Lecture 21: Loss function framework for point estimation

24 March 2016

So far we have discussed Mean Square Error performance of estimators. In this lecture, we shall see the loss function framework for the evaluation of estimators.

1 Ingredients of a general loss function framework

- Parameter space: Θ (e.g. $\mathbb{R}$)
- Observation space: $\mathscr{X}$
- Family of distributions indexed by Θ: $\{f(x|\theta),\ \theta \in \Theta\}$
- Action/Decision/Output space: $\mathscr{A}$
  (typically $\mathscr{A} \supseteq \Theta$, because estimator can give output $\notin \Theta$)
- Loss function:

  $$L : \Theta \times \mathscr{A} \ \rightarrow\ \mathbb{R}_+$$

  $L(\theta, a)$: "cost" suffered when estimating θ to be equal to a. (Ideally, if $\mathscr{A} = \Theta$; then L(θ,a)=0 when a=θ).

  Below are some examples of loss functions for $\Theta = \mathscr{A} = \mathbb{R}$

  1. Absolute loss:

     $$L(\theta, a) = |\theta - a|$$

  2. Square loss (corresponds to MSE)

     $$L(\theta, a) = (a - \theta)^2$$

3. Zero-One loss:

$$L(\theta, a) = \mathbb{1}_{\{\theta \neq a\}}$$

4. p-norm loss:

$$L(\theta, a) = |\theta - a|^p$$

Given an estimator $W(X)$, $(W : X \to \mathscr{A})$ of $\theta \in \Theta$, $\{X \sim f(x|\theta)\}$, its *risk function* at $\theta \in \Theta$ is given as,

$$\begin{aligned} R(\theta, W) &= \mathbb{E}_\theta[L(\theta, W(X))], \\ &= \int_{\mathscr{X}} L(\theta, W(X))\ f(x|\theta)\ dx. \end{aligned} \tag{1}$$

(If L is square loss, then the above risk R gives the mean square error). Our goal is to design W to minimize $R(\theta, W)$ over "all or most $\theta \in \Theta$".

Now given two estimators over the parameter space Θ, how do we compare their performance and choose the best?. Consider the figure shown below (fig.1). The x-axis represents the parameter space $\theta \in \Theta$ and y-axis represents the risk $R(\theta, W)$, for an estimator W w.r.t θ.

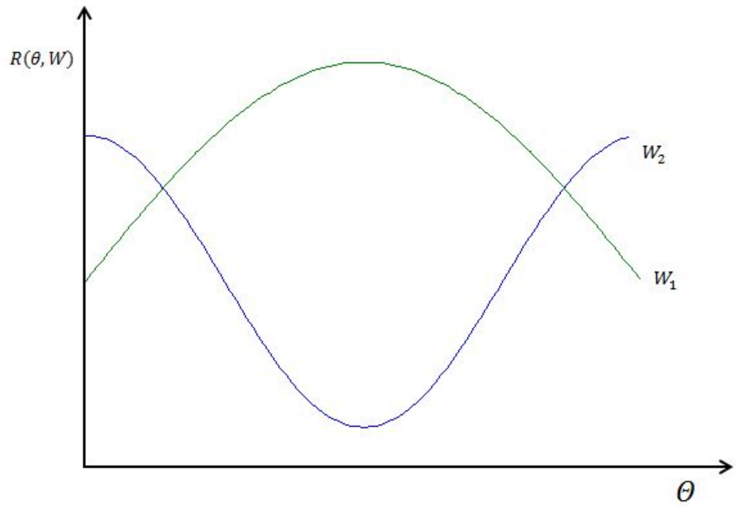

Figure 1: Risk v/s Θ for different estimators

One way to decide on the best estimator W^* would be to choose the one having smaller peak. One can see that this is equivalent to the minimax estimator (as we are choosing the W with minimum $\max_\theta R(\theta, W)$). Another option is to choose W that minimizes the area under the $R(\theta, W)$ function. This is equivalent to the Bayesian estimator.

2 Notions of Optimality (Rule to compare estimators)

1. Bayes Risk: Assume the a-priori probability distribution π over the parameter space Θ is given. The Bayes risk of $W = W(X)$ is,

$$B_\pi(W) = \int_\Theta R(\theta, W)\ \pi(\theta)\ d\theta. \tag{2}$$

   Any estimator W that minimizes $B_\pi(.)$ over all estimators is called a Bayes estimator (denoted by W_π^*)

2. Max Risk (No prior necessary):

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

   Estimator minimizing $\overline{R}(.)$ is a minimax estimator.

2.1 Bayes Estimators

Bayes risk under prior π:

$$\begin{aligned} B_\pi(W) &= \int_\Theta R(\theta, W)\ \pi(\theta)\ d\theta, & (4)\\ &= \int_\Theta \int_{\mathscr{X}} L(\theta, W(X))\ f(x|\theta) dx\ \pi(\theta) d\theta, & (5)\\ &= \int_{\mathscr{X}} \left[\int_\Theta L(\theta, W(X))\ \pi(\theta|x)\ d\theta \right] m(x)\ dx, & (6) \end{aligned}$$

where, we have used $f(x|\theta)\ \pi(\theta) = \pi(\theta|x)\ m(x)$. We have defined $m(x)$ as marginal of x,

$$m(x) = \int_\Theta \pi(\theta').f(x|\theta')d\theta',$$

and Posterior density of θ given x

$$\pi(\theta|x) = \frac{\pi(\theta).f(x|\theta)}{m(x)}. \tag{7}$$

Note that the quantity inside [.] in eqn. (6), is a function of only x (and not θ). This implies that, to minimize $B_\pi(W)$, we should choose,

$$\forall x \in \mathscr{X} : W(X) \in \arg\min_{a \in \mathscr{A}} \int_\Theta L(\theta, a)\pi(\theta|x)d\theta \tag{8}$$

i.e., a Bayes estimator minimizes the posterior expected loss given the data x.

Example 2.1 (Bayes estimator for square-loss function). Let $\Theta = \mathscr{A} = \mathbb{R}$, and $L(\theta, a) = (a-\theta)^2$. The posterior expected loss is,

$$\int_{\mathbb{R}} (a-\theta)^2 \pi(\theta|x) dx. \tag{9}$$

Then the Bayes estimator is $W(X) = \int_\Theta \theta \; \pi(\theta|x) \; dx$ i.e., the posterior mean.

Example 2.2 (Bayes estimator for absolute loss function). Let $\Theta = \mathscr{A} = \mathbb{R}$, and $L(\theta, a) = |a-\theta|$. The posterior expected loss is

$$\int_{\mathbb{R}} |a-\theta| \pi(\theta|x) dx. \tag{10}$$

Here the Bayes estimator returns $W(X) = median(\pi(.|x))$.

Proof. The posterior expected loss is given by

$$\begin{aligned} \mathbb{E}|x-a| &= \int_{\mathbb{R}} |x-a| \pi(\theta|x) \; dx, \\ &= \int_{-\infty}^{a} -(x-a)\pi(\theta|x)dx + \int_{a}^{\infty} (x-a)\pi(\theta|x) \; dx. \end{aligned} \tag{11}$$

The Bayes estimator is given by

$$W(x) = \arg\min_a \mathbb{E}|x-a|. \tag{12}$$

Minimum can be obtained by computing the derivative and equating to 0.

$$\frac{d}{da}\mathbb{E}|x-a| = \int_{-\infty}^{a} \pi(\theta|x)dx - \int_{a}^{\infty} \pi(\theta|x)dx \tag{13}$$

Equating this equation to zero gives the result as $a = median(\pi(.|x))$ □

(Similarly a $0-1$ loss function returns $W(X) = mode(\pi(.|x))$)

2.2 Minimax Estimator

It turns out that minimax estimation is complicated. The main take-away here is that the Bayes estimator with constant risk over Θ is minimax.

Definition 2.3. A prior π over Θ is a *least favorable prior*, if it has the highest Bayes risk, i.e., $B_\pi(W^*_\pi) \geq B_{\pi'}(W^*_{\pi'}), \; \forall \; \pi'$ on Θ .

Theorem 2.4. *Suppose W is the Bayes estimator for some prior π over Θ, if $L(\theta, W)$ is a constant $\forall \; \theta \in \Theta$, then,*

1. *π is a least favorable prior*
2. *W is a minimax estimator.*

3 Asymptotic Evaluation of Estimators

The goal here is to study what happens to the quality of estimation as the number of samples tend to infinity.

Definition 3.1. Let $W_n \equiv W_n(X_1, ..., X_n)$ for $n \geq 1$, be a sequence of estimators, for θ, and assuming $X_i \overset{iid}{\sim} f(x|\theta)$, then W_n is *consistent* for estimating θ, if $\forall \theta \in \Theta,\ W_n \overset{P_\theta}{\to} \theta$, i.e., $\forall\ \theta \in \Theta,\ \epsilon > 0,\ lim_{n\to\infty}\ P[|\ W_n - \theta| \geq \epsilon] = 0$.

Note 1. Consistency is equivalent to convergence to quantity being estimated.

Note 2. Need convergence in probability $\forall\ \theta \in \Theta$.

Since mean-square convergence implies convergence in probability, $\forall\ \theta \in \Theta,\ E_\theta[(W_n - \theta)^2]\ \to \infty$ as $n \to \infty$ is enough to show that W_n is consistent.

Theorem 3.2. *If* $W_n \equiv W_n(X_1, ..., X_n)$ *is a sequence of estimators, such that* $\forall\ \theta$*,*

1. $lim_{n\to\infty}\ var_\theta[W_n] = 0$*,*
2. $lim_{n\to\infty}\ \mathbb{E}_\theta[W_n] - \theta = 0$*,*

then W_n *is consistent.*

Example 3.3 (Consistency of sample mean). Let $X_1, ..., X_n \overset{iid}{\sim} f(x|\theta)$, for $\theta \in \Theta \subseteq \mathbb{R}$, and $\forall\ \theta \in \Theta,\ \mathbb{E}_\theta[|X_1|] < \infty$, let $W_n = \frac{1}{n}\sum_{i=1}^n X_i;\ \forall\ n \geq 1$. $\{W_n\}$ is consistent for estimating $\mathbb{E}_\theta[X]$ since, $\frac{1}{n}\sum_{i=1}^n X_i \overset{P_\theta}{\to} \mathbb{E}_\theta[X_1] = g(\theta)$, due to the weak law of large numbers.

3.1 Consistency of Maximum Likelihood Estimator

Recall that $X_1, ..., X_n \overset{iid}{\sim} f(x|\theta)$, for $\theta \in \Theta \subseteq \mathbb{R}$; the MLE of θ is $\arg\max_{\theta\in\Theta} \prod_{i=1}^n f(x_i|\theta)$ or we can say,

$$W_{MLE} \in \arg\max_{\theta\in\Theta} \sum_{i=1}^n log(f(x_i|\theta)). \tag{14}$$

Theorem 3.4 (Consistency of MLE). *Suppose* $X_1, ..., X_n \overset{iid}{\sim} f(x|\theta)$*, for* $\theta \in \Theta \subseteq \mathbb{R}$*, and* $f(x|\theta \in \Theta)$ *satisfies some regularity conditions, then* $\forall \theta \in \Theta,\ W_{MLE}^{(n)} \overset{P_\theta}{\to} \theta$*.*