STA 561: Homeworks 5 (Due April 1 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. (Some simple calculations.) Consider a binary classification problem with features $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \{-1, 1\}$. Show that the misclassification rate of the Bayes error is $\mathbb{E} \min \{q(\boldsymbol{X}), 1 - q(\boldsymbol{X})\}$, where $q(\boldsymbol{x}) \triangleq P(Y = 1 | \boldsymbol{X} = \boldsymbol{x})$.

2. (An unusual example.) Suppose we have data of the form $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ drawn i.i.d. such that $P(\boldsymbol{X} = 1) = P(\boldsymbol{X} = -1) = 1/4$ and $P(\boldsymbol{X} = 0) = 1/2$ while $Y = \text{sign}(|\boldsymbol{X}| - 1/2)$. Suppose we estimate a linear classifier using least squares, i.e., $\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n (Y - \beta_0 - \beta_1 \boldsymbol{X})^2$.

   - Derive $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^\intercal$ in this example.
   - Derive the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)$.
   - Describe the behavior of the test error $P 1_{Y \boldsymbol{X}^\intercal \widehat{\boldsymbol{\beta}}_n < 0}$ as n grows.