the best method in statistical estimation. However, minimization and expectation cannot be commutative.

$$E[\min_w K_n(w)] \neq \min_w E[K_n(w)] = \min_w K(w). \tag{1.6}$$

Hence maximization of $L_n(w)$ does not mean minimization of $K(w)$. This is the basic reason why statistical learning does not result in a simple optimization problem. To clarify the difference between $K(w)$ and $K_n(w)$, we have to study the meaning of the convergence $K_n(w) \to K(w)$ in a functional space. There are many nonequivalent functional topologies. For example, sup-norm, L^p-norm, weak topology of Hilbert space L^2, Schwartz distribution topology, and so on. It strongly depends on the topology of the function space whether the convergence $K_n(w) \to K(w)$ holds or not. The Bayes estimation corresponds to the Schwartz distribution topology, whereas the maximum likelihood or *a posteriori* method corresponds to the sup-norm. This difference strongly affects the learning results in singular models.

1.1.3 Fisher information matrix

Definition 1.4 (Fisher information matrix) For a given statistical model or a learning machine $p(x|w)$, where $x \in \mathbb{R}^N$ and $w \in \mathbb{R}^d$, the Fisher information matrix

$$I(w) = \{I_{jk}(w)\} \quad (1 \leq j, k \leq d)$$

is defined by

$$I_{jk}(w) = \int \left(\frac{\partial}{\partial w_j} \log p(x|w)\right) \left(\frac{\partial}{\partial w_k} \log p(x|w)\right) p(x|w)\, dx$$

if the integral is finite.

By the definition, the Fisher information matrix is always symmetric and positive semi-definite. It is not positive definite in general. In some statistics textbooks, it is assumed that the Fisher information matrix is positive definite, and that the Cramer–Rao inequality is proven; however, there are a lot of statistical models and learning machines in which Fisher information matrices

have zero eigenvalue. The Fisher information matrix is positive definite if and only if

$$\left\{ \frac{\partial}{\partial w_j} \log p(x|w) \right\}_{j=1}^{d}$$

is linearly independent as a function of x on the support of $p(x|w)$. Since

$$\frac{\partial}{\partial w_j} \log p(x|w) = -\frac{\partial}{\partial w_j} f(x, w),$$

the Fisher information matrix is positive definite if and only if

$$\left\{ \frac{\partial}{\partial w_j} f(x, w) \right\}_{j=1}^{d}$$

is linearly independent as a function of x. By using $\int p(x|w)dx = 1$ for an arbitrary w, it is easy to show that

$$I_{jk}(w) = -\int \left(\frac{\partial^2}{\partial w_j \partial w_k} \log p(x|w) \right) p(x|w)\, dx.$$

If $q(x) = p(x|w_0)$, then

$$I_{jk}(w_0) = \frac{\partial^2}{\partial w_j \partial w_k} K(w_0).$$

Therefore, the Fisher information matrix is equal to the Hessian matrix of the Kullback–Leibler distance at the true parameter.

Remark 1.4 If the Fisher information matrix is positive definite in a neighborhood of the true parameter w_0, $K(w) > 0$ $(w \neq w_0)$ holds, and the Kullback–Leibler distance can be approximated by the positive definite quadratic form, then

$$K(w) \approx \tfrac{1}{2}(w - w_0) \cdot I(w_0)(w - w_0),$$

where $u \cdot v$ shows the inner product of two vectors u, v. If the Fisher information matrix is not positive definite, then $K(w)$ cannot be approximated by any quadratic form in general. This book establishes the mathematical foundation for the case when the Fisher information matrix is not positive definite.

Remark 1.5 (Cramer–Rao inequality) Assume that random samples $\{X_i; i = 1, 2, \ldots, n\}$ are taken from the probability density function $\prod_{i=1}^{n} p(x_i|w)$, where $w = (w_1, w_2, \ldots, w_d) \in \mathbb{R}^d$. A function from random samples to the parameter space

$$\{u_j(x_1, x_2, \ldots, x_n); j = 1, 2, \ldots, d\} \in \mathbb{R}^d$$

is called an unbiased estimator if it satisfies

$$E[u_j(X_1, X_2, \ldots, X_n) - w_j] \equiv \int (u_j(x_1, x_2, \ldots, x_n) - w_j) \times \prod_{i=1}^{n} p(x_i|w) dx_i = 0$$

for arbitrary $w \in \mathbb{R}^d$. Under certain are conditions which ensure that $\int dx_j$ and $(\partial/\partial w_k)$ are commutative for arbitrary j, k,

$$\begin{aligned} 0 &= \frac{\partial}{\partial w_k} E[u_j(X_1, X_2, \ldots, X_n) - w_j] \\ &= E[(u_j - w_j) \sum_{i=1}^{n} \frac{\partial}{\partial w_k} \log p(X_i, w)] - \delta_{jk}. \end{aligned}$$

Therefore,

$$\delta_{jk} = E[(u_j - w_j) \sum_{i=1}^{n} \frac{\partial}{\partial w_k} \log p(X_i, w)].$$

For arbitrary d-dimensional vectors $\mathbf{a} = (a_j)$, $\mathbf{b} = (b_k)$,

$$(\mathbf{a} \cdot \mathbf{b}) = E\Big[\Big(\sum_{j=1}^{d} a_j(u_j - w_j)\Big)\Big(\sum_{k=1}^{d}\sum_{i=1}^{d} b_k \frac{\partial}{\partial w_k} \log p(X_i, w)\Big)\Big].$$

By applying the Cauchy–Schwarz inequality

$$(\mathbf{a} \cdot \mathbf{b})^2 \leq n\ (\mathbf{a} \cdot V\mathbf{a})(\mathbf{b} \cdot I(w)\mathbf{b}), \tag{1.7}$$

where $V = (V_{jk})$ is the covariance matrix of $u - w$,

$$V_{jk} = E[(u_j - w_j)(u_k - w_k)]$$

and $I(w)$ is the Fisher information matrix. If $I(w)$ is positive definite, by putting

$$\begin{aligned} \mathbf{a} &= I(w)^{1/2}\mathbf{c}, \\ \mathbf{b} &= I(w)^{-1/2}\mathbf{c}, \end{aligned}$$

it follows that

$$\|\mathbf{c}\|^2 \leq n\ (\mathbf{c} \cdot I(w)^{1/2} V I(w)^{1/2}\mathbf{c})$$

holds for arbitrary vector $\mathbf{c}$, hence

$$V \geq \frac{I(w)^{-1}}{n}. \tag{1.8}$$

This relation, the Cramer–Rao inequality, shows that the covariance matrix of any unbiased estimator cannot be made smaller than the inverse of the Fisher information matrix. If $I(w)$ has zero eigenvalue and $\mathbf{b}$ is an eigenvector for zero eigenvalue, eq.(1.7) shows that either V is not a finite matrix or no unbiased estimator exists. For statistical models which have a degenerate Fisher information matrix, we have no effective unbiased estimator in general.

1.2 Statistical models and learning machines

1.2.1 Singular models

Definition 1.5 (Identifiablity) A statistical model or a learning machine $p(x|w)$ ($x \in \mathbb{R}^N$, $w \in W \subset \mathbb{R}^d$) is called identifiable if the map

$$W \ni w \mapsto p(\ |w)$$

is one-to-one, in other words,

$$p(x|w_1) = p(x|w_2) \quad (\forall x \in \mathbb{R}^d) \Longrightarrow w_1 = w_2.$$

A model which is not identifiable is called nonidentifiable or unidentifiable.

Definition 1.6 (Positive definite metric) A statistical model or a learning machine $p(x|w)$ ($x \in \mathbb{R}^N$, $w \in W \subset \mathbb{R}^d$) is said to have a positive definite metric if its Fisher information matrix $I(w)$ is positive definite for arbitrary $w \in \mathbf{W}$. If a statistical model does not have a positive definite metric, it is said to have a degenerate metric.

Definition 1.7 (Singular statistical models) Assume that the support of the statistical model $p(x|w)$ is independent of w. A statistical model $p(x|w)$ is said to be regular if it is identifiable and has a positive definite metric. If a statistical model is not regular, then it is called strictly singular. The set of singular statistical models consists of both regular and strictly singular models.

Mathematically speaking, identifiability is neither a necessary nor a sufficient condition of positive definiteness of the Fisher information matrix. In fact, if $p(x|a)$ ($x, a \in \mathbb{R}^1$) is a regular statistical model, then $p(x|a^3)$ is identifiable but has a degenerate Fisher information matrix. Also $p(x|a^2)$ ($|a| > 1$) has a nondegenerate Fisher information matrix but is nonidentifiable. These are trivial examples in which an appropriate transform or restriction of a parameter makes models regular.

However, a lot of statistical models and learning machines used in information science have simultaneously nonidentifiability and a degenerate metric.