Tools and Techniques in Machine Learning
Homework 1: Missing data and inverse propensity weighting

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. It's preferred that you write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or Jupyter), though if you need to you may scan handwritten work. For submission, you can also export your Jupyter notebook and merge that PDF with your PDF for the written solutions into one file. **Don't forget to complete the Jupyter notebook as well, for the programming part of this assignment**.

1 Estimators for missing at random (MAR)

All questions below pertain to the missing at random (MAR) setting. Let's review the MAR setup: $(X, R, Y), (X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$ are i.i.d. with covariate $X \in \mathcal{X}$, response indicator $R \in \{0, 1\}$, and response $Y \in \mathbb{R}$. Under MAR, we assume that $R \perp\!\!\!\perp Y \mid X$, and the response probability is given by $\mathbb{P}(R = 1 \mid X = x) = \pi(x) \in (0, 1]$, where $\pi(x)$ is the propensity score function. The Y_i's corresponding to $R_i = 0$ are unobserved. The missing data problem is to estimate $\mathbb{E}Y$ without using the unobserved Y_i's, that is, using only $(X_1, R_1, R_1Y_1), \ldots, (X_n, R_n, R_nY_n)$.

1.1 Total inverse propensity weight for observations has expectation n

Let $W_i = \frac{1}{\pi(X_i)}$ be the inverse propensity weight for Y_i.

1. Show that
$$\mathbb{E}\left[\sum_{i=1}^{n} W_i R_i\right] = n.$$

1.2 Complete case estimator is not consistent for MAR setting.

1. The complete case mean is defined as $\hat{\mu}_{\text{cc}} = \sum_{i=1}^{n} R_i Y_i / \sum_{i=1}^{n} R_i$. Show that under the MAR assumption,
$$\hat{\mu}_{\text{cc}} \xrightarrow{P} \frac{\mathbb{E}[\pi(X)\mu(X)]}{\mathbb{E}[\pi(X)]},$$
where $\mu(x) = \mathbb{E}[Y|X = x]$. Assume[1] that $\mathbb{E}|RY| < \infty$ and $\mathbb{E}|R| < \infty$. [Hint: The weak law of large numbers (WLLN) states that if $Y, Y_1, \ldots, Y_n$ are i.i.d. with $\mathbb{E}|Y| < \infty$, then

[1] The first first inequality is true if Y is bounded, which is reasonable in all our applications, and the second inequality is clearly true since $R \in \{0, 1\}$.

$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{P} \mathbb{E}Y$. Apply the WLLN on the numerator and denominator separately, and then apply Slutsky's Theorem, which states that if $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{P} b$ for constants a and $b \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{P} \frac{a}{b}$.]

2. Recall the SeaVan1 distribution from lecture (which is based on Example 1 in [SV18]):

$$\begin{aligned} X &\sim \operatorname{Unif}(\{0,1,2\}) \\ Y \mid X = x &\sim \mathcal{N}(x,1) \\ R \mid X = x &\sim \operatorname{expit}(4-4x), \end{aligned}$$

where $\operatorname{expit}(x) = 1/\left(1+e^{-x}\right)$. What does the complete case mean converge to for the SeaVan1 distribution [with at least 2 decimal places accuracy]? What is $\mathbb{E}Y$?. The large gap between the two is why we need to develop more sophisticated estimators to handle response bias.

1.3 IPW estimator is not equivariant

Suppose $\mathcal{D}$ represents the dataset $(X_1, R_1, R_1Y_1), \ldots, (X_n, R_n, R_nY_n)$. For any $a \in \mathbb{R}$, we'll write $\mathcal{D}+a$ for the dataset $(X_1, R_1, R_1(Y_1+a)), \ldots, (X_n, R_n, R_n(Y_n+a))$, which is the same as $\mathcal{D}$, but with each Y value shifted by a. We say that an estimator $\hat{\mu}(\mathcal{D})$ is **equivariant** if $\hat{\mu}(\mathcal{D}+a) = \hat{\mu}(\mathcal{D})+a$ for any $\mathcal{D}$. In other words, adding a to all the responses Y_i just shifts the estimate by the same amount a. (This definition is based on [LC98, Ch 3].)

1. Show that the self-normalized IPW estimator $\hat{\mu}_{\text{sn\_ipw}}$ is equivariant. Explain why this implies the complete case estimator $\hat{\mu}_{\text{cc}}$ is also equivariant.

2. Show that $\hat{\mu}_{\text{ipw}}(\mathcal{D}+a) = \hat{\mu}_{\text{ipw}}(\mathcal{D}) + \frac{a}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}$ and demonstrate that $\hat{\mu}_{\text{ipw}}$ is generally not equivariant (though it is if $\pi(x) \equiv 1$).

3. Consider the estimator $\hat{\mu}_{\text{ipw}-a}(\mathcal{D}) := \hat{\mu}_{\text{ipw}}(\mathcal{D}+a) - a$. Show that $\hat{\mu}_{\text{ipw}-a}(\mathcal{D})$ is an unbiased estimator of $\mathbb{E}Y$. [Hint: We already know that $\hat{\mu}_{\text{ipw}}(\mathcal{D})$ is an unbiased estimator of $\mathbb{E}Y$.]
Remark: By varying a, we can get a whole collection of unbiased estimators of $\mathbb{E}Y$. Some will be better than others. We'll revisit this setup in our next homework, where we'll view $\hat{\mu}_{\text{ipw}-a}(\mathcal{D})$ as a control variate adjustment of $\hat{\mu}_{\text{ipw}}$, with the hope that a judicious choice of a will lead to an estimator with reduced variance.

References

[LC98] E.L. Lehmann and George Casella, *Theory of point estimation*, 2nd ed., Springer, 1998. 1.3

[SV18] Shaun R. Seaman and Stijn Vansteelandt, *Introduction to double robust methods for incomplete data*, Statistical Science **33** (2018), no. 2, 184–197. 2