Stats 300A session 1

TA: Song Mei

September 27, 2018

1 The decision theory framework

In this section, we go over the concepts in the decision theory framework using the example "biased coin flips", or say "Bernoulli random experiment".

1.1 Decision theory concepts

- Statistical model. A statistical model is a family of distribution $\mathcal{P}$, indexed by a parameter $\boldsymbol{\theta} \in \Theta$. We write
$$\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}.$$
Each $\mathsf{P}_{\boldsymbol{\theta}}$ are probability distribution on the same sample space $\mathcal{X}$. We call the random variable $\boldsymbol{X} \sim \mathsf{P}_{\boldsymbol{\theta}}$ a sample under measure $\mathsf{P}_{\boldsymbol{\theta}}$.

- Loss function. Denote $\mathcal{A} \subset \mathbb{R}^k$ to be the space of possible decision. When decision a is made, and the actual parameter is $\boldsymbol{\theta}$, we incur a loss $L(a, \boldsymbol{\theta})$, where
$$\begin{aligned} L : \mathcal{A} \times \Theta &\to \overline{\mathbb{R}}, \\ (a, \boldsymbol{\theta}) &\mapsto L(a, \boldsymbol{\theta}). \end{aligned}$$

- Statistical procedure/estimator. A statistical procedure is a mapping $A : \mathcal{X} \to \mathcal{A}$. When $\mathcal{A} = \Theta$ and A is an estimate of $\boldsymbol{\theta}$, we also write A as $\hat{\boldsymbol{\theta}}$.

- Risk function. We measure the quality of the procedure A under a loss function L, by the risk function $R(A; \boldsymbol{\theta})$, defined as
$$R(A; \boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}} L(A(\boldsymbol{X}), \boldsymbol{\theta}) = \int_{\mathcal{X}} L(A(\boldsymbol{x}), \boldsymbol{\theta}) \mathsf{P}_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{x}).$$

- Bayes optimality. The risk function is a function of $\boldsymbol{\theta}$, so that we cannot compare different procedure using the risk function directly. Given a prior Q on the space Θ, we define the expected risk function as
$$R_{\mathrm{B}}(A; \mathsf{Q}) = \int_{\Theta} R(A; \boldsymbol{\theta}) \mathsf{Q}(\mathrm{d}\boldsymbol{\theta}).$$
The procedure $A_\star$ is Bayes optimal if it achieves the minimum expected risk
$$R_{\mathrm{B}}(A_\star; \mathsf{Q}) = \inf_A R_{\mathrm{B}}(A; \mathsf{Q}).$$
The minimum expected risk is also called Bayes risk under prior Q.

- Minimax optimality. The worst risk function for a procedure A on parameter space Θ is defined as
$$R_{\mathrm{M}}(A; \Theta) = \sup_{\boldsymbol{\theta} \in \Theta} R(A; \boldsymbol{\theta}).$$
The procedure $A_\star$ is minimax optimal if it achieves the minimum worst risk
$$R_{\mathrm{M}}(A_\star; \Theta) = \inf_A R_{\mathrm{M}}(A; \Theta).$$
The minimum worst risk is also called minimax risk within parameter space Θ.

- Sufficient statistics. Given a statistical model $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ with sample space $(\mathcal{X}, \mathcal{F})$, we say that $\boldsymbol{T} : \mathcal{X} :\to \mathbb{R}^k$ is a sufficient statistics for model $\mathcal{P}$, if the conditional distribution $\boldsymbol{X}$ given $T(\boldsymbol{X})$ is independent of $\boldsymbol{\theta}$. FN factorization criteria: $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(\boldsymbol{T}(\boldsymbol{x}); \boldsymbol{\theta}) f(\boldsymbol{x})$ for some function g and f.

- Randomized statistical procedure (less important for now). Given a probability space $(\mathcal{U}, \mathcal{G}, \mathbb{P})$, a randomized statistical procedure is a map $A : \mathcal{X} \times \mathcal{U} \to \mathcal{A}$. It's risk function is

$$R(A; \boldsymbol{\theta}) = \mathbb{E}\mathsf{E}_{\boldsymbol{\theta}} L(A(\boldsymbol{X}, U), \boldsymbol{\theta}).$$

1.2 Biased coin flip

We observe a sequence of coin flips $(X_1, \ldots, X_n)$ take values in $\{0,1\}^n$, where 0 encodes tails and 1 encodes heads. We denote the probability for head to be $\theta \in [0,1]$.

- Statistical model. In this model, $\Theta = [0,1]$ encodes the space of head probability. For each $\theta \in \Theta = [0,1]$, P_θ is a probability distribution on the sample space $\mathcal{X} = \{0,1\}^n$, such that for each $(x_1, \ldots, x_n) \in \{0,1\}^n$,

$$\mathsf{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}.$$

- Loss function. We will take the action space to be $\mathcal{A} = \Theta = [0,1]$. There are many choice of loss functions. For example, the square loss,

$$L_1(a, \theta) = (a-\theta)^2, \qquad a, \theta \in [0,1].$$

  Another natural choice of loss function is the entropy loss

$$L_2(a, \theta) = \theta \log[\theta/a] + (1-\theta)\log[(1-\theta)/(1-a)] \qquad a, \theta \in [0,1]$$

- Statistical procedure/estimator. The most natural estimator is the sample mean

$$\hat{\theta}_1(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n x_i.$$

  Another estimator gives the sample mean with prior

$$\hat{\theta}_{a,m}(\boldsymbol{x}) = \frac{1}{m+n}\Big[a + \sum_{i=1}^n x_i\Big],$$

  for some m and a. We will see later that this estimator is natural when talking about Bayesian estimator.

- Risk function. It is easier to analytically calculate the risk function under the square loss L_1. The risk function for the sample mean estimator $\hat{\theta}_1$ under square loss L_1 gives

$$R(\hat{\theta}_1, \theta) = \mathsf{E}_\theta\Big[\Big(\frac{1}{n}\sum_{i=1}^n X_i - \theta\Big)^2\Big] = \frac{1}{n^2}\sum_{i=1}^n \mathsf{E}_\theta[(X_i - \theta)^2] = \frac{\theta(1-\theta)}{n}.$$

  The risk function for the other estimator gives

$$\begin{aligned} R(\hat{\theta}_{a,m}, \theta) =& \mathsf{E}_\theta\Big[\Big(\frac{1}{m+n}\Big[a + \sum_{i=1}^n X_i\Big] - \theta\Big)^2\Big] = \frac{1}{(n+m)^2}\sum_{i=1}^n \mathsf{E}_\theta\Big[(X_i-\theta)^2 + \frac{(m\theta - a)^2}{n^2}\Big] \\ =& \frac{n\theta(1-\theta) + (m\theta - a)^2}{n(n+m)^2}. \end{aligned}$$

  For a fixed estimator, the risk function is a function of θ.

- Bayes optimality. The natural prior for this problem gives the beta distribution $\mathsf{Q} \sim \text{Beta}(a, m-a)$, with two parameters a and m. The density of beta distribution gives

$$\mathsf{Q}(\mathrm{d}\theta) = \frac{1}{B(a, m-a)} \theta^a (1-\theta)^{m-a} \mathbf{1}\{\theta \in [0,1]\} \mathrm{d}\theta,$$

where (it is better to get familiarized with the Beta/Gamma function calculus)

$$B(a,b) = \int_0^1 \theta^a (1-\theta)^b \mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

The expected risk of estimator $\hat{\theta}_1$ under prior Q gives

$$R_{\mathrm{B}}(\hat{\theta}_1; \mathsf{Q}) = \frac{1}{n} \int_0^1 \frac{1}{B(a, m-a)} \theta^{a+1} \theta^{m-a+1} \mathrm{d}\theta = \frac{1}{n} \frac{B(a+1, m-a+1)}{B(a, m-a)} = \frac{a(m-a)}{nm}.$$

The expected risk of estimator $\hat{\theta}_{a,m}$ under prior Q gives

$$R_{\mathrm{B}}(\hat{\theta}_{a,m}; \mathsf{Q}) = \frac{1}{n} \int_0^1 \frac{n\theta(1-\theta) + (m\theta - a)^2}{n(n+m)^2} \frac{1}{B(a, m-a)} \theta^a \theta^{m-a} \mathrm{d}\theta = f(n, a, m),$$

for some function f (exercise).

Claim (exercise): $\hat{\theta}_{a,m}$ is the Bayes estimator under the loss L_1 and the prior $\mathsf{Q} \sim \text{Beta}(a, m-a)$.

- Minimax optimality. The worst risk of estimator $\hat{\theta}_1$ within space $\Theta = [0,1]$ gives

$$R_{\mathrm{M}}(\hat{\theta}_1; [0,1]) = \sup_{\theta \in [0,1]} \frac{\theta(1-\theta)}{n} = \frac{1}{4n}.$$

The worst risk of estimator $\hat{\theta}_{a,m}$ within space $\Theta = [0,1]$ gives

$$R_{\mathrm{M}}(\hat{\theta}_{a,m}; [0,1]) = \sup_{\theta \in [0,1]} \frac{n\theta(1-\theta) + (m\theta - a)^2}{n(n+m)^2} = f(n, a, b),$$

for some function of f (exercise).

Claim (exercise): $\hat{\theta}_{\sqrt{n}/2, \sqrt{n}}$ is the minimax estimator under the loss L_1 and the parameter space $[0,1]$, with risk

$$R_{\mathrm{M}}(\hat{\theta}_{\sqrt{n}/2, \sqrt{n}}; [0,1]) = \frac{1}{4(\sqrt{n}+1)^2}.$$

- Sufficient statistics. Denote $T(\boldsymbol{x}) = \sum_{i=1}^n x_i$. We claim that $T(\boldsymbol{x})$ is a sufficient statistics for this model. Now let's check $T(\boldsymbol{x})$ satisfies the definition of sufficient statistics. The conditional distribution of $[\boldsymbol{X} | T(\boldsymbol{X}) = k]$ gives

$$\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = k) = \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = k)}{\mathsf{P}_\theta(T(\boldsymbol{X}) = k)} = \frac{1}{\binom{n}{k}} \mathbf{1}\Big\{ \sum_{i=1}^n x_i = k \Big\}$$

does not depend on θ. To check the FN factorization criteria, the distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ gives

$$\mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \theta^{T(\boldsymbol{x})} (1-\theta)^{n - T(\boldsymbol{x})}.$$

This is in the form of FN factorization criteria for $g(T; \theta) = \theta^T (1-\theta)^{n-T}$ and $f(x) = 1$.

- Randomized statistical procedure. Let $U \sim \text{Unif}([0,1])$ be a uniform random variable independent of $\mathsf{P}_{\boldsymbol{\theta}}$. The randomized statistical procedure is defined as

$$A(\boldsymbol{X}, U) = U.$$

In words, no matter what we observe, we perform action generated by a uniform random number.

Why do we care about randomized procedure?

Claim (see Mackey's notes 2015, lecture 10): under the loss function

$$L(a,\theta) = \begin{cases} 0, & \text{if } |\theta - a| \le \alpha, \\ 1, & \text{otherwise,} \end{cases}$$

for $\alpha < 1/[2(n+1)]$, this randomized statistical procedure is minimax optimal.

2 More on exponential families

Let ν be some reference measure on $\mathbb{R}^n$, and $T_1, \dots, T_d : \mathbb{R}^n \to \mathbb{R}$ be measurable functions. We will write $\boldsymbol{T}(\boldsymbol{x}) = [T_1(\boldsymbol{x}), \dots, T_d(\boldsymbol{x})]^{\mathsf{T}}$. Define partition functions

$$\begin{aligned} Z(\boldsymbol{\theta}) &= \int \exp\{\langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\theta}\rangle\} \nu(\mathrm{d}\boldsymbol{x}), \\ \phi(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta}). \end{aligned}$$

Let $\Theta \subseteq \Theta_\star = \{\boldsymbol{\theta} \in \mathbb{R}^d : Z(\boldsymbol{\theta}) < \infty\}$. Then we can define the statistical model $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where

$$\mathsf{P}_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{\langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\theta}\rangle\} \nu(\mathrm{d}\boldsymbol{x}) = \exp\{\langle \boldsymbol{T}(\boldsymbol{x}), \boldsymbol{\theta}\rangle - \phi(\boldsymbol{\theta})\} \nu(\mathrm{d}\boldsymbol{x}).$$

This is called the exponential family in canonical form.

In homework 1, we are asked to prove that

$$\begin{aligned} \frac{\partial \phi}{\partial \theta_i}(\boldsymbol{\theta}) &= \mathsf{E}_{\boldsymbol{\theta}}[T_i(\boldsymbol{x})], \\ \frac{\partial^2 \phi}{\partial \theta_i \partial \theta_j}(\boldsymbol{\theta}) &= \mathrm{Cov}_{\boldsymbol{\theta}}[T_i(\boldsymbol{x}), T_j(\boldsymbol{x})]. \end{aligned}$$

These are some of the most important properties of exponential families. The second identity implies that $\nabla^2 \phi(\boldsymbol{\theta}) \succeq 0$, so that ϕ is a convex function in $\boldsymbol{\theta}$. Instead of proving these identities, we will show them using some simple examples.

2.1 Examples: Biased coin flip

Recall the statistical model for biased coin flips gives (here instead of using θ as the notation of parameter, we use p)

$$\mathsf{P}_p(\boldsymbol{X} = \boldsymbol{x}) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}.$$

This is not written in the form of canonical exponential family. To write it in the canonical form, we rewrite

$$\mathsf{P}_p(\boldsymbol{X} = \boldsymbol{x}) = \exp\Big\{ \sum_{i=1}^n x_i \log p + \Big(n - \sum_{i=1}^n x_i\Big) \log(1-p) \Big\} = \exp\Big\{ \sum_{i=1}^n x_i \log[p/(1-p)] + n \log(1-p) \Big\}.$$

Denote $\theta = \log[p/(1-p)]$ to be the natural parameter, then $p = e^\theta/(1+e^\theta)$. Let $T(\boldsymbol{x}) = \sum_i x_i$ be the sufficient statistics, $\phi(\theta) = n \log(1+e^\theta)$ be the partition function, $\nu(\mathrm{d}\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \{0,1\}^n} \delta(\boldsymbol{x})$ to be the reference measure. Then we write binomial distribution in its canonical form

$$\mathsf{P}_\theta(\mathrm{d}\boldsymbol{x}) = \exp\{\langle T(\boldsymbol{x}), \theta\rangle - \phi(\theta)\} \nu(\mathrm{d}\boldsymbol{x}).$$

Calculating the mean and variance of $T(\boldsymbol{x})$, we have

$$\mathsf{E}_\theta[T(\boldsymbol{x})] = \mathsf{E}_\theta\Big[\sum_{i=1}^n x_i\Big] = np = ne^\theta/(1+e^\theta),$$

and

$$\begin{aligned}\mathrm{Var}_\theta[T(x)] =& \mathrm{Var}_\theta[x] = \mathsf{E}_\theta[x^2] - \mathsf{E}_\theta[x]^2 \\ &= \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} p^k (1-p)^{n-k} - n^2 p^2 = np(1-p) = ne^\theta/(1+e^\theta)^2.\end{aligned}$$

Calculating the derivatives of the partition function $\phi(\theta)$, we get

$$\phi'(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}[n \log(1+e^\theta)] = ne^\theta/(1+e^\theta),$$

and

$$\phi''(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}[ne^\theta/(1+e^\theta)] = ne^\theta/(1+e^\theta)^2.$$