STA 561: Homeworks 7 & 8 (Due April 17 at midnight)

Reminder: work together! Share ideas, brainstorm, explain/verify your answers but write up your own work. Your homework should be submitted as pdf file generated using either latex or an python notebook.

1. (A simple decision model.) Consider the following generative model for (X, A, Y)

   - $X^{1/5} \sim \text{Normal}(0, 1)$
   - $A = 1_{|X| \leq \tau}$
   - $Y = \beta_0 + \beta_1 A + \beta_2 X + \beta_3 AX + \epsilon$, where $\epsilon \sim \text{Normal}(0, 1)$.

   Suppose that $\beta_0 = \beta_1 = \beta_2 = 1$ and $\beta_3 = 0.5$. What is the optimal decision rule? I.e., mapping $\pi : \mathbb{R} \rightarrow \{0, 1\}$ such that if decisions are assigned according to π the value $V(\pi)$ is maximized. Generate 1000 data sets of size $n = 500$ from this model for $\tau = 0.01$ and $\tau = 0.025$ and use OLS to estimate $\beta_0, \ldots, \beta_3$. In what proportion of your data sets was the p-value for β_3 significant? What's happening here? Are the standard causal assumptions verified?

2. (Run on sentence, run on.) Suppose that we have a black-box regression model that inputs data of the form $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ and outputs an estimator $\widehat{f}_n(\boldsymbol{x})$ of $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. Our goal in this problem to explore approximating $\widehat{f}_n$ with a kernel. For this problem you'll be exploring two approaches for constructing kernels: (i) born-again random forests in which you will generate many inputs $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_B$ (where B is large) from the convex hull of the support of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, create outputs $\widehat{f}_n(\boldsymbol{Z}_1), \ldots, \widehat{f}_n(\boldsymbol{Z}_B)$, then fit $\left\{(\boldsymbol{Z}_b, \widehat{f}_n(\boldsymbol{Z}_b))\right\}_{b=1}^{B}$ using a random forest from which you can extract the random forest kernel; (ii) ad-hoc kernels in which you will repeat the following steps for $b = 1, \ldots, B$

   - bootstrap the data
   - randomly select a subset of the predictors (columns of $\boldsymbol{X}$) by drawing M entries without replacement from $\{1, \ldots, p\}$
   - apply the black box to the bootstrapped and column-subset data to obtain $\widehat{f}_n^{(b)}$
   - draw random seeds $y_1, \ldots, y_L$ uniformly from $Y_1, \ldots, Y_n$ and corresponding Voronoi partition of $\mathbb{R}$

then for any $\boldsymbol{x} \in \mathbb{R}^p$ define $A_n^{(b)}(\boldsymbol{x})$ to the partition to which $\widehat{f}_n^{(b)}(\boldsymbol{x})$ belongs and define the kernel distance between two points $\boldsymbol{x}, \boldsymbol{z}$ to be

$$K(\boldsymbol{x}, \boldsymbol{z}) \triangleq \frac{1}{B} \sum_{b=1}^{B} 1_{\boldsymbol{x} \in A_n^{(b)}(\boldsymbol{z})}.$$

Note that the ad hoc kernel depends on L, K, and B which you will need to tune/adjust.

Implement the two kernel methods and conduct a simulation study comparing local linear models fit using these kernel functions when the black box model is random forests and boosting (implemented in xgboost).