Assignment 2

(*Note:* 1. Each problem needs to be coded in R only.)

(*Note:* 2. Do not use the libraries available for the methods (e.g: ridge regression, lasso, LDA). You are expected to code them. However, you are free to use other libraries that are not associated with methods (e.g: optimization). Also, you can use *glmnet* package for logistic regression.)

1. (a) Construct a data set for regression as follows. Generate $(X_1, X_2, \ldots, X_{100}) \in \mathbb{R}^{20}$ such that they are i.i.d. $\sim \mathcal{N}([0, 0, \ldots, 0], I_{20})$. Generate four numbers $i_1, i_2, i_3, i_4 \in \{1, 2, \ldots, 20\}$ uniformly at random without replacement. Also generate four numbers $a, b, c, d \sim \mathcal{N}(0, 0.25)$. Generate $(Y_1, Y_2, \ldots, Y_{100}) \in \mathbb{R}$ as given below:

$$Y_k = aX_{ki_1} + bX_{ki_2} + cX_{ki_3} + dX_{ki_4} + n_k,$$

where $n_k \sim \mathcal{N}(0, 0.01)$ for $k = 1, 2, \ldots, 100$. **Print** i_1, i_2, i_3, i_4.

   (b) Compute and **print** the coefficient vector $\hat{\beta}^{ls}$ using linear regression.

   (c) Consider the *forward selection* method. Compute the best four (out of 20) parameters that minimizes the residual sum of squares. **Print** both the best set of parameters and the coefficient vector corresponding to the best set.

   (d) Consider *ridge regression*. Center the data, fix $\lambda = 0.01$, and compute the coefficient vector $\hat{\beta}^{ridge}$ corresponding to this value of λ. **Print** $\hat{\beta}^{ridge}$. Sort the coefficients $\hat{\beta}^{ridge}$ in decreasing order of their absolute value, and then **print the indices** that correspond to the five highest coefficients.

   (e) Consider *lasso method*. Repeat what you did for ridge regression.

2. (a) Construct the data set for a 3-class classification as follows. Generate $(X_1, X_2, \ldots, X_{50}) \in \mathbb{R}^2$ such that they are i.i.d. $\sim \mathcal{N}([0, 0], I_2)$. Label them all in bin 1. Generate $(X_{51}, X_{52}, \ldots, X_{100}) \in \mathbb{R}^2$ such that they are i.i.d. $\sim \mathcal{N}([2, 0], I_2)$. Label them all in bin 2. Generate $(X_{101}, X_{102}, \ldots, X_{150}) \in \mathbb{R}^2$ such that they are i.i.d. $\sim \mathcal{N}([1, \sqrt{3}], I_2)$. Label them all in bin 3. **Plot** the generated features in the form of a scatterplot. Represent the features having different labels with different colors.

   (b) Compute the classifier using linear classifier with indicator matrices. **Plot** the classifying lines along with the scatter plot of the generated features.

(c) Repeat the same using LDA.

(d) Repeat the same using logistic regression. Use glmnet package for logistic regression.