Chapter 2
General Framework of Mathematics

With the explosive growth of data nowadays, a young and interdisciplinary field, *data science*, has emerged, which uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured. This data science field is becoming popular and needs to be developed urgently so that it can serve and guide for the industry of the society. Rigorously, applied *data science* is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Within the field of data analytics, *machine learning* is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. The name *machine learning* was coined in 1959 by Arthur Samuel, which evolved from the study of pattern recognition and computational learning theory in artificial intelligence. *Computational statistics*, which also focuses on prediction-making through the use of computers, is a closely related field and often overlaps with *machine learning*.

The name, *computational statistics*, implies that it is composed of two indispensable parts, statistics inference models and the corresponding algorithms implemented in computers. Based on the different kinds of hypotheses, statistics inference can be divided into two schools, frequentist inference school and Bayesian inference school. Here, we describe each one briefly. Let $\mathcal{P}$ be a premise and $\mathcal{O}$ be an observation which may give evidence for $\mathcal{P}$. The priori $P(\mathcal{P})$ is the probability that $\mathcal{P}$ is true before the observation is considered. Also, the posterior $P(\mathcal{P}|\mathcal{O})$ is the probability that $\mathcal{P}$ is true after the observation $\mathcal{O}$ is considered. The likelihood $P(\mathcal{O}|\mathcal{P})$ is the chance of observation $\mathcal{O}$ when evidence $\mathcal{P}$ exists. Finally, $P(\mathcal{O})$ is the total probability, calculated in the following way:

B. Shi, S. S. Iyengar, *Mathematical Theories of Machine Learning - Theory and Applications*, https://doi.org/10.1007/978-3-030-17076-9_2

$$P(\mathcal{O}) = \sum_{\mathcal{P}} P(\mathcal{O}|\mathcal{P})\, P(\mathcal{P}).$$

Connecting the probabilities above is the significant Bayes' formula in the theory of probability

$$P(\mathcal{P}|\mathcal{O}) = \frac{P(\mathcal{O}|\mathcal{P})\, P(\mathcal{P})}{P(\mathcal{O})} \sim P(\mathcal{O}|\mathcal{P})\, P(\mathcal{P}), \tag{2.1}$$

where $P(\mathcal{O})$ can be calculated automatically if we have known the likelihood $P(\mathcal{O}|\mathcal{P})$ and $P(\mathcal{P})$. If we presume that some hypothesis (parameter specifying the conditional distribution of the data) is true and that the observed data is sampled from that distribution, that is,

$$P(\mathcal{P}) = 1,$$

only using conditional distributions of data given the specific hypotheses are the view of the frequentist school. However, if there is no presumption that some hypothesis (parameter specifying the conditional distribution of the data) is true, that is, there is a prior probability for the hypothesis $\mathcal{P}$,

$$\mathcal{P} \sim P(\mathcal{P}),$$

summing up the information from the prior and likelihood is the view from the Bayesian school. Apparently, the view from the frequentist school is a special case of the view from the Bayesian school, but the view from the Bayesian school is more comprehensive and requires more information.

Take the Gaussian distribution with known variance for the likelihood as an example. Without loss of generality, we assume the variance $\sigma^2 = 1$. In other words, the data point is viewed as a random variable $\mathbf{X}$ following the rule below:

$$\mathbf{X} \sim P(x|\mathcal{P}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}},$$

where the hypothesis is $\mathcal{P} = \{\mu | \mu \in (-\infty, \infty) \text{ is some fixed real number}\}$. Let the data set be $\mathcal{O} = \{x_i\}_{i=1}^n$. The frequentist school requires us to compute maximum likelihood or maximum log-likelihood, that is,

$$\begin{aligned}
\underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} f(\mu) &= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log P(\mathcal{O}|\mathcal{P}) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \left(\log \prod_{i=1}^{n} P(x_i \in \mathcal{O}|\mathcal{P})\right) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log\left[\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}}\right] \\
&= -\underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + n\log\sqrt{2\pi}\right],
\end{aligned} \tag{2.2}$$

which has been shown in the classical textbooks, such as [RS15], whereas the Bayesian school requires to compute maximum posterior estimate or maximum log-posterior estimate, that is, we need to assume reasonable prior distribution.

- If the prior distribution is a Gauss distribution $\mu \sim \mathcal{N}(0, \sigma_0^2)$, we have

$$\begin{aligned}
&\underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} f(\mu) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log P(\mathcal{O}|\mathcal{P})\, P(\mathcal{P}) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log\left(\prod_{i=1}^{n} \log P(x_i \in \mathcal{O}|\mathcal{P})\right) P(\mathcal{P}) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log\left\{\left[\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}}\right]\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right) e^{-\frac{\mu^2}{2\sigma_0^2}}\right\} \\
&= -\underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{2\sigma_0^2}\cdot\mu^2 + n\log\sqrt{2\pi} + \log\sqrt{2\pi}\sigma_0\right].
\end{aligned} \tag{2.3}$$

- If the prior distribution is a Laplace distribution $\mu \sim \mathcal{L}(0, \sigma_0^2)$, we have

$$\begin{aligned}
&\max_{\mu\in(-\infty,\infty)} f(\mu) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log P(\mathcal{O}|\mathcal{P})\, P(\mathcal{P}) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log\left(\prod_{i=1}^{n} \log P(x_i \in \mathcal{O}|\mathcal{P})\right) P(\mathcal{P}) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \log\left\{\left[\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}}\right]\cdot\left(\frac{1}{2\sigma_0^2}\right) e^{-\frac{|\mu|}{\sigma_0^2}}\right\} \\
&= -\underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{\sigma_0^2}\cdot|\mu| + n\log\sqrt{2\pi} + \log 2\sigma_0^2\right].
\end{aligned} \tag{2.4}$$

- If the prior distribution is the mixed distribution combined with Laplace distribution and Gaussian distribution $\mu \sim \mathcal{M}(0, \sigma_{0,1}^2, \sigma_{0,2}^2)$, we have

$$
\begin{aligned}
&\underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \; f(\mu) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \; \log P(\mathcal{O}|\mathcal{P})\, P(\mathcal{P}) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \; \log \left(\prod_{i=1}^{n} \log P(x_i \in \mathcal{O}|\mathcal{P}) \right) P(\mathcal{P}) \\
&= \underset{\mu\in(-\infty,\infty)}{\operatorname{argmax}} \; \log \left\{ \left[\left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}} \right] \right. \\
&\qquad\qquad \left. \cdot\, C(\sigma_{0,1}, \sigma_{0,2})^{-1} e^{-\frac{|\mu|}{\sigma_{0,1}^2} - \frac{\mu^2}{2\sigma_{0,2}^2}} \right\} \\
&= - \underset{\mu\in(-\infty,\infty)}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{\sigma_0^2} \cdot |\mu| + \frac{1}{2\sigma_{0,2}^2} \cdot \mu^2 \right. \\
&\qquad\qquad \left. + n \log \sqrt{2\pi} + \log C(\sigma_{0,1}, \sigma_{0,2}) \right],
\end{aligned}
\tag{2.5}
$$

  where $C = 2\sqrt{2\pi}\sigma_{0,1}^2\sigma_{0,2}$.

2.1 Concluding Remarks

In summary, based on the description in this chapter, a statistical problem can be solved by transforming it into an optimization problem. The required proof to validate this statement was outlined and provided in this chapter. In the following chapter we discuss the problem further by identifying how it is formulated and we develop an approach to tackle the problem.