CHAPTER

STATISTICAL ESTIMATION 9

CHAPTER CONTENTS

So far, various properties of random variables and probability distributions have been discussed. However, in practice, probability distributions are often unknown and only samples are available. In this chapter, an overview of *statistical estimation* for identifying an underlying probability distribution from samples is provided.

9.1 FUNDAMENTALS OF STATISTICAL ESTIMATION

A quantity estimated from samples is called an *estimator* and is denoted with a "hat." For example, when the expectation $\boldsymbol{\mu}$ of a probability distribution is estimated by the sample average, its estimator is denoted as

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i .$$

An estimator is a function of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and thus is a random variable. On the other hand, if particular values are plugged in the estimator, the obtained value is called an *estimate*.

A set of probability mass/density functions described with a finite-dimensional parameter $\boldsymbol{\theta}$ is called a *parametric model* and is denoted by $g(\boldsymbol{x};\boldsymbol{\theta})$. In the notation $g(\boldsymbol{x};\boldsymbol{\theta})$, $\boldsymbol{x}$ before the semicolon is a random variable and $\boldsymbol{\theta}$ after the semicolon is a parameter. For example, a parametric model corresponding to the d-dimensional

An Introduction to Statistical Machine Learning. DOI: 10.1016/B978-0-12-802121-7.00020-0

normal distribution,

$$g(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

has expectation vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ as parameters.

An approach to statistical estimation by identifying the parameter in a parametric model is called a *parametric method*, while a *nonparametric method* does not use parametric models or uses a parametric model having infinitely many parameters.

Below, samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ are assumed i.i.d. with $f(\boldsymbol{x})$ (see Section 7.3).

9.2 POINT ESTIMATION

Point estimation gives a best estimate of an unknown parameter from samples. Since methods of point estimation will be extensively explored in Part 3 and Part 4, only a brief overview is provided in this section.

9.2.1 PARAMETRIC DENSITY ESTIMATION

Maximum likelihood estimation determines the parameter value so that samples at hand, $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$, are generated most probably. The *likelihood* is the probability that the samples $\mathcal{D}$ are generated:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\boldsymbol{x}_i;\boldsymbol{\theta}),$$

and maximum likelihood estimation maximizes the likelihood:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}),$$

where $\operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ is the maximizer of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. See Chapter 12, details of maximum likelihood estimation.

The parameter $\boldsymbol{\theta}$ is regarded as a deterministic variable in maximum likelihood estimation, while it is regarded as a random variable in *Bayesian inference*. Then the following probabilities can be considered:

$$\begin{aligned} \text{Prior probability: } & p(\boldsymbol{\theta}),\\ \text{Likelihood: } & p(\mathcal{D}|\boldsymbol{\theta}),\\ \text{Posterior probability: } & p(\boldsymbol{\theta}|\mathcal{D}). \end{aligned}$$

Typical Bayesian point-estimators are given as the posterior expectation or the posterior mode:

$$\begin{aligned} \text{Posterior expectation: } & \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta},\\ \text{posterior mode: } & \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathcal{D}). \end{aligned}$$

Estimating the posterior mode is often called *maximum a posteriori probability estimation*. The posterior probability can be computed by *Bayes' theorem* explained in Section 5.4 as

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'}.$$

Thus, given the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ and the prior probability $p(\boldsymbol{\theta})$, the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ can be computed. However, the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ depends on subjective choice of the prior probability $p(\boldsymbol{\theta})$ and its computation can be cumbersome if the posterior probability $p(\boldsymbol{\theta}|\mathcal{D})$ has a complex profile. See Chapter 17, Section 17.3, and Section 19.3 for the details of Bayesian inference.

Maximum likelihood estimation is sometimes referred to as *frequentist inference* when it is contrasted to Bayesian inference.

9.2.2 NONPARAMETRIC DENSITY ESTIMATION

Kernel density estimation (KDE) is a nonparametric technique to approximate the probability density function $f(\boldsymbol{x})$ from samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ as

$$\widehat{f}_{\mathrm{KDE}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K(\boldsymbol{x},\boldsymbol{x}_i),$$

where $K(\boldsymbol{x},\boldsymbol{x}')$ is a *kernel function*. Typically, the Gaussian kernel function,

$$K(\boldsymbol{x},\boldsymbol{x}') = \frac{1}{(2\pi h^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2h^2}\right),$$

is used, where $h > 0$ is the *bandwidth* of the Gaussian function and d denotes the dimensionality of $\boldsymbol{x}$, and $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ denotes the *Euclidean norm*.

Nearest neighbor density estimation (NNDE) is another nonparametric method given by

$$\widehat{f}_{\mathrm{NNDE}}(\boldsymbol{x}) = \frac{k\Gamma(\frac{d}{2}+1)}{n\pi^{\frac{d}{2}}\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}_k\|^d},$$

where $\widetilde{\boldsymbol{x}}_k$ denotes the kth nearest sample to $\boldsymbol{x}$ among $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ and $\Gamma(\cdot)$ denotes the gamma function explained in Section 4.3.

See Chapter 16 for the derivation and properties of nonparametric density estimation.

9.2.3 REGRESSION AND CLASSIFICATION

Regression is a problem to estimate a function from d-dimensional input $\boldsymbol{x}$ to a real scalar output y based on input-output paired samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.

The method of *least squares* (LS) fits a regression model $r(\boldsymbol{x};\boldsymbol{\alpha})$ to data by minimizing the squared sum of residuals:

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{LS}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \big(y_i - r(\boldsymbol{x}_i;\boldsymbol{\alpha})\big)^2.$$

A nonparametric Gaussian kernel model is a popular choice as a regression model:

$$r(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{j=1}^{n} \alpha_j \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_j\|^2}{2h^2}\right),$$

where $h > 0$ is the *bandwidth* of the Gaussian kernel. To avoid overfitting to noisy samples, *regularization* (see Chapter 23) is effective:

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{RLS}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \big(y_i - r(\boldsymbol{x}_i;\boldsymbol{\alpha})\big)^2 + \lambda\|\boldsymbol{\alpha}\|^2\right],$$

where $\lambda \geq 0$ is the regularization parameter to control the strength of regularization. The LS method is equivalent to maximum likelihood estimation if output y is modeled by the normal distribution with expectation $r(\boldsymbol{x};\boldsymbol{\alpha})$:

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\big(y - r(\boldsymbol{x};\boldsymbol{\alpha})\big)^2}{2\sigma^2}\right).$$

Similarly, the regularized LS method is equivalent to Bayesian maximum *a posteriori* probability estimation if the normal prior probability,

$$\frac{1}{(2\pi\lambda^2)^{n/2}} \exp\left(-\frac{\|\boldsymbol{\alpha}\|^2}{2\lambda^2}\right),$$

is used for parameter $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)^\top$. See Chapter 22, Chapter 23, Chapter 24, and Chapter 25 for the details of regression.

When output value y takes c discrete *categorical* value, the function estimation problem is called *classification*. When $c = 2$, setting $y = \pm 1$ allows us to naively use (regularized) LS regression in classification. See Chapter 26, Chapter 27, Chapter 30, and Chapter 28 for the details of classification.

9.2.4 MODEL SELECTION

The performance of statistical estimation methods depends on the choice of tuning parameters such as the regularization parameters and the Gaussian bandwidth. Choosing such tuning parameter values based on samples is called *model selection*.

In the frequentist approach, *cross validation* is the most popular model selection method: First, samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ are split into k disjoint subsets $\mathcal{D}_1,\ldots,\mathcal{D}_k$. Then statistical estimation is performed with $\mathcal{D}\backslash\mathcal{D}_j$ (i.e., all samples without

$\mathcal{D}_j$), and its estimation error (such as the log-likelihood in density estimation, the squared error in regression, and the misclassification rate in classification) for $\mathcal{D}_j$ is computed. This process is repeated for all $j = 1, \ldots, k$ and the model that minimizes the average estimation error is chosen as the most promising one. See Chapter 14 for the details of frequentist model selection.

In the Bayesian approach, the model $\mathcal{M}$ that maximizes the *marginal likelihood*,

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) \mathrm{d}\boldsymbol{\theta},$$

is chosen as the most promising one. This approach is called *type-II maximum likelihood estimation* or the *empirical Bayes* method. See Section 17.4 for the details of Bayesian model selection.

9.3 INTERVAL ESTIMATION

Since an estimator $\widehat{\theta}$ is a function of samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$, its value depends on the realizations of the samples. Thus, it would be practically more informative if not only a point-estimated value but also its reliability is provided. The interval that an estimator $\widehat{\theta}$ is included with probability at least $1 - \alpha$ is called the *confidence interval* with *confidence level* $1 - \alpha$. In this section, methods for estimating the confidence interval are explained.

9.3.1 INTERVAL ESTIMATION FOR EXPECTATION OF NORMAL SAMPLES

For one-dimensional i.i.d. samples $x_1, \ldots, x_n$ with normal distribution $N(\mu, \sigma^2)$, if the expectation μ is estimated by the sample average,

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

the standardized estimator

$$z = \frac{\widehat{\mu} - \mu}{\sigma / \sqrt{n}}$$

follows the standard normal distribution $N(0, 1)$. Thus, the confidence interval of $\widehat{\mu}$ with confidence level $1 - \alpha$ can be obtained as

$$\left[\widehat{\mu} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \widehat{\mu} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right],$$

where $[-z_{\alpha/2}, +z_{\alpha/2}]$ corresponds to the middle $1 - \alpha$ probability mass of the standard normal density (see Fig. 9.1). However, to compute the confidence interval in practice, knowledge of the standard deviation σ is necessary.

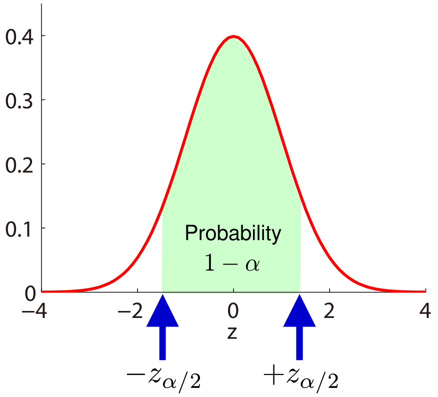

FIGURE 9.1

Confidence interval for normal samples.

When σ is unknown, it is estimated from samples as

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2}.$$

In this case, an estimator standardized with $\widehat{\sigma}$,

$$t = \frac{\widehat{\mu} - \mu}{\widehat{\sigma}/\sqrt{n}},$$

follows the t-distribution with $n-1$ degrees of freedom (Section 4.6). For this reason, standardization with $\widehat{\sigma}$ is sometimes called *Studentization*.

The middle $1-\alpha$ probability mass of the t-density (see Section 4.6) gives the confidence interval with confidence level $1-\alpha$ as

$$\left[\widehat{\mu} - \frac{\widehat{\sigma}}{\sqrt{n}}t_{\alpha/2}, \widehat{\mu} + \frac{\widehat{\sigma}}{\sqrt{n}}t_{\alpha/2}\right],$$

where $[-t_{\alpha/2}, +t_{\alpha/2}]$ corresponds to the middle $1-\alpha$ probability mass of the t-density with $n-1$ degrees of freedom (as in Fig. 9.1). As shown in Fig. 4.9, the t-density has heavier tails than the normal density.

9.3.2 BOOTSTRAP CONFIDENCE INTERVAL

The above method for computing the confidence interval is applicable only to the average of normal samples. For statistics other than the expectation estimated from samples following a non-normal distribution, the probability distribution of the

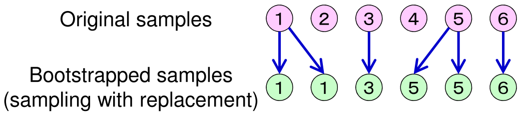

FIGURE 9.2

Bootstrap resampling by sampling with replacement.

estimator cannot be explicitly obtained in general. In such a situation, the use of *bootstrap* allows us to numerically compute the confidence interval.

In the bootstrap method, n pseudo samples $\mathcal{D}' = \{\boldsymbol{x}_i'\}_{i=1}^n$ are gathered by *sampling with replacement* from the original set of samples $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$. Because of sampling with replacement, some samples in the original set $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ may be selected multiple times and others may not be selected in $\mathcal{D}' = \{\boldsymbol{x}_i'\}_{i=1}^n$ (Fig. 9.2). From the bootstrapped samples $\mathcal{D}' = \{\boldsymbol{x}_i'\}_{i=1}^n$, an estimator $\widehat{\theta}'$ of the target statistic is computed. These resampling and estimation procedures are repeated many times and the histogram of the estimator $\widehat{\theta}'$ can be constructed. Extracting the middle $1 - \alpha$ probability mass of the histogram (as in Fig. 9.1) gives the confidence interval $[-b_{\alpha/2}, +b_{\alpha/2}]$ with confidence level $1 - \alpha$.

As illustrated above, the bootstrap method allows us to construct the confidence interval for any statistic and any probability distribution. Furthermore, not only the confidence interval but also any statistics such as the variance and higher-order moments of any estimator can be numerically evaluated by the bootstrap method. However, since the resampling and estimation procedures need to be repeated many times, the computation cost of the bootstrap method can be expensive.

9.3.3 BAYESIAN CREDIBLE INTERVAL

In Bayesian inference, the middle $1 - \alpha$ probability mass of the posterior probability $p(\theta|\mathcal{D})$ corresponds to the confidence interval with confidence level $1 - \alpha$. This is often referred to as the *Bayesian credible interval*. Thus, in Bayesian inference, the confidence interval can be naively obtained without additional computation. However, if the posterior probability $p(\theta|\mathcal{D})$ has a complex profile, computation of the confidence interval can be cumbersome. Moreover, dependency of the confidence interval on subjective choice of the prior probability $p(\theta)$ can be an issue in practice.