L E C T U R E 12

§17. LOSS FUNCTIONS AND BAYESIAN ESTIMATORS.

The outcome of a Bayesian analysis is the posterior distribution, which combines the prior information and the information from data. However, sometimes we may want to summarize the posterior information with a scalar, for example the mean, median or mode of the posterior distribution. In the following, we show how the use of scalar estimator can be justified using statistical decision theory.

In the classical approach to point estimation, we adopted the criterion that selects the decision function that is most efficient. That is, we choose from all possible unbiased estimators the one with the smallest variance as our best estimator. In decision theory we also take into account the rewards for making correct decision and the penalties for making incorrect decision.

It is convenient now to introduce a **loss function** whose values depend on the true value of the parameter θ and the estimator $\hat{\theta}$.

Let $L(\theta, \hat{\theta})$ denote the loss function which gives the cost of using $\hat{\theta} = \hat{\theta}(x_1, x_2, ..., x_n)$ as an estimate for θ. We define that $\hat{\theta}$ is a Bayes estimate of θ if it minimizes the posterior expected loss

$$E[L(\theta, \hat{\theta})|x_1, x_2, ..., x_n] = \int L(\theta, \hat{\theta}) p(\theta|x_1, x_2, ..., x_n)\, d\theta. \tag{17.1}$$

On the other hand, the expectation of the loss function over the sampling distribution of $x_1, x_2, ..., x_n$ is called risk function:

$$R(\theta, \hat{\theta}) = E[L(\theta, \hat{\theta})|\theta] = \int L(\theta, \hat{\theta}) p(x_1, x_2, ..., x_n|\theta)\, dx_1 dx_2 ... dx_n. \tag{17.2}$$

Further, the expectation of the risk function over the prior distribution of θ,

$$E[R(\theta, \hat{\theta})] = \int R(\theta, \hat{\theta}) p(\theta)\, d\theta, \tag{17.3}$$

is called Bayes risk.

By changing the order of integration one can see that the Bayes risk

$$\int R(\theta, \hat{\theta})\, p(\theta)\, d\theta = \int p(\theta) \int L(\theta, \hat{\theta})\, p(x_1, x_2, ..., x_n|\theta)\, dx_1 dx_2 ... dx_n\, d\theta =$$

$$= \int p(x_1, x_2, ..., x_n) \int L(\theta, \hat{\theta})\, p(\theta|x_1, x_2, ..., x_n)\, d\theta\, dx_1 dx_2 ... dx_n \tag{17.4}$$

is minimized when the inner integral in (17.4) is minimized for each $dx_1 dx_2 ... dx_n$, that is, when a Bayes estimator is used.

In the following, we introduce the Bayes estimators for three simple loss functions.

Zero-one loss function:

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{when } |\hat{\theta} - \theta| < a \\ 1, & \text{when } |\hat{\theta} - \theta| \geq a. \end{cases}$$

We should minimize

$$\int_{-\infty}^{\infty} L(\theta, \hat{\theta}) p(\theta|x_1, x_2, ..., x_n)\, d\theta = \int_{-\infty}^{\hat{\theta}-a} p(\theta|x_1, x_2, ..., x_n)\, d\theta + \int_{\hat{\theta}+a}^{\infty} p(\theta|x_1, x_2, ..., x_n)\, d\theta =$$

$$= 1 - \int_{\hat{\theta}-a}^{\hat{\theta}+a} p(\theta|x_1, x_2, ..., x_n)\, d\theta,$$

or maximize

$$\int_{\hat{\theta}-a}^{\hat{\theta}+a} p(\theta|x_1, x_2, ..., x_n)\, d\theta.$$

If $p(\theta|x_1, x_2, ..., x_n)$ is unimodal, maximization is achieved by choosing $\hat{\theta}$ to be the midpoint of the interval of length $2a$ for which $p(\theta|x_1, x_2, ..., x_n)$ has the same value at both ends. If we let $a \to 0$, then $\hat{\theta}$ tends to the mode of the posterior distribution.

Absolute error loss function:

$$L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|.$$

In general, if η is a random variable, then the expectation $E(|\eta - d|)$ is minimized by choosing d to be the median of the distribution of η. Thus, the Bayes estimate of θ is the posterior median.

Quadratic loss function:

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2.$$

In general, if η is a random variable, then the expectation $E[(\eta - d)^2]$ is minimized by choosing d to be the mean of the distribution of η.

Theorem 17.1. For any random variable η

$$\min_{d\in R} E[(\eta - d)^2] = E(\eta - E(\eta))^2 = \mathrm{Var}_\theta(\eta).$$

Proof: We have

$$E[(\eta - d)^2] = E[(\eta - E(\eta) + E(\eta) - d)^2] =$$

$$= E[(\eta - E(\eta)^2] + 2[E(\eta) - d]\, E[\eta - E(\eta)] + [E(\eta) - d]^2 =$$

$$= E[(\eta - E(\eta)^2] + [E(\eta) - d]^2 = \mathrm{Var}_\theta(\eta) + [E(\eta) - d]^2.$$

Since $[E(\eta) - d]^2 \geq 0$, the proof is complete.

Thus, the Bayes estimate of θ is the posterior mean.

Quadratic loss function

Theorem 17.2. If

$$\hat{\theta} = \arg\min_{\hat{\theta}} \int (\theta - \hat{\theta})^2 p(\theta|x_1, x_2, ..., x_n)\, d\theta$$

then $\hat{\theta}$ is a posterior distribution mean.

Proof for continuous probability distributions.

Calculate the first derivative with respect to $\hat{\theta}$, we get

$$\frac{\partial}{\partial\hat{\theta}} \int (\theta - \hat{\theta})^2 p(\theta|x_1, x_2, ..., x_n)\, d\theta =$$

$$\int \frac{\partial}{\partial\hat{\theta}} \left[(\theta - \hat{\theta})^2\, p(\theta|x_1, x_2, ..., x_n)\right] d\theta = \int -2(\theta - \hat{\theta})\, p(\theta|x_1, x_2, ..., x_n)\, d\theta.$$

Equating to 0, we get

$$\int -2(\theta - \hat{\theta})\, p(\theta|x_1, x_2, ..., x_n)\, d\theta = 0$$

$$\Leftrightarrow \int 2\hat{\theta} p(\theta|x_1, x_2, ..., x_n) d\theta = \int 2\theta p(\theta|x_1, x_2, ..., x_n)\, d\theta \Leftrightarrow \hat{\theta} \underbrace{\int p(\theta|x_1, x_2, ..., x_n)\, d\theta}_{=1} =$$

$$= \int \theta p(\theta|x_1, x_2, ..., x_n)\, d\theta \Leftrightarrow \hat{\theta} = \int \theta p(\theta|x_1, x_2, ..., x_n)\, d\theta$$

Thus, we get MMSE estimate (Minimum Mean Square Error) posterior distribution mean

$$\hat{\theta}_{MMSE} = \int \theta p(\theta|x_1, x_2, ..., x_n)\, d\theta = E(\theta|x_1, x_2, ..., x_n)$$