Chapter 4: Statistical Decision Theory *

4 Statistical Decision Theory

4.1 Basic Framework and Concepts

- To the usual statistical modeling framework from earlier

$$X, \quad \Theta, \quad \mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

  we add the following elements

  1. some "action space" $\mathcal{A}$ with σ-algebra ϵ,
  2. a suitably measurable "loss function"

$$L(\theta, a) : \Theta \times \mathcal{A} \to [0, \infty),$$

  3. and (non-randomized) decision rules

$$\delta(x) : (\mathcal{X}, \mathcal{B} \to (\mathcal{A}, \epsilon))$$

  For data X, $\delta(x)$ is the action taken based on X.

  To identify "good" devusuib rules δ, we have to average our X, which naturally leads to expectation.

- Risk function The mapping from $\Theta \to [0, \infty)$ given by

$$R(\theta, \delta) \equiv R_\theta L(\theta, \delta(X)) = \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x)$$

  is call the risk function for θ.

  - δ is *at least as good as* δ' if $R(\theta, \delta) \leq R(\theta, \delta')$ for all $\theta \in \Theta$
  - δ is *better than* δ' if $R(\theta, \delta) \leq R(\theta, \delta')$ for all $\theta \in \Theta$, and $R(\theta_0, \delta) < R(\theta_0, \delta')$ for some θ_0
  - δ and δ' are *risk equivalent* if $R(\theta, \delta) = R(\theta, \delta')$ for all $\theta \in \Theta$.
  - δ is *best in a class of decision rules* Δ if $\delta \in \Delta$, and δ is at least as good as any other $\delta' \in \Delta$
  - Example: $X \sim N(\theta, 1), \theta \in \mathbb{R}$ with $\Delta =$"the class of all estimators of θ". There is no best element here. Prove by proposing two constant estimators and zero-one loss.

- If there is no best estimator,
  - Try a smaller and appropriate Δ, e.g. unbiased estimators.
  - Reduce the risk function $R(\theta, \delta)$ to a number and compare numbers for different δ's, e.g.: averaging over θ according to some distribution G on Θ is a way to make "Bayes Risk" and look for "Bayes optimal " decision rules.
  - Maximize $R(\theta, \delta)$ over θ and seek to minimize over δ's, i.e. mini-max procedures.

- Inadmissible: δ is inadmissible in Δ if there exists $\delta' \in \Delta$ that is better than δ.

- Admissible: δ is admissible in Δ if it is not inadmissible in Δ.

  Note: One may never want to use an inadmissible rule, but there are decision problems where every rule is inadmissible.

- Behavorial decision rule: If for each $x \in \mathcal{X}, \phi_x$ is a distribution on $(\mathcal{A}, \epsilon)$, then ϕ_x is called a behavorial decision rule.

*STA643: Advanced Theory of Statistical Inference. Instructed by Dr. Daniel Nordman. Arranged by Zhiling Gu

- $\mathcal{D}^* \equiv \{\phi_x\} \equiv$ the class of behavorial decision rules
- $\mathcal{D} \subset \mathcal{D}^*$ where

$$\mathcal{D} \equiv \{\delta(x)\} \equiv \text{the class of non-randomized decision rules } \delta : \mathcal{X} \to \mathcal{A}$$

- The risk function of a behaorial decision rule is defined as

$$R(\theta, \phi) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) d\phi_x(a) dP_\theta(x)$$

- Randomized decision rule: A randomized decision rule ψ is a probability measure on $(\mathcal{D}, \mathcal{F})$ (δ, with a distribution ψ, becomes a random object and we take decision $\delta(X)$.) Notes:
  - Let $\mathcal{D}_* \equiv \{\psi\} \equiv$ the class of randomized decision rules.
  - It's possible to think of

$$\mathcal{D} \subset \mathcal{D}_*$$

    by associating with $\delta \in \mathcal{D}$ a randomized decision rule ψ_δ which places mass 1 on δ (i.e. , $\psi_\delta(\{\delta\}) = 1$)
  - The risk function of a randomized decision rule is defined as

$$R(\theta, \psi) = \int_{\mathcal{D}} R(\theta, \delta) d\psi(\delta) = \int_{\mathcal{D}} \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x) d\psi(\delta)$$

- Among $(\mathcal{D}, \mathcal{D}^*, \mathcal{D}_*)$, $\mathcal{D}^*$ is perhaps the most natural, while $\mathcal{D}_*$ is the easiest to deal with in some proofs. A natural question is "When are $\mathcal{D}^*$ and $\mathcal{D}_*$ equivalent in termes of generating the same set of risk functions?' It is typically the case under certion space, distribution and loss functions conditions.
  - Example: $(\mathcal{D}, \mathcal{D}^*, \mathcal{D}_*)$ where Behavioural rule and Randomized rule has the same risk function.

$$\begin{aligned}
&X \sim \text{ Bernoulli } (p), \text{ Estimation of } p \in \Theta \equiv [0,1] \equiv \mathcal{A} \\
&\mathcal{X} = \{0,1\}, \quad \mathcal{A} = [0,1], \quad \delta \in D \iff (\delta(0), \delta(1)) \in [0,1] \times [0,1] \equiv \mathcal{A}_0 \\
&\mathcal{D} = \{\delta(x) : \mathcal{X} \to \mathcal{A}\} = \{\delta(x) \mid x = 0,1 \text{ and } \delta(0), \delta(1) \in [0,1]\} \\
&\mathcal{D}^* = \{\phi_x \mid x = 0,1 \text{ and } \phi_0, \phi_1 \text{ are distributions on } \mathcal{A} \equiv [0,1]\} \\
&\mathcal{D}_* = \{\psi \mid \psi \text{ is a probability measure on } (\mathcal{D}, \mathcal{F})\}
\end{aligned}$$

    * $\delta(0) = 0.3$, $\delta(1) = 0.7$ is non-randomized rule
    * $\phi_{X=0} \sim U(0, 0.5), \phi_{X=1} \sim U(0.5, 1)$ then $\phi_X \in D^*$
    * ψ on D, where ψ has a uniform distribution on $(0, 0.5) \times (0.5, 1)$
      Note: if $\tilde{\delta}$ is randomly chosen according to ψ then we observe $X \in \{0,1\}$, we take $\tilde{\delta}(0)$ if $X = 0$, $\tilde{\delta}(1)$ if $X = 1$, so $\psi \in D_*$. That is, first determine the rule, then plug in the observed X.
    * Remark: ϕ_X and ψ in this case are equivalent because

$$\tilde{\delta}(0) \sim U(0, 0.5) \quad \tilde{\delta}(1) \sim U(0.5, 1)$$

- When D^*, D_* contain better tules that are better than those in D? For convex loss functions, rules in D^*, D_* are typically no better.
  - Lemma 51: Suppose that $\mathcal{A}$ is a convex subset of $\mathbb{R}^d$ and ϕ_x is a behavioral decision rule. Define a non-randomized decision rule by

$$\delta(x) = \int_{\mathcal{A}} a d\phi_x(a)$$

    assuming the integral exists. (In the case that $d > 1$, interpret $\delta(x)$ as vector-valued, and the integral as a vector of integrals over d coordinates of $a \in \mathcal{A}$.)
    1. If $L(\theta, \cdot) : \mathcal{A} \to [0, \infty)$ is convex, then

$$R(\theta, \delta) \le R(\theta, \phi)$$

    2. If $L(\theta, \cdot) : \mathcal{A} \to [0, \infty)$ is strictly convex, $R(\theta, \phi) < \infty$ and $P_\theta(\{x \mid \phi_x \text{ is non-degenerate }\}) > 0$, then

$$R(\theta, \delta) < R(\theta, \phi)$$

    Prove by Jensen's Inequality. This lemma shows randomization does not hlep in picking the best decisions. Next two lemmas shows averaging out the randomzation will improve convex loss function.

- Corollary 52: Suppose that $\mathcal{A}$ is a convex subset of $\mathbb{R}^d, \phi_x$ is a behavioral decision rule, and

$$\delta(x)=\int_{\mathcal{A}} a d \phi_x(a)$$

assuming the integral exists.
1. If $L(\theta, a): \mathcal{A} \rightarrow[0, \infty)$ is convex in a for all θ, then δ is at least as good as ϕ
2. If $L(\theta, a): \mathcal{A} \rightarrow[0, \infty)$ is convex in a for all θ and, for some θ_0, the function $L\left(\theta_0, a\right): \mathcal{A} \rightarrow[0, \infty)$ is strictly convex in $a, R\left(\theta_0, \phi\right)<\infty$ and $P_{\theta_0}(\{x \mid \phi_x$ is non-degenerate $\})>0$, then δ is better than ϕ

4.2 Finite Dimensional Geometry of Decision Theory

- A helpful device for understanding some of the basics of decision theory is the geometry involved when

$$\Theta=\left\{\theta_1, \ldots, \theta_k\right\}$$

  Assume that $R(\theta, \psi)<\infty$ for all $\theta \in \Theta$ and $\psi \in \mathcal{D}_*$. Note that in this case

$$R(\cdot, \psi): \Theta \rightarrow[0, \infty)$$

  corresponds to a k -vector in $[0, \infty)^k$

  Let $\mathcal{S}=\left\{y_\psi=\left(y_1, y_2, \ldots, y_k\right) \in \mathbb{R}^k \mid y_i=R\left(\theta_i, \psi\right)\right.$ for all i and some $\left.\psi \in \mathcal{D}_*\right\}=$ the set of all randomized risk vectors.

- Theorem 53: $\mathcal{S}$ is a convex set in $\mathbb{R}^k$
- Let $\mathcal{S}^0=\left\{y_\delta=\left(y_1, y_2, \ldots, y_k\right) \in \mathbb{R}^k \mid y_i=R\left(\theta_i, \delta\right)\right.$ for all i and some $\left.\delta \in \mathcal{D}\right\}=$ the set of all non-randomized risk vectors. It turns out that $\mathcal{S}$ is the convex hull of $\mathcal{S}^0$ (or, equivalently, the smallest convex set containing $\mathcal{S}^0$ or the set of all convex combinations of points in $\mathcal{S}^0$ or the intersection of all convex sets containing $\mathcal{S}^0$)
- Lower Quadrant: Definition 54: For $x=\left(x_1, \ldots, x_k\right) \in \mathbb{R}^k$, the lower quadrant of x is

$$Q_x=\left\{z=\left(z_1, \ldots, z_k\right) \in \mathbb{R}^k \mid z_i \leq x_i \text { for all } i=1, \ldots, k\right\}$$

- Theorem 55: $y \in \mathcal{S}$ (or the decision rule giving rise to y) is admissible if and only if

$$Q_y \cap \mathcal{S}=\{y\}$$

- Definition 56: For $\overline{\mathcal{S}}$ the closure of $\mathcal{S}$, the lower boundary of $\mathcal{S}$ is

$$\lambda(\mathcal{S})=\left\{y \in \mathbb{R}^k \mid Q_y \cap \overline{\mathcal{S}}=\{y\}\right\}$$

- Definition 57 : $\mathcal{S}$ is closed from below if $\lambda(\mathcal{S}) \subset \mathcal{S}$ Denote the set of admissible risk points as

$$A(\mathcal{S})=\left\{y \in \mathbb{R}^k \mid Q_y \cap \mathcal{S}=\{y\}\right\}$$

- Theorem 58: If $\mathcal{S}$ is closed (i.e., $\mathcal{S}=\overline{\mathcal{S}}$), then $\lambda(\mathcal{S})=A(\mathcal{S})$
- Theorem 59: If $\mathcal{S}$ is closed from below, then $\lambda(\mathcal{S})=A(\mathcal{S})$.

4.3 Complete Classes of Decision Rules

- Complete Class Definition 60: A class of decision rules $\mathcal{C} \subset \mathcal{D}^*$ is a complete class (for $\mathcal{D}^*$) if, for any given $\phi \notin \mathcal{C}$, there exists $\phi^{\prime} \in \mathcal{C}$ such that $\phi^{\prime}$ is better than ϕ.

  Remark: This indicates $\mathcal{C}$ contains the best rules that we should focus on.

- Essentially Complete Class Definition 61 : $\mathcal{C} \subset \mathcal{D}^*$ is a called an essentially complete class (for $\mathcal{D}^*$) if, for any given $\phi \notin \mathcal{C}$, there exists $\phi^{\prime} \in \mathcal{C}$ such that $\phi^{\prime}$ is at least as good as ϕ.
- Minimal Complete Class Definition 62 : $\mathcal{C} \subset \mathcal{D}^*$ is a minimal complete class for $\mathcal{D}^*$ if $\mathcal{C}$ is complete and is a subset of any other complete class for $\mathcal{D}^*$. Denote the set of admissible rules in $\mathcal{D}^*$ as $A\left(\mathcal{D}^*\right)$ in the following results.
- Theorem 63: If a minimal complete class $\mathcal{C}$ exists, then $\mathcal{C}=A\left(\mathcal{D}^*\right)$
- Theorem 64: If $A\left(\mathcal{D}^*\right)$ is a complete class, then $A\left(\mathcal{D}^*\right)$ is a minimal complete class.

  Note: The statement " $A\left(\mathcal{D}^*\right)$ is a minimal complete class" is, in general, incorrect. Minimimal complete class does not always exists, when $\mathcal{S}$ is not closed and does not contain the minimum Q_y.