Lecture-15: Statistical decision theory

1 Setting

Definition 1.1. Consider an *observation space* $\mathcal{X}$ and *parameter space* Θ. The set of all probability distributions on the observation space is defined as $\mathcal{M}(\mathcal{X}) \triangleq \left\{P \in [0,1]^{\sigma(X)} : P \text{ satisfies probability axioms}\right\}$. Let $P_\theta \in \mathcal{M}(\mathcal{X})$ for each $\theta \in \Theta$. A *statistical model* refers to a collection

$$\mathcal{P}(\Theta) \triangleq \{P_\theta \in \mathcal{M}(\mathcal{X}) : \theta \in \Theta\}. \tag{1}$$

Remark 1. Without loss of generality, all statistical models can be expressed in the parametric form (1).

Definition 1.2. A statistical model is called *parametric* if Θ is a finite-dimensional Euclidean space so that each distribution is specified by finitely many parameters, and *nonparametric* if Θ is an infinite-dimensional space.

Assumption 1.3. Let $\mathcal{X}, \mathcal{Y}, \Theta$ be the observation, output, and parameter spaces respectively. Let *estimand map* be $T : \Theta \to \mathcal{Y}$ and $\mathcal{P}(\Theta)$ be a statistical model parametrized over parameter space Θ. The observation random variable $X : \Omega \to \mathcal{X}$ is assumed to be generated by distribution $P_\theta \in \mathcal{P}(\Theta)$ and the goal is to estimate $T(\theta)$ based on the observation X.

Example 1.4. Some examples of estimand $T(\theta)$ are $\theta, \mathbb{1}_{\{\theta>0\}}$,sign$(\theta)$, or $\|\theta\|_p$ for some $p \geqslant 1$. If $\Theta \subseteq \mathbb{R}^d$, then an interesting estimand is $T(\theta) \triangleq \max\{\theta_i : i \in [d]\}$. If $\Theta \subseteq \mathbb{R}^{d\times d}$, then an interesting estimand is $T(\theta) \triangleq \max\{\lambda_i : i \in [d]\}$ where $(\lambda_1, \ldots, \lambda_d)$ are eigenvalues of θ.

Example 1.5 (Binary classification). Let $\mathcal{X} \triangleq \mathbb{R}^d, \Theta = \mathcal{Y} \triangleq \{-1, 1\}$, estimand $T(\theta) = \theta$, and an independent labeled training sample $Z \in (\mathcal{X} \times \mathcal{Y})^m$. Defining $I_\theta \triangleq \{i \in [m] : Y_i = \theta\}$, we observe that $(I_\theta : \theta \in \Theta)$ partitions $[m]$. For any parameter $\theta \in \Theta$, we define random vector $X_{I_\theta} \triangleq (X_i : i \in I_\theta)$, which is *i.i.d.* with a common distribution P_θ. We note that this assumption is different than assuming $x \in \mathcal{X}^m$ is *i.i.d.*. However, assuming a prior distribution $\pi \in \mathcal{M}(\Theta)$ we can define a distribution $D \in \mathcal{M}(\mathcal{X})$ for each $B \in \sigma(X)$, as $D(B) \triangleq \mathbb{E}_{\theta\sim\pi} P_\theta(B) = \int_{\theta\in\Theta} d\pi(\theta) P_\theta(B)$. This assumption ensures that $X \in \mathcal{X}^m$ is *i.i.d.* with common .

Definition 1.6. Let $\mathcal{Y}'$ be the prediction space which need not be same as the output space $\mathcal{Y}$. An *estimator* is a random map $\hat{T} : \Omega \to (\mathcal{Y}')^{\mathcal{X}}$ that provides a random estimate $\hat{T}(W)$ for $T(\theta) \in \mathcal{Y}'$ given any observation $W \in \mathcal{X}$. We can denote a random estimator by a probability distribution $P_{\hat{T}(W)|W} \in \mathcal{M}(\mathcal{Y}')$.

Assumption 1.7. We assume that the random estimator $\hat{T}$ is a map such that $\hat{T}(W)$ conditioned on observation W is a random variable independent of everything else and models external randomness.

Remark 2. A deterministic estimator $\hat{T}(W)$ is constant for any given observation W, i.e. $P_{\hat{T}(W)|W} = 1$.

Remark 3. Recall the previous setup where any sample $W : \Omega \to \mathcal{X}$ is *i.i.d.* with an unknown distribution $D \in \mathcal{M}(\mathcal{X})$ and label $Y = c(W)$ for a concept $c \in \mathcal{Y}^{\mathcal{X}}$. We compare the previous setup to this setup where observation W is sampled from distribution $P_\theta \in \mathcal{M}(\mathcal{X})$ under true parameter $\theta \in \Theta$, and the the joint distribution of label $Y \triangleq T(\theta)$ and observation W is defined as $dP_{W,Y}(w, y) = d\pi(\theta) dP_\theta(w) \mathbb{1}_{\{\theta\in T^{-1}(y)\}}$ such that the dependance of the label Y on observation W is given as

$$dP_{Y|W=w}(y) = \frac{\int_{\theta\in T^{-1}(y)} d\pi(\theta) dP_\theta(w)}{\int_{\theta\in\Theta} d\pi(\theta) dP_\theta(w)}.$$

Consider the case when $w = f(\theta)$ for a deterministic and invertible map $f : \Theta \to \mathcal{X}$, then $P_\theta(w) = \mathbb{1}_{\{w=f(\theta)\}}$ and each observation w comes from a deterministic parameter $f^{-1}(w) \in \Theta$ and hence the label $y = T(\theta) = (T \circ f^{-1})(w)$ and the concept $c \triangleq T \circ f^{-1} : \mathcal{X} \to \mathcal{Y}$ in this case.

Example 1.8. $\hat{T}$ may be a confidence interval that aims to contain the scalar $T(\theta)$.

Example 1.9 (Binary classification). We take prediction space $\mathcal{Y}' = \mathcal{Y}$, observation space $\mathcal{X} \subseteq \mathbb{R}^d$, and define linear estimator $\hat{T} : \Omega \to \mathcal{Y}^{\mathcal{X}}$ as $\hat{T}(X) \triangleq \operatorname{sign}\langle w, X\rangle$ for some $w \in \mathbb{R}^d$. This is a deterministic estimator and not depending on the external randomness.

Definition 1.10. To measure the quality of an estimator $\hat{T}$, we introduce a loss function $\ell : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ such that $(T(\theta), \hat{T}(X)) \mapsto \ell(T(\theta), \hat{T}(X))$ is the loss incurred by map $\hat{T}$ in estimating T at parameter θ by observing X.

Remark 4. Since we are dealing with loss, all the negative or converse results are lower bounds and all the positive or achievable results are upper bounds. Note that X is a random variable, so is the estimate $\hat{T}(X)$ and the loss $\ell(T(\theta), \hat{T}(X))$ even for a deterministic estimator.

Definition 1.11. The risk of estimator $\hat{T}$ at a parameter θ under loss ℓ is defined as

$$R_\theta(T, \hat{T}) \triangleq \mathbb{E}[\ell(T(\theta), \hat{T}(X)) \mid \theta] = \mathbb{E}[\mathbb{E}[\ell(T(\theta), \hat{T}(X)) \mid X, \theta] \mid \theta] = \int dP_\theta(x) \int_{y' \in \mathcal{Y}'} \ell(T(\theta), y') dP_{\hat{T}(X)|X=x}(y').$$

Example 1.12 (Binary classification for general parameter space). We take the following parameter space, estimand, prediction space, and loss functions for $\Theta_0 \cap \Theta_1 = \emptyset$,

$$\Theta \triangleq \Theta_0 \cup \Theta_1, \qquad T(\theta) \triangleq \mathbb{1}_{\Theta_1}(\theta), \qquad \mathcal{Y}' \triangleq \{0, 1\}, \qquad \ell(T(\theta), \hat{T}(x)) \triangleq \mathbb{1}_{\left\{T(\theta) \neq \hat{T}(x)\right\}}.$$

Denoting the random set $\mathcal{X}_{\hat{T}(x)} \triangleq \left\{x \in \mathcal{X} : \hat{T}(x) = T(\theta)\right\}$ given observation $X = x$, we can write the expected risk as the probability of error

$$R_\theta(T, \hat{T}) = \mathbb{E}[\mathbb{1}_{\left\{T(\theta) \neq \hat{T}(X)\right\}} \mid \theta] = P_\theta\left\{X \notin \mathcal{X}_{\hat{T}(X)}\right\}.$$

Example 1.13 (Confidence interval estimation). Consider the problem of inference where the goal is to output a confidence interval or region which covers the true parameter with high probability. In this case, $\mathcal{Y} = \Theta = \mathcal{X} \subseteq \mathbb{R}^d$, estimand $T(\theta) = \theta$, prediction space $\mathcal{Y}' \triangleq \mathcal{P}(\Theta)$ and random estimator $\hat{T} : \Omega \to (\mathcal{Y}')^{\mathcal{X}}$. Random estimate $\hat{\theta} \triangleq \hat{T}(x)$ is a subset of parameter Θ for observation $X = x$ and external randomness. The loss function $\ell : \mathcal{Y} \times \mathcal{Y}'$ is defined as $\ell(\theta, \hat{\theta}) \triangleq \mathbb{1}_{\Theta \setminus \hat{\theta}}(\theta) + \lambda \left|\hat{\theta}\right|$ where $\left|\hat{\theta}\right|$ is the volume of region $\hat{\theta}$ and $\lambda > 0$ is some regularization parameter.

Remark 5 (Randomized versus deterministic estimators). Although most of the estimators used in practice are deterministic, there are a number of reasons to consider randomized estimators.

(a) For certain formulations, such as the minimizing worst-case risk (minimax approach), deterministic estimators are suboptimal and it is necessary to randomize. On the other hand, if the objective is to minimize the average risk (Bayes approach), then it does not lose generality to restrict to deterministic estimators.

(b) The space of randomized estimators (viewed as Markov kernels) is convex which is the convex hull of deterministic estimators. This convexification is needed for example for the treatment of minimax theorems.

Lemma 1.14. *If the loss function $\ell : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ is convex in the second argument, then the best estimator is deterministic.*

Proof. Let X be an observation samples from distribution $P_\theta \in \mathcal{M}(\mathcal{X})$ for some parameter $\theta \in \Theta$, we denote label $y \triangleq T(\theta)$, and random prediction $\hat{T}(X)$. From conditional Jensen's inequality applied to the second argument of loss function ℓ, it follows that $R_\theta(T, \hat{T}) = \mathbb{E}[\ell(y, \hat{T}(X)) \mid \theta] = \mathbb{E}[\mathbb{E}[\ell(y, \hat{T}(X)) \mid X, \theta] \mid \theta] \geqslant \mathbb{E}[\ell(y, \mathbb{E}[\hat{T}(X) \mid X, \theta]) \mid \theta]$. □

Remark 6. For any randomized estimator $\hat{T}(X)$, we can derandomize it by considering its conditional expectation $\mathbb{E}[\hat{T}(X) \mid X, \theta]$, which is a deterministic estimator. For convex loss functions, the risk for deterministic estimator dominates that of the random estimator at every parameter θ.

2 Gaussian location model (GLM)

Definition 2.1 (Gaussian location model (GLM) or normal mean model). Consider parameter space $\Theta \subseteq \mathbb{R}^d$ where I_d denotes the d-dimensional identity matrix. *Gaussian location model (GLM)* is the collection of d-dimensional Gaussian distributions parameterized by mean θ and variance σ^2, and denoted

$$\mathcal{P}(\Theta) \triangleq \left\{ \mathcal{N}(\theta, \sigma^2 I_d) : \theta \in \Theta \right\}.$$

Remark 7. For an observation $X : \Omega \to \mathbb{R}^d$ generated by GLM on parameter space $\Theta \subseteq \mathbb{R}^d$, we can write the observation as $X = \theta + Z$ where Z is a zero-mean Gaussian random variable $\mathcal{N}(0, \sigma^2 I_d)$.

Example 2.2 (Parametric spaces $\Theta \subseteq \mathbb{R}^d$ for GLM). Following are some of the examples.

(a) **Unconstrained**. $\Theta = \mathbb{R}^d$.

(b) **ℓ_p-norm balls.** $\Theta = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_p \leqslant 1 \right\}$.

(c) **k-sparse vectors.** $\Theta = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_0 \leqslant k \right\}$ where $\|\theta\|_0 \triangleq |\{i \in [d] : \theta_i \neq 0\}|$ is the size of the support of θ.

(d) **r-rank matrices.** $\Theta = \left\{ \theta \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank} \theta \leqslant r \right\}$. A matrix $\theta \in \mathbb{R}^{d_1 \times d_2}$ can be vectorized into a $d = d_1 \times d_2$ dimensional vector.

Remark 8. Let parameter space $\Theta \subseteq \mathbb{R}^d$, observation space $\mathcal{X} \triangleq \mathbb{R}^d$, and observation X sampled with a Gaussian distribution with mean θ and covariance matrix $\sigma^2 I_d$. Then, we can write the conditional density of X given parameter θ as $\frac{dP_\theta(x)}{dx} \triangleq \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|x-\theta\|_2^2}$.

Example 2.3 (Loss functions and estimators for GLM). Let $\mathcal{Y} = \mathcal{Y}' = \Theta$ where $T(\theta) = \theta$ and denote $\hat{\theta} = \hat{T}(X)$. We consider following loss functions $\ell : \Theta \times \Theta \to \mathbb{R}_+$ defined for all $(\theta, \hat{\theta}) \in \Theta \times \Theta$.

(a) A p-norm loss function defined as $\ell(\theta, \hat{\theta}) \triangleq \left\|\theta - \hat{\theta}\right\|_p^\alpha$ for $p \geqslant 1$ and $\alpha > 0$.

(b) We define the log likelihood loss function defined for each $\theta, \hat{\theta}$ as $\ell(\theta, \hat{\theta}) \triangleq -\ln \frac{dP_{\hat{\theta}}(X)}{dx}$. The resulting estimator that minimizes the loss is called the maximum likelihood estimator (MLE) and denoted by $\hat{\theta}_{\mathrm{ML}}$. From the definition of the log likelihood loss function, minimizing log likelihood loss for GLM is equivalent to maximizing log likelihood of sample X, which for GLM is given by

$$-\ln \frac{dP_{\hat{\theta}}(X)}{dx} = \frac{d}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \left\|X - \hat{\theta}\right\|_2^2.$$

We observe that $\hat{\theta}_{\mathrm{ML}} = X$ maximizes the log-likelihood for GLM.

(c) The resulting estimator based on shrinkage is called the James-Stein estimator $\hat{\theta}_{\mathrm{JS}} \triangleq \left(1 - \frac{(d-2)\sigma^2}{\|X\|_2^2}\right) X$.