STA 561: Homework 4.1 (Due March 8 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. (Streaming Kaplan-Meier) Suppose that you have data $\{(T_i, \delta_i)\}_{i=1}^n$ where T denotes the observation time and δ is a censoring indicator. Recall that the Kaplan-Meier estimator is of the surival function is given by

$$\widehat{S}_n(t) = \prod_{j\,:\,t_j \le t} \left(1 - \frac{d_j}{n_j}\right),$$

where j indexes the unique observed event times (i.e., $j \in \{1, \ldots, n\}$ such that $\delta_j = 0$), $n_j = \sum_{i=1}^n 1_{T_i \ge t_j}$ are the number of subjects still in the study at time t_j, and $d_j = \sum_{i=1}^n 1_{T_i = t_j}$ are the number who died at time t_j. Propose and evaluate a streaming version of the Kaplan-Meier estimator. What challenges did you encounter? (Note, you don't need to know anything about survival analysis to complete this problem. It's just a nice example of a streaming problem with tricky accounting!)

2. (Lasso again!) Suppose we observe $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ comprising n i.i.d. copies of $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^p$ are the inputs and $Y \in \mathbb{R}$ are the outputs. Our goal, as usual, is to predict the inputs from the outputs. Let $\circ$ denote elementwise product, i.e., for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ we have $\mathbf{u} \circ \mathbf{v} = (u_1 * v_1, u_2 * v_2, \ldots, u_p * v_p)$. Let $\lambda > 0$ and define

$$(\widehat{\mathbf{u}}_n, \widehat{\mathbf{v}}_n) = \arg\min_{(\mathbf{u},\mathbf{v})} \mathbb{P}_n \left\{Y - (\mathbf{u} \circ \mathbf{v})^\intercal \boldsymbol{\beta}\right\}^2 + \lambda ||\mathbf{u}||^2/2 + \lambda ||\mathbf{v}||^2/2,$$

show that $\widehat{\beta}_n^\lambda = \widehat{\mathbf{u}}_n \circ \widehat{\mathbf{v}}_n$ is the lasso estiator with tuning parameter λ. (This is challenging problem, try and get as far as you can.) Show how, using this representation, you can implement lasso using a series of alternating ridge regressions.