BAYESIAN ROBUSTNESS OF THE QUANTILE LOSS IN STATISTICAL DECISION THEORY

Julián de la Horra

Departamento de Matemáticas
Universidad Autónoma de Madrid
E-mail: julian.delahorra@uam.es

Introduction and notation

Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample taken from a sample space $\mathcal{X}$ with density function $f(\mathbf{x}|\theta) = \Pi_{i=1}^{n} f(x_i|\theta)$, where θ is an unknown parameter taking values in a parameter space Θ.

Let us denote by a any action we can take from an action space $\mathcal{A}$.

Let us denote by $L : \mathcal{A} \times \Theta \rightarrow \Re$, the loss function measuring (in some suitable way) the loss we incur when the action a is taken and the true value of the parameter is θ.

Of course, we are interested in choosing an action depending on the sample $\mathbf{x} = (x_1, ..., x_n)$ we have observed. So, let us denote by $\delta : \mathcal{X} \rightarrow \mathcal{A}$, any decision rule we can use for choosing an action depending on the different samples we can obtain.

Introduction and notation

In any decision problem, we have to evaluate (in some suitable way) the different decision rules in order to choose one of them and, for doing that, we have to choose a criterion for comparing random variables, because the loss of a decision rule δ is a random variable.

1. Frequentist context:
   For each θ fixed, the loss is the random variable $L : \mathcal{X} \to \Re$ defined as $L(\delta(\mathbf{X}), \theta)$, where the probability model over $\mathcal{X}$ is given by the sampling density $f(\mathbf{x}|\theta)$.
2. Bayesian context:
   Let $\pi(\theta)$ be the prior density summarizing the prior information we have on the parameter θ. Now, the loss is the random variable $L : \mathcal{X} \times \Theta \to \Re$ defined as $L(\delta(\mathbf{X}), \theta)$, where the probability model over $\mathcal{X} \times \Theta$ is given by the joint density $f(\mathbf{x}|\theta)\pi(\theta)$ (obtained from the sampling density and the prior density).

Introduction and notation

Von Neumann and Morgenstern (1947) (frequentist context) and Savage (1954) (Bayesian context) provided axiomatic justifications for comparing these random variables by means of their expectations. In other words, they gave axiomatic justifications for minimizing expected loss. Moreover, expectations are very friendly to handle and, therefore, minimizing expected loss became the standard criterion.

But, of course, there are other possibilities ...

Three medloss criteria

Manski (1988) suggested the optimization of quantiles (instead of expectations) for comparing random variables. The use of quantiles is very popular in finance and has different benefits: it is not necessary to assume the existence of moments (and so, heavy tailed distributions are not a problem), and the behavior of quantiles is usually more robust. Rostek (2007) provided an axiomatic justification for the optimization of quantiles, making use of the previous work by Machina and Schmeidler (1992). Before that, De la Horra (1981) provided an axiomatic justification for the case in which the parameter space, Θ, is finite, exploiting the previous work by Rios (1967).

Yu and Clarke (2011, J. S. P. I.) proposed three different criteria based on comparing medians of random variables (of course, the extension from medians to quantiles is immediate).

Frequentist medloss criterion

(a) For each θ fixed, let us consider the random variable $L : \mathcal{X} \to \Re$ defined as $L(\delta(\mathbf{X}), \theta)$, where the probability model over $\mathcal{X}$ is given by the sampling density $f(\mathbf{x}|\theta)$.

(b) For each θ fixed, let us consider the median of the previous random variable:

$$M_{f(\mathbf{x}|\theta)}(L(\delta(\mathbf{X}), \theta)).$$

(c) We say that the decision rule δ_1 is better than the decison rule δ_2 (in the sense of the frequentist medloss criterion) when:

$$M_{f(\mathbf{x}|\theta)}(L(\delta_1(\mathbf{X}), \theta)) \leq M_{f(\mathbf{x}|\theta)}(L(\delta_2(\mathbf{X}), \theta)), \qquad \forall \theta \in \Theta.$$

Bayes medloss criterion

(a) For each θ fixed, let us consider the random variable $L : \mathcal{X} \to \Re$ defined as $L(\delta(\mathbf{X}), \theta)$, where the probability model over $\mathcal{X}$ is given by the sampling density $f(\mathbf{x}|\theta)$.

(b) For each θ fixed, let us consider the median of the previous random variable:

$$M_{f(\mathbf{x}|\theta)}(L(\delta(\mathbf{X}), \theta)).$$

(c) Let $\pi(\theta)$ be the prior density summarizing the prior information we have on the parameter θ, and let us consider the random variable $h : \Theta \to \Re$ defined as $h(\theta) = M_{f(\mathbf{x}|\theta)}(L(\delta(\mathbf{X}), \theta))$, where the probability model over Θ is given by the prior density $\pi(\theta)$.

Bayes medloss criterion

(d) Let us consider the median of the previous random variable:

$$M_{\pi(\theta)}(h(\theta)) = M_{\pi(\theta)}\left[M_{f(\mathbf{x}|\theta)}(L(\delta(\mathbf{X}), \theta))\right].$$

(e) We say that the decision rule δ_1 is better than the decison rule δ_2 (in the sense of the Bayes medloss criterion) when:

$$M_{\pi(\theta)}\left[M_{f(\mathbf{x}|\theta)}(L(\delta_1(\mathbf{X}), \theta))\right] \leq M_{\pi(\theta)}\left[M_{f(\mathbf{x}|\theta)}(L(\delta_2(\mathbf{X}), \theta))\right].$$

Posterior medloss criterion

(a) Let $\pi(\theta)$ be the prior density summarizing the prior information we have on the parameter θ. For the sample $\mathbf{x} = (x_1, ..., x_n)$ we have actually observed, let us compute the posterior density in the usual way:

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f(\mathbf{x}|\theta)\pi(\theta)d\theta} \ .$$

(b) For $\mathbf{x} = (x_1, ..., x_n)$ fixed (the sample we have actually observed), let us consider the random variable $L : \Theta \to \Re$ defined as $L(\delta(\mathbf{x}), \theta)$, where the probability model over Θ is given by the posterior density $\pi(\theta|\mathbf{x})$.

Posterior medloss criterion

(c) For $\mathbf{x} = (x_1, ..., x_n)$ fixed (the sample we have actually observed), let us consider the median of the previous random variable:

$$M_{\pi(\theta|\mathbf{x})}(L(\delta(\mathbf{x}), \theta)).$$

(d) We say that the decision rule δ_1 is better than the decison rule δ_2 (in the sense of the posterior medloss criterion) when:

$$M_{\pi(\theta|\mathbf{x})}(L(\delta_1(\mathbf{x}), \theta)) \leq M_{\pi(\theta|\mathbf{x})}(L(\delta_2(\mathbf{x}), \theta)).$$

Bayes quantile loss

(a) Let $\pi(\theta)$ be the prior density summarizing the prior information we have on the parameter θ. Now, let us consider the random variable $L : \mathcal{X} \times \Theta \to \Re$ defined as $L(\delta(\mathbf{X}), \theta)$, where the probability model over $\mathcal{X} \times \Theta$ is given by the joint density $f(\mathbf{x}|\theta)\pi(\theta)$ (obtained from the sampling density and the prior density).

(b) Now, let us consider the median of the previous random variable:

$$M_{f(\mathbf{x}|\theta)\pi(\theta)}(L(\delta(\mathbf{X}), \theta)).$$

(c) We say that the decision rule δ_1 is better than the decision rule δ_2 (in the sense of this new Bayes medloss criterion) when:

$$M_{f(\mathbf{x}|\theta)\pi(\theta)}(L(\delta_1(\mathbf{X}), \theta)) \leq M_{f(\mathbf{x}|\theta)\pi(\theta)}(L(\delta(\mathbf{X}), \theta)).$$

Bayes quantile loss

Comments

1. This criterion is close to Bayes medloss criterion but it is not the same. The technical reason for this difference is that, although for expectations we have

$$E_{\pi(\theta)}[E_{f(\mathbf{x}|\theta)}(L(\delta(\mathbf{X}),\theta))] = E_{f(\mathbf{x}|\theta)\pi(\theta)}(L(\delta(\mathbf{X}),\theta)),$$

   for medians we have (in general)

$$M_{\pi(\theta)}[M_{f(\mathbf{x}|\theta)}(L(\delta(\mathbf{X}),\theta))] \neq M_{f(\mathbf{x}|\theta)\pi(\theta)}(L(\delta(\mathbf{X}),\theta)).$$

2. The median is nothing but a specific quantile, and there is no reason for constrain our attention to the median (in fact, quantiles have been very used in finance). Therefore, the new criterion is next defined through quantiles. All the elements in the decision problem are recalled in the following definition.

Bayes quantile loss

Definition (Bayes ε-quantloss criterion)
The Bayes ε-quantloss of a decision rule δ is the quantile (at level ε) of the random variable $L : \mathcal{X} \times \Theta \longrightarrow \Re$ defined as $L(\delta(\mathbf{X}), \theta)$, where the probability model over $\mathcal{X} \times \Theta$ is given by the joint density $f(\mathbf{x}|\theta)\pi(\theta)$ (obtained from the sampling density and the prior density):

$$\begin{aligned} Q^{\varepsilon}(\delta, \pi) &= \max\{t \in \Re : Pr_{f(\mathbf{x}|\theta)\pi(\theta)}(L(\delta(\mathbf{X}), \theta) < t) \leq \varepsilon\} \\ &= \max\{t : Pr(L(\delta, \pi) < t) \leq \varepsilon\}. \end{aligned}$$

We say that the decision rule δ_1 is better than the decison rule δ_2 (in the sense of Bayes ε-quantloss criterion) when:

$$Q^{\varepsilon}(\delta_1, \pi) \leq Q^{\varepsilon}(\delta_2, \pi).$$

Bayesian robustness

In general, it is complicated to elicit the exact prior distribution. The usual solution to this problem is to consider a class of priors (instead of a single prior) and to evaluate the differences we find when the prior ranges over this class. This is the usual analysis of Bayesian robustness.

Therefore, if we need to carry out a study of Bayesian robustness for the Bayes ε-quantloss, we first choose a suitable class of prior distributions, Γ, and we then compute:

$$\inf_{\pi\in\Gamma} Q^{\varepsilon}(\delta,\pi) = \inf_{\pi\in\Gamma} [\max\{t : Pr(L(\delta,\pi) < t) \leq \varepsilon\}], \tag{1}$$

$$\sup_{\pi\in\Gamma} Q^{\varepsilon}(\delta,\pi) = \sup_{\pi\in\Gamma} [\max\{t : Pr(L(\delta,\pi) < t) \leq \varepsilon\}]. \tag{2}$$

When the difference between these two values is small, we say that the Bayes ε-quantloss shows a robust behavior.

Bayesian robustness

Theorem

For any class Γ*, we have:*

$$\inf_{\pi\in\Gamma} Q^{\varepsilon}(\delta,\pi) = \max\left\{t : \sup_{\pi\in\Gamma} Pr(L(\delta,\pi) < t) \leq \varepsilon\right\},$$

$$\sup_{\pi\in\Gamma} Q^{\varepsilon}(\delta,\pi) = \max\left\{t : \inf_{\pi\in\Gamma} Pr(L(\delta,\pi) < t) \leq \varepsilon\right\}.$$

Corollary

For the case in which all the degenerate distributions are included in the class Γ*, we have:*

$$\inf_{\pi\in\Gamma} Q^{\varepsilon}(\delta,\pi) = \max\left\{ t : \sup_{\theta\in\Theta} Pr(L(\delta,\theta) < t) \leq \varepsilon \right\},$$

$$\sup_{\pi\in\Gamma} Q^{\varepsilon}(\delta,\pi) = \max\left\{ t : \inf_{\theta\in\Theta} Pr(L(\delta,\theta) < t) \leq \varepsilon \right\}.$$

These theoretical results may be useful for developing algorithms.