Lecture Notes 14
36-705

We continue with our discussion of decision theory.

1 Decision Theory

Suppose we want to estimate a parameter θ using data $X^n = (X_1, \ldots, X_n)$. What is the best possible estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of θ? Decision theory provides a framework for answering this question.

1.1 The Risk Function

Let $\widehat{\theta} = \widehat{\theta}(X^n)$ be an estimator for the parameter $\theta \in \Theta$. We start with a **loss function** $L(\theta, \widehat{\theta})$ that measures how good the estimator is. For example:

$$
\begin{array}{ll}
L(\theta,\widehat{\theta}) = (\theta - \widehat{\theta})^2 & \text{squared error loss,} \\
L(\theta,\widehat{\theta}) = |\theta - \widehat{\theta}| & \text{absolute error loss,} \\
L(\theta,\widehat{\theta}) = |\theta - \widehat{\theta}|^p & L_p \text{ loss,} \\
L(\theta,\widehat{\theta}) = 0 \text{ if } \theta = \widehat{\theta} \text{ or } 1 \text{ if } \theta \neq \widehat{\theta} & \text{zero–one loss,} \\
L(\theta,\widehat{\theta}) = I(|\widehat{\theta} - \theta| > c) & \text{large deviation loss,} \\
L(\theta,\widehat{\theta}) = \int \log\left(\frac{p(x;\theta)}{p(x;\widehat{\theta})}\right) p(x;\theta)dx & \text{Kullback–Leibler loss.}
\end{array}
$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector then some common loss functions are

$$L(\theta,\widehat{\theta}) = \|\theta - \widehat{\theta}\|^2 = \sum_{j=1}^{k} (\widehat{\theta}_j - \theta_j)^2,$$

$$L(\theta,\widehat{\theta}) = \|\theta - \widehat{\theta}\|_p = \left(\sum_{j=1}^{k} |\widehat{\theta}_j - \theta_j|^p\right)^{1/p}.$$

When the problem is to predict a $Y \in \{0, 1\}$ based on some classifier $h(x)$ a commonly used loss is

$$L(Y, h(X)) = I(Y \neq h(X)).$$

For real valued prediction a common loss function is

$$L(Y, \widehat{Y}) = (Y - \widehat{Y})^2.$$

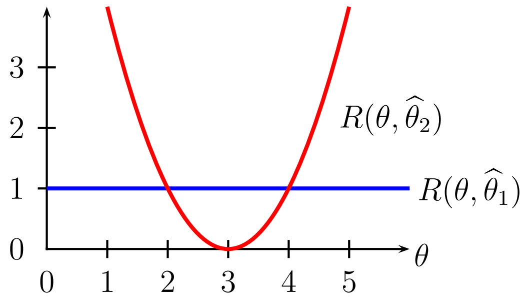

Figure 1: *Comparing two risk functions. Neither risk function dominates the other at all values of θ.*

The **risk** of an estimator $\widehat{\theta}$ is

$$R(\theta,\widehat{\theta}) = \mathbb{E}_\theta\Big(L(\theta,\widehat{\theta})\Big) = \int L(\theta,\widehat{\theta}(x_1,\ldots,x_n))p(x_1,\ldots,x_n;\theta)dx. \tag{1}$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta,\widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta}-\theta)^2 = \mathsf{Var}_\theta(\widehat{\theta}) + \text{bias}^2. \tag{2}$$

If we do not state what loss function we are using, assume the loss function is squared error.

1.2 Comparing Risk Functions

To compare two estimators, we compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

Example 1 *Let $X \sim N(\theta,1)$ and assume we are using squared error loss. Consider two estimators: $\widehat{\theta}_1 = X$ and $\widehat{\theta}_2 = 3$. The risk functions are $R(\theta,\widehat{\theta}_1) = \mathbb{E}_\theta(X-\theta)^2 = 1$ and $R(\theta,\widehat{\theta}_2) = \mathbb{E}_\theta(3-\theta)^2 = (3-\theta)^2$. If $2 < \theta < 4$ then $R(\theta,\widehat{\theta}_2) < R(\theta,\widehat{\theta}_1)$, otherwise, $R(\theta,\widehat{\theta}_1) < R(\theta,\widehat{\theta}_2)$. Neither estimator uniformly dominates the other; see Figure 1.*

Example 2 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Consider squared error loss and let* $\widehat{p}_1 = \overline{X}$. *Since this has zero bias, we have that*

$$R(p, \widehat{p}_1) = \mathsf{Var}(\overline{X}) = \frac{p(1-p)}{n}.$$

Another estimator is

$$\widehat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$$

where $Y = \sum_{i=1}^n X_i$ *and* α *and* β *are positive constants.*[1] *Now,*

$$\begin{aligned} R(p, \widehat{p}_2) &= \mathsf{Var}_p(\widehat{p}_2) + (\text{bias}_p(\widehat{p}_2))^2 \\ &= \mathsf{Var}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2. \end{aligned}$$

Let $\alpha = \beta = \sqrt{n/4}$. *The resulting estimator is*

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and the risk function is

$$R(p, \widehat{p}_2) = \frac{n}{4(n + \sqrt{n})^2}.$$

The risk functions are plotted in Figure 2. As we can see, neither estimator uniformly dominates the other.

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

The **maximum risk** is

$$\overline{R}(\widehat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \tag{3}$$

and the **Bayes risk** under prior π is

$$B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta. \tag{4}$$

Example 3 *Consider again the two estimators in Example 2. We have*

$$\overline{R}(\widehat{p}_1) = \max_{0 \le p \le 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

[1] This is the posterior mean using a Beta (α, β) prior.

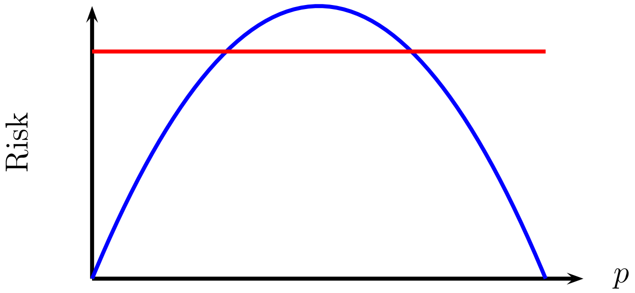

Figure 2: Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 2. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

and

$$\overline{R}(\widehat{p}_2) = \max_p \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when n is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like the maximum risk are imperfect.

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

An estimator $\widehat{\theta}$ that minimizes the Bayes risk is called a **Bayes estimator**. That is,

$$B_\pi(\widehat{\theta}) = \inf_{\tilde{\theta}} B_\pi(\tilde{\theta}) \tag{5}$$

where the infimum is over all estimators $\tilde{\theta}$. An estimator that minimizes the maximum risk is called a **minimax estimator**. That is,

$$\sup_\theta R(\theta, \widehat{\theta}) = \inf_{\tilde{\theta}} \sup_\theta R(\theta, \tilde{\theta}) \tag{6}$$

where the infimum is over all estimators $\tilde{\theta}$. We call the right hand side of (6), namely,

$$R_n \equiv R_n(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}), \tag{7}$$

the **minimax risk**. Statistical decision theory has two main goals: determine the minimax risk R_n and find an estimator that achieves this risk.

Once we have found the minimax risk R_n we want to find the minimax estimator that achieves this risk:

$$\sup_{\theta\in\Theta} R(\theta,\widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R(\theta,\widehat{\theta}). \tag{8}$$

1.3 Bayes Estimators

Let π be a prior distribution. After observing $X^n = (X_1,\ldots,X_n)$, the posterior distribution is, according to Bayes' theorem,

$$\mathbb{P}(\theta\in A|X^n) = \frac{\int_A p(X_1,\ldots,X_n|\theta)\pi(\theta)d\theta}{\int_\Theta p(X_1,\ldots,X_n|\theta)\pi(\theta)d\theta} = \frac{\int_A \mathcal{L}(\theta)\pi(\theta)d\theta}{\int_\Theta \mathcal{L}(\theta)\pi(\theta)d\theta} \tag{9}$$

where $\mathcal{L}(\theta) = p(x^n;\theta)$ is the likelihood function. The posterior has density

$$\pi(\theta|x^n) = \frac{p(x^n|\theta)\pi(\theta)}{m(x^n)} \tag{10}$$

where $m(x^n) = \int p(x^n|\theta)\pi(\theta)d\theta$ is the **marginal distribution** of X^n. Define the **posterior risk** of an estimator $\widehat{\theta}(x^n)$ by

$$r(\widehat{\theta}|x^n) = \int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta. \tag{11}$$

Theorem 4 *The Bayes risk $B_\pi(\widehat{\theta})$ satisfies*

$$B_\pi(\widehat{\theta}) = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n. \tag{12}$$

Let $\widehat{\theta}(x^n)$ be the value of θ that minimizes $r(\widehat{\theta}|x^n)$. Then $\widehat{\theta}$ is the Bayes estimator.

Proof:

Let $p(x,\theta) = p(x|\theta)\pi(\theta)$ denote the joint density of X and θ. We can rewrite the Bayes risk as follows:

$$\begin{aligned}
B_\pi(\widehat{\theta}) &= \int R(\theta,\widehat{\theta})\pi(\theta)d\theta = \int\left(\int L(\theta,\widehat{\theta}(x^n))p(x|\theta)dx^n\right)\pi(\theta)d\theta \\
&= \int\int L(\theta,\widehat{\theta}(x^n))p(x,\theta)dx^n d\theta = \int\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)m(x^n)dx^n d\theta \\
&= \int\left(\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta\right)m(x^n)\,dx^n = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n.
\end{aligned}$$

If we choose $\widehat{\theta}(x^n)$ to be the value of θ that minimizes $r(\widehat{\theta}|x^n)$ then we will minimize the integrand at every x and thus minimize the integral $\int r(\widehat{\theta}|x^n)m(x^n)dx^n$.

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

Theorem 5 *If $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x^n) = \int \theta \pi(\theta|x^n) d\theta = \mathbb{E}(\theta|X = x^n). \tag{13}$$

If $L(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$ then the Bayes estimator is the median of the posterior $\pi(\theta|x^n)$. If $L(\theta, \widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $\pi(\theta|x^n)$.

Proof:

We will prove the theorem for squared error loss. The Bayes estimator $\widehat{\theta}(x^n)$ minimizes $r(\widehat{\theta}|x^n) = \int (\theta - \widehat{\theta}(x^n))^2 \pi(\theta|x^n) d\theta$. Taking the derivative of $r(\widehat{\theta}|x^n)$ with respect to $\widehat{\theta}(x^n)$ and setting it equal to zero yields the equation $2\int (\theta - \widehat{\theta}(x^n))\pi(\theta|x^n)d\theta = 0$. Solving for $\widehat{\theta}(x^n)$ we get 13.

Example 6 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is known. Suppose we use a $N(a, b^2)$ prior for μ. The Bayes estimator with respect to squared error loss is the posterior mean, which is*

$$\widehat{\theta}(X_1, \ldots, X_n) = \frac{b^2}{b^2 + \frac{\sigma^2}{n}} \overline{X} + \frac{\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}} a. \tag{14}$$

It is worth keeping in mind the trade-off: Bayes estimators although easy to compute are very subjective; they depend strongly on the prior π. Minimax estimators, although more challenging to compute are not subjective, but do have the drawback that they are protecting against the worst-case which might lead to pessimistic conclusions.

2 Minimax Estimators through Bayes Estimators

Our goal is to compute a minimax estimator $\widehat{\theta}$ that satisfies:

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \leq \inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}).$$

We will let θ_{minimax} denote a minimax estimator.

2.1 Bounding the Minimax Risk

One strategy to find the minimax estimator is by finding (upper and lower) bounds on the minimax risk that match. Then the estimator that achieves the upper bound is a minimax estimator.

Upper bounding the minimax risk is straightforward. Given an estimator $\widehat{\theta}_{\text{up}}$ we can compute its maximum risk and use it to upper bound the minimax risk, i.e.

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta}_{\text{up}}).$$

The Bayes risk of the Bayes estimator for any prior π lower bounds the minimax risk. Fix a prior π and suppose that $\widehat{\theta}_{\text{low}}$ is the Bayes estimator with respect to π, then we have that:

$$B_\pi(\widehat{\theta}_{\text{low}}) \leq B_\pi(\theta_{\text{minimax}}) \leq \sup_{\theta} R(\theta, \theta_{\text{minimax}}) = \inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}).$$

Let us see an example of this in action.

Example: We will prove a classical result that if we observe independent draws from a d-dimensional Gaussian, $X_1, \ldots, X_n \sim N(\theta, I_d)$, then the average:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

is a minimax estimator of θ with respect to the squared loss.

Let R_n denote the minimax risk. First, let us compute the upper bound on R_n. We note that,

$$\widehat{\theta} \sim N(\theta, I_d/n),$$

so that its risk:

$$R(\theta, \widehat{\theta}) = \mathbb{E}[\sum_{i=1}^{d} (\widehat{\theta}_i - \theta_i)^2] = \mathbb{E}[\sum_{i=1}^{d} Z_i^2],$$

where $Z_i \sim N(0, 1/n)$. This yields that,

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta}) = \frac{d}{n}.$$

Now we lower bound the minimax risk using the Bayes risk. Let us take the prior to be zero-mean Gaussian, i.e. we take $\pi = N(0, c^2 I_d)$. By sufficiency, we can replace the data with $\widehat{\theta}$. We can write:

$$\theta \sim N(0, c^2 I_d)$$
$$\widehat{\theta} | \theta \sim N(\theta, I_d/n).$$

We can write this as

$$\theta = c\epsilon$$
$$\widehat{\theta} = \frac{1}{\sqrt{n}}\, Z$$

where $\epsilon, Z \sim N(0, I_d)$. Hence,

$$\begin{pmatrix} \theta \\ \widehat{\theta} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} c^2 I_d & c^2 I_d \\ c^2 I_d & (c^2 + 1/n) I_d \end{bmatrix} \right]$$

We can now compute the posterior (using standard conditional Gaussian formulae), and obtain its mean:

$$\mathbb{E}[\theta|\widehat{\theta}] = \frac{c^2}{c^2 + 1/n}\widehat{\theta}.$$

Now,

$$R(\theta, \widehat{\theta}) = \mathbb{E}\left|\left|\frac{c^2}{c^2 + 1/n}\widehat{\theta} - \theta\right|\right|^2.$$

Write $\widehat{\theta} = \theta + W$, where $W \sim N(0, I_d/n)$. Then

$$R(\theta, \widehat{\theta}) = \mathbb{E}_W\left|\left|\frac{c^2}{c^2 + 1/n}Z - \frac{\theta}{n(c^2 + 1/n)}\right|\right|^2.$$

Let us denote $\beta := c^2 + 1/n$. Then we obtain that,

$$R(\theta, \widehat{\theta}) = \frac{\|\theta\|_2^2}{n^2\beta^2} + \frac{c^4}{\beta^2}\mathbb{E}\|W\|_2^2 = \frac{\|\theta\|_2^2}{n^2\beta^2} + \frac{c^4}{\beta^2}\frac{d}{n}.$$

The Bayes risk further averages this over $\theta \sim N(0, c^2 I_d)$ to obtain that,

$$B_\pi\left(\frac{c^2}{c^2 + 1/n}\widehat{\theta}\right) = \frac{c^2 d}{n^2\beta^2} + \frac{c^4}{\beta^2}\frac{d}{n} = \frac{c^2 d}{n\beta} = \frac{d}{n(1 + 1/(nc^2))}.$$

We conclude that

$$\frac{d}{n(1 + 1/(nc^2))} \le R_n \le \frac{d}{n}.$$

This is true for every $c > 0$. Since c was arbitrary we can take the limit as $c \to \infty$ to obtain that the minimax risk is upper and lower bounded by d/n and hence, $R_n = d/n$ and the sample average $\widehat{\theta}$ is minimax.

2.2 Least Favorable Prior

The other way to obtain Bayes estimators is by constructing what are called least favorable priors.

Theorem 7 *Let $\widehat{\theta}$ be the Bayes estimator for some prior π. If*

$$R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \ \text{ for all } \theta \tag{15}$$

then $\widehat{\theta}$ is minimax and π is called a **least favorable prior***.*

Proof:

Suppose that $\widehat{\theta}$ is not minimax. Then there is another estimator $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta})$. Since the average of a function is always less than or equal to its maximum, we have that $B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \tag{16}$$

which is a contradiction.

Theorem 8 *Suppose that $\widehat{\theta}$ is the Bayes estimator with respect to some prior π. If the risk is constant then $\widehat{\theta}$ is minimax.*

Proof:

The Bayes risk is $B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all θ. Now apply the previous theorem.

Example 9 *Consider the Bernoulli model with squared error loss. We showed previously that the estimator*

$$\widehat{p} = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes estimator, for the prior Beta(α, β) *with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax.*