ORF 524: Statistical Theory and Methods
Homework 3

Lecturer: Samory Kpotufe
AI: Zhuoran Yang
Due: Oct. 28^{th}

Exercise 1 (10 points). *Suppose $\theta \in \mathbb{R}^d$. Show that the following loss functions $\mathcal{L}$ are convex.*

(1). $\mathcal{L}(a) = \|a - \theta\|_p,\ p \geq 1,$

(2). $\mathcal{L}(a) = \|a - \theta\|_p^q,\ p, q \geq 1.$

Exercise 2 (10 points). *Let $X \sim Ber^n(p)$ where $p \in (0,1)$. Consider the naive estimator $\widehat{p} = X_1$, and let $\widetilde{p} = \mathbb{E}(\widehat{p}|T)$, where $T(X) = \sum_{i=1}^n X_i$.*

(1). Derive $\widetilde{p}$.

(2). Compute and compare $\mathbb{E}(\widehat{p} - p)^2$ and $\mathbb{E}(\widetilde{p} - p)^2$.

Exercise 3 (10 points). *Let us prove Rao-Blackwell theorem in a different way. Denote the ℓ_2 loss by $R(\widehat{\theta}) = \mathbb{E}\|\widehat{\theta} - \theta\|^2$.*
(1). Show that for any two random variables X, Y defined on the same space $(\Omega, \Sigma, \mathbb{P})$,

$$\mathrm{Var}(X) = \mathrm{Var}[\mathbb{E}(X|Y)] + \mathbb{E}[\mathrm{Var}(X|Y)]. \tag{1}$$

Recall that by definition, $\mathbb{E}(X|Y)$ is a function of Y, and $\mathrm{Var}(X|Y)$ is defined to be $\mathbb{E}\{[X - \mathbb{E}(X|Y)]^2|Y\}$.
Hint: use the fact that $\mathbb{E}X = \mathbb{E}[\mathbb{E}(X|Y)]$.
(2). Use the decomposition of variance in (1) to prove Rao-Blackwell theorem.

Exercise 4 (10 points). *In K-means algorithm, we do the following steps:*

- *Step 1: start with $\{c_j\}^0$, with each of K initial centers in $\mathbb{R}^d$.*
- *Step 2: for $l = 0, 1, \ldots$*
  - *Assign each x_j to the nearest center $c_j^l \in \{c_j\}^l$. In a tie appears, x_j is assigned to one of its nearest centers in an arbitrary way.*
  - *Updating the centers: compute the mean of points in class C_j^l, where C_j^l consists of points that are assigned to the center c_j^l. Let c_j^{l+1} be that mean.*

Show that K-means algorithm converges (not necessarily to its global minimizer), i.e. $\lim_{l \to \infty} \phi(\{c_j\}^l)$ exists, where

$$\phi(\{c_j\}^l) = \sum_{j=1}^K \sum_{x_i \in C_j^l} \|x_i - c_j^l\|^2.$$

Exercise 5 (Cramer-Rao Theorem in the general case (10points)). *Suppose that $\theta \in \mathbb{R}^p$, and a statistic $T(X)$ is also in $\mathbb{R}^p$, with $\mathbb{E}_\theta T(X) = g(\theta)$ for some function $g \colon \mathbb{R}^d \to \mathbb{R}^d$. Cramer-Rao Theorem states that under the same regularity condition as in the univariate case, we have*

$$\mathrm{Cov}(T) \succeq \nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta),$$

where $A \succeq B$ means $A - B$ is a positive semi-definite matrix, $(\nabla_\theta g(\theta))_{ij} = \frac{\partial}{\partial \theta_i} g_j(\theta)$, and

$$I(\theta) = \mathbb{E} \nabla_\theta \log f_\theta(x) (\nabla_\theta \log f_\theta(x))^T. \qquad (\textit{assuming existence and invertibility})$$

You can prove this theorem using the following ideas.

(1). Let $a(x) = \nabla_\theta \log f_\theta(x)$. Derive $\mathrm{Cov}((T, a)^T)$ in terms of $\mathrm{Cov}(T)$, $\nabla_\theta g(\theta)$ and $I(\theta)$.

(2). Find a matrix B such that

$$B^T \,\mathrm{Cov} \begin{pmatrix} T \\ a \end{pmatrix} B = \mathrm{Cov}(T) - \nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta).$$

(3). Conclude the proof of Cramer-Rao Theorem.

Exercise 6 (10 points). *Suppose that $f_\theta(x)$ is the density function of $P_\theta \in \mathcal{P}$, where $\theta \in \mathbb{R}^k$. We assume that $f_\theta(x)$ is twice differentiable in θ, satisfying the regular condition, meaning that*

$$\nabla_\theta \int h_\theta(x)\, dx = \int \nabla_\theta h_\theta(x)\, dx$$

holds for $h_\theta = f_\theta$ and $h_\theta = \nabla_\theta f_\theta$. Show that

$$I(\theta) = -\mathbb{E}\big[\nabla_\theta^2 \log f_\theta(X)\big].$$

Exercise 7 (10 points). *Let*

$$f_\alpha(x) = h(x) l(\alpha) e^{\alpha T(x)} \qquad x \in \mathbb{R}$$

be the density function of probability measure P_α in the exponential family $\mathcal{P} = \{P_\alpha : \alpha \in \mathcal{A}\}$. Suppose that $\mathcal{A}$ is an open set in $\mathbb{R}$.
(1) Show that

$$\left| \frac{e^{az} - 1}{z} \right| \le \frac{e^{\delta |a|}}{\delta}$$

holds for $|z| \le \delta$.
(2) Use the Dominated Convergence Theorem to show the following. Let g be a Borel function such that $\mathbb{E} g < \infty$. Show that $\frac{d}{d\alpha} \mathbb{E} g = \int g(x) \frac{d}{d\alpha} f_\alpha(x)\, dx$ (you may assume the l.h.s. is differentiable).

Exercise 8 (10points). *Consider the fixed design model $y = X\beta + \eta$, where $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^d$, and $X \in \mathbb{R}^{n \times d}$. Here, 'fixed design' means that X is a deterministic matrix. Suppose that $n \ge d$, $X^T X$ is invertible, and $\eta \sim N(0, \sigma^2 I_n)$, where σ^2 is known.*
(1) Show that $\widehat{\beta} = (X^T X)^{-1} X^T y$ is a UMVUE of β.
(2) Derive the mean square error $R_2(\widehat{\beta})$ in terms of σ^2 and $(X^T X)^{-1}$ only. What would $R_2(\widehat{\beta})$ be if $n^{-1} X^T X$ is equal to the identity matrix?

Exercise 9 (10 points). *In a Bayesian setting where the prior distribution of θ is π. The Bayesian risk is defined as $\overline{R}_{\mathcal{L}}(\widehat{\theta}) = \mathbb{E}_{\theta} R_{\mathcal{L}}(\widehat{\theta}, \theta)$. Recall that the risk function $R_{\mathcal{L}}(\widehat{\theta}, \theta)$ is defined as $R_{\mathcal{L}}(\widehat{\theta}, \theta) = \mathbb{E}_{X|\theta} \mathcal{L}(\widehat{\theta}(X), \theta)$. We assume that all the integrals exist. Answer the following questions.*

(1) Let X be a continuous random variable with distribution P_X. Show that $\mathbb{E}|X - c|$ is minimized at $c = \mathit{Median}(P_X)$, where $\mathit{Median}(P_X)$ is the median of P_X. Assume for simplicity that the cdf is strictly increasing, in which case the median is unique, and is just $F_X^{-1}(1/2)$.

(2) Let θ, X be jointly continuous, where $\theta \in \mathbb{R}$, and $X \in \mathbb{R}^d$. Find a minimizer $\widehat{\theta}$ for $\overline{R}_{\mathcal{L}}(\widehat{\theta})$ where

$$\mathcal{L}(\widehat{\theta}(X), \theta) = |\widehat{\theta}(X) - \theta|.$$

(3) Let $\theta \in \Theta$ and $X \in \mathbb{R}^d$, where $\Theta = \{1, 2, \ldots, K\}$. Find a minimizer $\widehat{\theta}$ for $\overline{R}_{\mathcal{L}}(\widehat{\theta})$ where

$$\mathcal{L}(\widehat{\theta}(X), \theta) = 1\{\widehat{\theta}(X) \neq \theta\}.$$

(4) Let $\theta \in \mathbb{R}^k$, and $X \in \mathbb{R}^d$. Find a minimizer $\widehat{\theta}$ for $\overline{R}_{\mathcal{L}}(\widehat{\theta})$ where $\mathcal{L}(\widehat{\theta}(X), \theta) = \|\widehat{\theta}(X) - \theta\|_2^2$.

Exercise 10 (10 points). *Let $\mathcal{P} = \{\mathbb{P}_\theta = \mathit{Unifrom}[0, \theta], \theta > 0\}$, that is, $\mathcal{P}$ is the family of uniform distributions. Let $X = \{X_i\}_{i=1}^n$ be n i.i.d. realizations of some $\mathbb{P}_\theta \in \mathcal{P}$. Show that*

$$T(X) = \max_{1 \le i \le n} X_i$$

is both sufficient and complete.

Hint: consider $\frac{\mathrm{d}}{\mathrm{d}\theta} \mathbb{E}[h(T)]$ for some Borel measurable function h.

Exercise 11 (10 points). *An ancillary statistic S for a family $\mathcal{P} = \{\mathbb{P}\}$ is one that has no information on $\mathbb{P}$. That is, the distribution of $S(X)$, when $X \sim \mathbb{P}$, is the same for all $\mathbb{P} \in \mathcal{P}$. Show that if T is complete for $\mathcal{P}$, then T and any ancillary statistic S are uncorrelated. Assume that both S and T are in $\mathbb{R}^d$.*

Note: As an example of ancillary statistic, we consider

$$\mathcal{P} = \{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \textit{ fixed and known}\}.$$

Then statistic

$$T(X) = \frac{(n-1)S_{n-1}^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \sim \chi_{n-1}^2,$$

which is an ancillary statistic.

Hint: Consider any statistic S' that is an unbiased estimator of zero vector, i.e., $\mathbb{E}S' = 0 \in \mathbb{R}^d$ for any $\mathbb{P} \in \mathcal{P}$. First show that $\mathrm{Cov}(S', T) = \mathbb{E}(S' - \mathbb{E}S')(T - \mathbb{E}T)^\top = 0$. Now noticing that $\mathbb{E}S$ is a constant for any $\mathbb{P} \in \mathcal{P}$, conclude that $\mathrm{Cov}(S, T) = 0$.