STAT 535, Homework 5

Due date: Dec 8 Thursday 23:59:59. Submit the homework through Canvas in a PDF file. If the questions involved programming, please include your codes.

1. *Simple bootstrap problem.* Assume that your data consists of $x_1, \cdots, x_n$, n values. When we generate the bootstrap sample, we sample with replacement of these n points to obtain a set of IID new points $X_1^*, \cdots, X_n^*$ such that

$$P(X_\ell^* = x_1) = P(X_\ell^* = x_2) = \cdots = P(X_\ell^* = x_n) = \frac{1}{n} \tag{1}$$

for each ℓ. This new dataset, $X_1^*, \cdots, X_n^*$, is called a bootstrap sample.

   (a) **(5 pts)** Show that the bootstrap sample is an IID random sample from $\widehat{F}_n$, where

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x),$$

   is the EDF formed by the original data points $x_1, \cdots, x_n$.

   (b) **(5 pts)** Assume we want to use the bootstrap to estimate the variance of the sample mean. It is well-known that the variance of the sample mean can be approximated by the sample variance divided by n, the sample size. Namely,

$$\widehat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

   Let $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$ be the sample mean of a *bootstrap sample*. Given the original data $x_1, \cdots, x_n$ being fixed, show that

$$\mathsf{Var}(\bar{X}_n^*) = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{n-1}{n^2} \cdot S_n^2.$$

   (this implies $\mathsf{Var}(\bar{X}_n^*) \approx S_n^2/n$ when the sample size is large)

2. *Bootstrap and contingency table.* In this problem, we will use the bootstrap to analyze the odds ratio of UC Berkeley's admission dataset, a built in dataset in R. In particular, we will focus on the department `A`. To obtain this dataset, use the command `UCBAdmissions[,,1]` in R. It is a 2 by 2 contingency table as the follows: The table is a summary of a set of observations. The original data will be a matrix like

	Male	Female
Admitted	512	89
Rejected	313	19

The product of the diagonal terms of the matrix (512 and 19) divided by the product of the off-diagonal terms (313 and 89), is called the odds ratio. In this case, the odds ratio $OR = \frac{512 \cdot 19}{313 \cdot 89} \approx 0.349212$.

   (a) **(3 pt)** Use the bootstrap to compute the MSE of the odds ratio OR.

ID	Gender	Outcome
001	Female	Admitted
002	Male	Admitted
003	Male	Rejected
004	Male	Rejected
005	Female	Rejected
006	Female	Admitted
⋮	⋮	⋮

(b) **(3 pt)** If there is no gender bias, the odds ratio will be 1. Use the bootstrap to compute the p-value of testing

$$H_0 : \text{no gender bias in this contingency table.}$$

(c) **(4 pt)** In this case, the parametric bootstrap (sampling from a fitted parametric model–in this case, the parametric model is a multinomial on the 4 categories) and the empirical bootstrap are the same procedure. Explain why.

3. *Uniform bounds.* Consider a non-linear regression problem where we observe

$$(X_1, Y_1) \cdots, (X_n, Y_n) \sim F$$

and both X, Y are univariate and are uniformly bounded by L (i.e.,$|X_i| \leq L$ and $|Y_i| \leq L$). We are attempting to fit a non-linear regression model $m(x) = \alpha + \beta x^\gamma$ so the empirical risk is

$$R_n(\alpha, \beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i^\gamma)^2.$$

Our estimator is

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) = \mathsf{argmin}_{(\alpha,\beta,\gamma)\in\Theta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i^\gamma)^2,$$

where $\Theta \subset \mathbb{R}^3$ is the parameter space and is assumed to be a compact set.

(a) **(2 pts)** A population that the estimator is approximating is

$$(\alpha^*, \beta^*, \gamma^*) = \mathsf{argmin}_{(\alpha,\beta,\gamma)\in\Theta} R(\alpha, \beta, \gamma)$$

and $R_n(\alpha, \beta, \gamma)$ is an unbiased estimator of $R(\alpha, \beta, \gamma)$. What is $R(\alpha, \beta, \gamma)$?

(b) **(5 pts)** Under appropriate conditions, for a given (α, β, γ),

$$\sqrt{n}(R_n(\alpha, \beta, \gamma) - R(\alpha, \beta, \gamma)) \xrightarrow{D} N(0, \sigma^2(\alpha, \beta, \gamma)).$$

Moreover, for any two pairs $(\alpha_1, \beta_1, \gamma_1)$ and $(\alpha_2, \beta_2, \gamma_2)$,

$$\sqrt{n} \begin{pmatrix} R_n(\alpha_1, \beta_1, \gamma_1) - R(\alpha_1, \beta_1, \gamma_1) \\ R_n(\alpha_2, \beta_2, \gamma_2) - R(\alpha_2, \beta_2, \gamma_2) \end{pmatrix} \xrightarrow{D} N\left(0, \Sigma(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)\right),$$

where $\Sigma(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)$ is a 2×2 matrix. Write down $\Sigma(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)$.

Note: in fact, you can easily generalize this to any set of k points in the parameter space. This gives you a hint about why $\sqrt{n}(R_n(\alpha, \beta, \gamma) - R(\alpha, \beta, \gamma))$ converges to a Gaussian process.

(c) **(3 pts)** One common way to show that the estimator $(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma})$ is as good as the optimal predictor $(\alpha^*, \beta^*, \gamma^*)$ is to argue that the performance in the population level (can be viewed as prediction for future data)

$$R(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) - R(\alpha^*, \beta^*, \gamma^*) \le \epsilon$$

for some given ϵ. Show that this can be established if we have

$$\sup_{(\alpha,\beta,\gamma)\in\Theta} |R_n(\alpha, \beta, \gamma) - R(\alpha, \beta, \gamma)| \le \frac{\epsilon}{2}.$$

Namely, if we have a uniform bound on the empirical process, then we have a bound on the excess risk.

(d) **(5 pts)** Here is a possible way to establish the uniform bound in the previous question. Define $f_\theta(x, y) = f_{\alpha,\beta,\gamma}(x, y) = (y - \alpha - \beta x^\gamma)^2$, where $\theta = (\alpha, \beta, \gamma)$. Using the notation of empirical process, we have $\widehat{\mathbb{P}}_n(f_\theta) = \frac{1}{n}\sum_{i=1}^n f_\theta(X_i, Y_i) = R_n(\theta)$ and $\mathbb{P}(f_\theta) = \mathbb{E}(f_\theta(X_i, Y_i)) = R(\theta)$. To simplify the problem, let B be the bound that

$$\sup_{\theta\in\Theta} |f_\theta(X, Y)| \le B.$$

Thus, the uniform bound can be written as

$$\sup_{(\alpha,\beta,\gamma)\in\Theta} |R_n(\alpha, \beta, \gamma) - R(\alpha, \beta, \gamma)| = \sup_{\theta\in\Theta} |\widehat{\mathbb{P}}_n(f_\theta) - \mathbb{P}(f_\theta)|.$$

Although the supremum is taken over the entire parameter space, we may approximate the entire space by a set of points $\theta_1, \cdots, \theta_N \in \Theta$ such that for any $\theta \in \Theta$, there exists one point θ_j with

$$|\widehat{\mathbb{P}}_n(f_\theta) - \widehat{\mathbb{P}}_n(f_{\theta_j})| \le \frac{\epsilon}{3}, \quad |\widehat{\mathbb{P}}(f_\theta) - \widehat{\mathbb{P}}(f_{\theta_j})| \le \frac{\epsilon}{3}.$$

Of course, the number of points $N = N(\epsilon)$ depends on the precision we enforce.

With this, we can then revise upper bound problem as

$$\begin{aligned}
&\sup_{(\alpha,\beta,\gamma)\in\Theta} |R_n(\alpha, \beta, \gamma) - R(\alpha, \beta, \gamma)| \\
&\qquad = \sup_{\theta\in\Theta} |\widehat{\mathbb{P}}_n(f_\theta) - \mathbb{P}(f_\theta)| \\
&\qquad \le \sup_{\theta\in\Theta} |\widehat{\mathbb{P}}_n(f_\theta) - \widehat{\mathbb{P}}_n(f_{\theta_j})| + \max_{j=1,\cdots,N} |\widehat{\mathbb{P}}_n(f_{\theta_j}) - \mathbb{P}(f_{\theta_j})| + \sup_{\theta\in\Theta} |\mathbb{P}(f_\theta) - \mathbb{P}(f_{\theta_j})| \\
&\qquad \le \frac{2}{3}\epsilon + \max_{j=1,\cdots,N} |\widehat{\mathbb{P}}_n(f_{\theta_j}) - \mathbb{P}(f_{\theta_j})|.
\end{aligned}$$

We can bound $\max_{j=1,\cdots,N} |\widehat{\mathbb{P}}_n(f_{\theta_j}) - \mathbb{P}(f_{\theta_j})|$ using the Hoeffding's inequality with the bound $\sup_{\theta\in\Theta} |f_\theta(X, Y)| \le B$.

Using the above result, provide a concentration bound on

$$P\left(\sup_{(\alpha,\beta,\gamma)\in\Theta} |R_n(\alpha, \beta, \gamma) - R(\alpha, \beta, \gamma)| \ge \epsilon\right).$$

Note: The quantity $N(\epsilon)$ is called the ϵ-*covering number* (in this case, we need the covering number with respect to L_∞ norm).

4. *Simple missing data.* Consider a problem where we have two random variables X, Y such that $X \in R$ and $Y \in \mathbb{R} \cup \{\texttt{NA}\}$. However, Y may not be observed (this occurs when $Y = \texttt{NA}$). So we introduce the binary random variable $R \in \{0, 1\}$ such that $R = 1$ if Y is observed ($Y \in \mathbb{R}$) and $R = 0$ if Y is missing ($Y = \texttt{NA}$).

   Let $(X_1, Y_1, R_1), \cdots, (X_n, Y_n, R_n)$ be IID random variables representing our data.

   (a) **(5 pts)** Consider a quantity

   $$\widehat{\mu}_{\mathsf{naive}} = \frac{\sum_{i=1}^n R_i Y_i}{\sum_{j=1}^n R_j}.$$

   Show that $\widehat{\mu}_{\mathsf{naive}}$ is a consistent estimator of $\mu = \mathbb{E}(Y)$ if Y and R are uncorrelated.

   (b) **(5 pts)** Suppose we want to estimate $\theta = \mathbb{E}(Xe^Y)$. To deal with missingness, we assume that $Y \perp R | X$. Derive an inverse probability weighting estimator. You need to make sure your estimator can be computed with missing data.

   (c) **(5 pts)** *Connection to transfer learning.* Assume $Y \perp R | X$. Suppose we are interested in estimating $\mu_0 = \mathbb{E}(Y|R = 0)$. It is known that $\mu_1 = \mathbb{E}(Y|R = 1)$ can be estimated by using $\mu_1 = \frac{\mathbb{E}(YR)}{\mathbb{E}(R)}$.

   This motivates us to consider the quantity

   $$\phi(g) = \frac{\mathbb{E}(YR \cdot g(X))}{\mathbb{E}(1 - R)}.$$

   It turns out that there exists a function $g_{1 \to 0}(x)$ such that

   $$\phi(g_{1 \to 0}) = \mu_0.$$

   Find the function $g_{1 \to 0}(x)$ and design a consistent estimator of it.