EE290 Mathematics of Data Science Lecture 1 - 08/29/2019

Lecture 1: Statistical Decision Theory

Lecturer: Jiantao Jiao *Scribe: Jiantao Jiao*

In this lecture, we discuss a unified theoretical framework of statistics proposed by Abraham Wald, which is named statistical decision theory. [1]. It was adapted from the notes of lecture 2 of EE378A at Stanford University taught by Jiantao Jiao and scribed by Andrew Hilger.

1 Goals

1. **Evaluation:** The theoretical framework should aid fair comparisons between algorithms (e.g., maximum entropy vs. maximum likelihood vs. method of moments).
2. **Achievability:** The theoretical framework should be able to inspire the constructions of statistical algorithms that are (nearly) optimal under the optimality criteria introduced in the framework.

2 Basic Elements of Statistical Decision Theory

1. **Statistical Experiment:** A family of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where θ is a parameter and P_θ is a probability distribution indexed by the parameter.
2. **Data:** $X \sim P_\theta$, where X is a random variable observed for some parameter value θ.
3. **Objective:** $g(\theta)$, e.g., inference on the entropy of distribution P_θ.
4. **Decision Rule:** $\delta(X)$. The decision rule need not be deterministic. In other words, there could be a probabilistically defined decision rule with an associated $P_{\delta|X}$.
5. **Loss Function:** $L(\theta, \delta)$. The loss function tells us how bad we feel about our decision once we find out the true value of the parameter θ chosen by nature.

Example: $P_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta)^2}{2}}$, $g(\theta) = \theta$, and $L(\theta, \delta) = (\theta - \delta)^2$. In other words, X is normally distributed with mean θ and unit variance $X \sim N(\theta, 1)$, and we are trying to estimate the mean θ. We judge our success (or failure) using mean-square error.

3 Risk Function

Definition 1 (Risk Function)**.**

$$R(\theta, \delta) \triangleq \mathbb{E}[L(\theta, \delta(X))] \tag{1}$$
$$= \int L(\theta, \delta(x)) P_\theta(dx) \tag{2}$$
$$= \iint L(\theta, \delta) P_{\delta|X}(d\delta|x) P_\theta(dx). \tag{3}$$

[1]See Wald, Abraham. "Statistical decision functions." In Breakthroughs in Statistics, pp. 342-357. Springer New York, 1992.

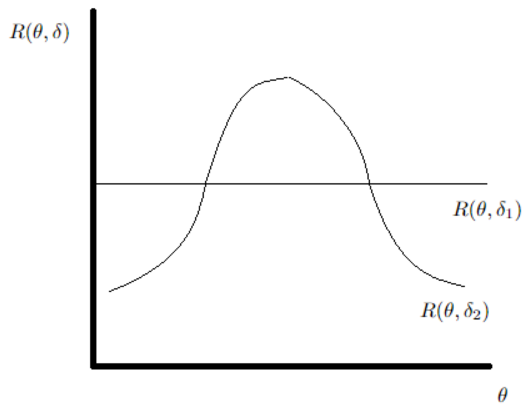

Figure 1: Example risk functions computed over a range of parameter values for two different decision rules.

A risk function evaluates a decision rule's success over a large number of experiments with fixed parameter value θ. By the law of large numbers, if we observe X many times independently, the average empirical loss of the decision rule δ will converge to the risk $R(\theta, \delta)$.

Even after determining risk functions of two decision rules, it may still be unclear which is better. Consider the example of Figure 1. Two different decision rules δ_1 and δ_2 result in two different risk functions $R(\theta, \delta_1)$ and $R(\theta, \delta_2)$ evaluated over different values on the parameter θ. The first decision rule δ_1 is inferior for low and high values of the parameter θ but is superior for the middle values. Thus, even after computing a risk function R, it can still be unclear which decision rule is better. We need new ideas to enable us to compare different decision rules.

4 Optimality Criterion of Decision Rules

Given the risk function of various decision rules as a function of the parameter θ, there are various approaches to determining which decision rule is optimal.

4.1 Restrict the Competitors

This is a traditional set of methods that were overshadowed by other approaches that we introduce later. A decision rule δ' is eliminated (or formally, is *inadmissible*) if there are any other decision rules δ that are strictly better, i.e., $R(\theta, \delta') \geq R(\theta, \delta)$ for any $\theta \in \Theta$ and the inequality becomes strict for at least one $\theta_0 \in \Theta$. However, the problem is that many decision rules cannot be eliminated in this way and we still lack a criterion to determine which one is better. Then to aid in selection, the rationale of the approach of restricting competitors is that only decision rules that are members of a certain decision rule class $\mathcal{D}$ are considered. The advantage is that, sometimes all but one decision rule in $\mathcal{D}$ is inadmissible, and we just use the only admissible one.

1. **Example 1:** Class of unbiased decision rules $\mathcal{D}' = \{\delta : \mathbb{E}[\delta(X)] = g(\theta), \forall \theta \in \Theta\}$
2. **Example 2:** Class of invariant estimators.

However, a serious drawback of this approach is that $\mathcal{D}$ may be an empty set for various decision theoretic problems.

4.2 Bayesian: Average Risk Optimality

The idea is to use averaging to reduce the risk function $R(\theta, \delta)$ to a single number for any given δ.

Definition 2 (Average risk under prior $\Lambda(d\theta)$)**.**

$$r(\Lambda, \delta) = \int R(\theta, \delta)\Lambda(d\theta) \tag{4}$$

Here Λ is the is the prior distribution, a probability measure on Θ. The Bayesians and the frequentists disagree about Λ; namely, the frequentists do not believe the existence of the prior. However, there do exist more justifications of the Bayesian approach than the interpretation of Λ as prior belief: indeed, the *complete class theorem* in statistical decision theory asserts that in various decision theoretic problems, all the admissible decision rules can be approximated by Bayes estimators. [2]

Definition 3 (Bayes estimator)**.**

$$\delta_\Lambda = \arg\min_{\delta} r(\Lambda, \delta) \tag{5}$$

The Bayes estimator δ_Λ can usually be found using the principle of computing posterior distributions. Note that

$$\begin{aligned} r(\Lambda, \delta) &= \iiint L(\theta, \delta) P_{\delta|X}(d\delta|x) P_\theta(dx)\Lambda(d\theta) && (6) \\ &= \int \Big(\iint L(\theta, \delta) P_{\delta|X}(d\delta|x) P_{\theta|X}(d\theta|x) \Big) P_X(dx) && (7) \end{aligned}$$

where $P_X(dx)$ is the *marginal* distribution of X and $P_{\theta|X}(d\theta|x)$ is the *posterior* distribution of θ given X. In Equation 6, $P_\theta(dx)\Lambda(d\theta)$ is the joint distribution of θ and X. In Equation 7, we only have to minimize the portion in parentheses to minimize $r(\Lambda, \delta)$ because $P_X(dx)$ doesn't depend on δ.

Theorem 4. [3] *Under mild conditions,*

$$\begin{aligned} \delta_\Lambda(x) &= \arg\min_{\delta} \mathbb{E}[L(\theta, \delta)|X = x] && (8) \\ &= \arg\min_{P_{\delta|X}} \iint L(\theta, \delta) P_{\delta|X}(d\delta|x) P_{\theta|X}(d\theta|x) && (9) \end{aligned}$$

Lemma 5. *If $L(\theta, \delta)$ is convex in δ, it suffices to consider deterministic rules $\delta(x)$.*

Proof Jensen's inequality:

$$\iint L(\theta, \delta) P_{\delta|X}(d\delta|x) P_{\theta|X}(d\theta|x) \ge \int L(\theta, \int \delta P_{\delta|X}(d\delta|x)) P_{\theta|X}(d\theta|x). \tag{10}$$

□

Examples

1. $L(\theta, \delta) = (g(\theta) - \delta)^2 \Rightarrow \delta_\Lambda(x) = \mathbb{E}[g(\theta)|X = x]$. In other words, the Bayes estimator under squared error loss is the conditional expectation of $g(\theta)$ given x.
2. $L(\theta, \delta) = |g(\theta) - \delta| \Rightarrow \delta_\Lambda(x)$ is any median of the posterior distribution $P_{g(\theta)|X=x}$.
3. $L(\theta, \delta) = \mathbb{1}(g(\theta) \neq \delta) \Rightarrow \delta_\Lambda(x) = \arg\max_{g(\theta)} P_{g(\theta)|x}(g(\theta)|X = x)$. In other words, an indicator loss function results in a maximum a posteriori (MAP) estimator decision rule[4].

[2] See Chapter 3 of Friedrich Liese, and Klaus-J. Miescke. "Statistical Decision Theory: Estimation, Testing, and Selection." (2009)

[3] See Theorem 1.1, Chapter 4 of Lehmann EL, Casella G. Theory of point estimation. Springer Science & Business Media; 1998

[4] Next week, we will cover special cases of P_θ and how to solve Bayes estimator in a computationally efficient way. In the general case, however, computing the posterior distribution may be difficult.

4.3 Frequentist: Worst-Case Optimality (Minimax)

Definition 6 (Minimax estimator)**.** *The decision rule δ^* is minimax among all decision rules in $\mathcal{D}$ iff*

$$\sup_{\theta\in\Theta} R(\theta,\delta^*) = \inf_{\delta\in\mathcal{D}} \sup_{\theta\in\Theta} R(\theta,\delta). \tag{11}$$

4.3.1 First observation

Since

$$R(\theta,\delta) = \iint L(\theta,\delta) P_{\delta|X}(d\delta|x) P_\theta(dx) \tag{12}$$

is linear in $P_{\theta|X}$, this is a convex function in δ. The supremum of a convex function is convex, so finding the optimal decision rule is a convex optimization problem. However, solving this convex optimization problem may be computationally intractable. For example, it may not even be computationally tractable to compute the supremum of the risk function $R(\theta,\delta)$ over $\theta \in \Theta$. Hence, finding the exact minimax estimator is usually *hard*.

4.3.2 Second observation

Due to the previous difficulty of finding the exact minimax estimator, we turn to another goal: we wish to find an estimator δ' such that

$$\inf_\delta \sup_\theta R(\theta,\delta) \le \sup_\theta R(\theta,\delta') \le c \cdot \inf_\delta \sup_\theta R(\theta,\delta) \tag{13}$$

where $c > 1$ is a constant. The left inequality is trivially true. For the right inequality, in practice one can usually choose some specific δ' and evaluate an upper bound of $\sup_\theta R(\theta,\delta')$ explicitly. However, it remains to find a lower bound of $\inf_\delta \sup_\theta R(\theta,\delta)$. To solve the problem (and save the world), we can use the minimax theorem.

Theorem 7 (Minimax Theorem (Sion-Kakutani))**.** *Let Λ, X be two compact, convex sets in some topologically vector spaces. Let function $H(\lambda,x) : \Lambda \times X \to \mathbb{R}$ be a continuous function such that:*

1. *$H(\lambda,\cdot)$ is convex for any fixed $\lambda \in \Lambda$*
2. *$H(\cdot,x)$ is concave for any fixed $x \in X$.*

Then

1. *Strong duality:* $\max_\lambda \min_x H(\lambda,x) = \min_x \max_\lambda H(\lambda,x)$
2. *Existence of Saddle point:*

$$\exists(\lambda^*,x^*): \qquad H(\lambda,x^*) \le H(\lambda^*,x^*) \le H(\lambda^*,x) \quad \forall\lambda\in\Lambda, x\in X.$$

The existence of saddle point implies the strong duality.

We note that other than the strong duality, the following weak duality is always true without assumptions on H:

$$\sup_\lambda \inf_x H(\lambda,x) \le \inf_x \sup_\lambda H(\lambda,x) \tag{14}$$

We define the quantity $r_\Lambda \triangleq \inf_\delta r(\Lambda,\delta)$ as the *Bayes risk* under prior distribution Λ. We have the following lines of arguments using weak duality:

$$\inf_\delta \sup_\theta R(\theta,\delta) = \inf_\delta \sup_\Lambda r(\Lambda,\delta) \tag{15}$$

$$\ge \sup_\Lambda \inf_\delta r(\Lambda,\delta) \tag{16}$$

$$= \sup_\Lambda r_\Lambda \tag{17}$$

Equation (17) gives us a strong tool for lower bounding the minimax risk: for any prior distribution Λ, the Bayes risk under Λ is a lower bound of the corresponding minimax risk. When the condition of the minimax theorem is satisfied (which may be expected due to the bilinearity of $r(\Lambda, \delta)$ in the pair (Λ, δ)), equation (16) achieves equality, which shows that there exists a sequence of priors such that the corresponding Bayes risk sequence converges to the minimax risk.

In practice, it suffices to choose some appropriate prior distribution Λ in order to solve the (nearly) minimax estimator.