CIS 7000 Spring 2024 Homework 2

Due March 18, 2024

In this problem, we will consider how to construct PAC prediction sets under covariate shift. In particular, suppose we are given a calibration dataset $Z \sim P^n$ (i.e., $Z = \{(x_1, y_1^*), ..., (x_n, y_n^*)\}$ consisting of i.i.d. samples from P), and we know the importance weights $w(x) = q(x)/p(x)$. You can assume that the cumulative distribution functions (CDFs) $F_P(x)$ and $F_Q(x)$ are invertible. In addition, you can assume that the importance weights satisfy $w(x) \leq w_{\max}$ for all $x \in \mathcal{X}$. We consider two different strategies for constructing PAC prediction sets.

1. Our first strategy is to devise a relationship between the optimal parameters for P and Q. In particular, let $\tau_P^*(\epsilon) = F_P^{-1}(\epsilon)$ and $\tau_Q^*(\epsilon) = F_Q^{-1}(\epsilon)$ be the optimal parameters for P and Q at error level ϵ, respectively.
   (a) Prove that for all $\alpha \in \mathbb{R}$, we have $F_Q(\alpha) \leq F_P(\alpha) \cdot w_{\max}$.
   (b) Using the previous result, prove that $\tau_Q^*(\epsilon) \geq \tau_P^*(\epsilon/w_{\max})$. [Hint: Note that F_P^{-1} and F_Q^{-1} are monotonically increasing.]
   (c) Describe how to translate these results into a PAC prediction set algorithm.
2. Our second strategy is to use rejection sampling to convert the set of i.i.d. samples $Z \sim P^n$ into a set of i.i.d. samples $Z' \sim Q^{n'}$, for some $n' \leq n$.
   (a) Suppose that P and Q are discrete distributions. Let the importance weights $w(x) = q(x)/p(x)$ be bounded, i.e., $w(x) \leq w_{\max}$ for all $x \in \mathcal{X}$. Consider the following rejection sampling strategy: (i) draw a random sample $z = (x, y^*) \sim P$, (ii) sample $b \sim \text{Bernoulli}(w(x)/w_{\max})$, and (iii) define

   $$z' = \begin{cases} z & \text{if } b = 1 \\ \varnothing & \text{if } b = 0. \end{cases}$$

   Show that the distribution of z is

   $$r(z') = \begin{cases} \frac{q(z')}{w_{\max}} & \text{if } z' \neq \varnothing \\ 1 - \frac{1}{w_{\max}} & \text{if } z' = \varnothing. \end{cases}$$

   (b) Based on the previous result, it can be shown (you do not need to do so) that if we use the described process for rejection sampling (i.e., resample z' until $z' \neq \varnothing$), then we obtain a single sample $z' \sim Q$. You can assume that this result also holds for continuous random variables. Describe how this result into a PAC prediction set algorithm.