Assignment 4

Yang Can, HKUST

Problem 1

Consider a linear model that relates variables $X_1, \ldots, X_p$ to the response Y:

$$Y = \sum_{j=1}^{p} X_j \beta_j + \epsilon,$$

where β_js are random effects, and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

Assume the prior $\beta_j \sim \mathcal{N}(0, \sigma_\beta^2), j = 1, \ldots, p$. Obtain the posterior of $\boldsymbol{\beta}$ using the mean-field approximation $q(\boldsymbol{\beta}) = \prod_{j=1}^{p} q(\beta_j)$, where the posterior mean and posterior variance denoted as $\boldsymbol{\mu}_{\text{mf}} \in \mathbb{R}^p$ and $\mathbf{S}_{\text{mf}}$ (a p-by-p diagonal matrix). Then compare the obtained mean-field approximation with the exact posterior distribution (e.g., obtained by the standard EM algorithm). You should check whether the approximated posterior mean is accurate and whether the posterior variance is underestimated. Please demonstrate your conclusion using simulation.

Problem 2

For a K-class classification problem, we can recode the class label c with a K-dimensional vector $\mathbf{y}$ with all entries equal to $-\frac{1}{K-1}$ except a 1 in position k if $c = k$, i.e.,

$$y_k = \begin{cases} 1, & \text{if } c = k, \\ -\frac{1}{K-1}, & \text{if } c \neq k. \end{cases}$$

Let $\mathbf{f} = (f_1(\mathbf{x}), \ldots, f_K(\mathbf{x}))^T$ with $\sum_{k=1}^{K} f_k(\mathbf{x}) = 0$, and define

$$L(\mathbf{y}, \mathbf{f}(\mathbf{x})) = \exp(-\frac{1}{K}\mathbf{y}^T \mathbf{f}(\mathbf{x})).$$

(a) Using Lagrange multipliers, derive the population minimizer $\mathbf{f}^*$ of $\mathbb{E}_{\mathbf{y}|\mathbf{x}}[L(\mathbf{y}, \mathbf{f}(\mathbf{x}))]$, subject to $\sum_{k=1}^{K} f_k(\mathbf{x}) = 0$, and relate these to the class probabilities.

(b) Derive a multiclass boosting algorithm using this loss function and verify that it covers the Adaboost algorithm as a special case ($K = 2$).

(c) Implement your derived algorithm, where you are allowed to call package of trees. Compare your implementation with the existing standard gradient boosting package on a multiclass classification problem. Make some discussion about what you observed.

Problem 3

We have a data set of $\mathcal{D} = \{\hat{\Gamma}_j, \hat{\gamma}_j, s^2_{Y,j}, s^2_{X,j}\}_{j=1}^M$ given in *data.txt*. Suppose the observed data is generated by the following model:

$$\begin{aligned}
&\hat{\Gamma}_j \sim \mathcal{N}(\hat{\Gamma}_j|\Gamma_j, s^2_{Y,j}),\\
&\hat{\gamma}_j \sim \mathcal{N}(\hat{\gamma}_j|\gamma_j, s^2_{X,j}),\\
&\Gamma_j = \beta\gamma_j,\\
&j = 1 \cdots M,
\end{aligned}$$

where β is the parameter of interest, and γ_j is a latent variable.

(a) Assuming that $\gamma_j \sim \mathcal{N}(\gamma_j|0, \sigma^2)$, please develop an algorithm to estimate β and perform statistical test to exam whether β is zero.

(b) Suppose the underlying true distribution of γ_j is given as $\gamma_j \sim q\delta(\gamma_j) + (1-q)\mathcal{N}(\gamma_j|0, \sigma^2)$, i.e, γ_j is either at 0 with probability q or distributed as $\mathcal{N}(\gamma_j|0, \sigma^2)$ probability $1-q$, where $\delta(.)$ is the Dirac delta function. However, you still use the algorithm developed in (a) to perform statistical inference on β. Will it give you inflated type I errors? You can perform some simulations first and then make some discussion based on your observation.

Problem 4

Consider the Lasso problem

$$\min_{\beta_1,\ldots,\beta_p} \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{1}$$

and denote its solution path as $\hat{\boldsymbol{\beta}}(\lambda)$. Suppose we are also interested in solving the following problem

$$\min_{u_1,\ldots,u_p;v_1,\ldots,v_j} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}u_jv_j)^2 + \lambda\sum_{j=1}^{p}(u_j^2 + v_j^2), \tag{2}$$

and denote its solution path as $(\hat{\mathbf{u}}(\lambda), \hat{\mathbf{v}}(\lambda))$. Let us define $\tilde{\beta}_j(\lambda) = \hat{u}_j(\lambda)\hat{v}_j(\lambda)$ for $j = 1, \ldots, p$. Please compare $\tilde{\boldsymbol{\beta}}(\lambda)$ with the well known Lasso solution path $\hat{\boldsymbol{\beta}}(\lambda)$. Make some discussion based on your observation. (Hint: You can obtain the solution paths by developing your own algorithm or calling some existing packages. Please make sure you know algorithm in the package well).

Requirement

- You need to submit a report, in which you should clearly describe your method and explain your idea. The code should also be included.
- You can use R or Python for coding.
- Your report should be in the **pdf** or **html** format, which is automatically generated by either R markdown or Jupyter notebook.
- The report is due to Dec, 1, 23:59 pm, 2021 (HK time).