ESI 4606 Analytics I - Foundations of Data Science
Homework 6
Due: November 23st (11:00AM), 2022

Problem 1 (1.5 points)

Questions in this problem should be answered using the same data set file of "HM5.txt". $x_1, x_2, ..., x_{10}$ are predictors and y is the response variable.

(a) Perform the ridge regression with λ selected based on 10-fold cross-validation (Set seed value as 2 for cross-validation). Report the coefficients of the model obtained.

(b) Perform the LASSO regression with λ selected based on 10-fold cross-validation (Set seed value as 2 for cross-validation). Report the coefficients of the model obtained.

Note: (1) $x_1, ..., x_{10}$ have already been standardized.

(2) use "data.matrix()" to transform data into matrix format before running ridge regression or lasso regression. To select λ based on 10-fold cross-validation, use "lambda.1se" as the best λ.

(3) **To get full points, include R codes in the appendix sections**

Problem 2 (2 points)

Given the first and second principle component loading vectors $\boldsymbol{\phi}_1 = [\phi_{11}, \phi_{21}]^{\mathrm{T}}$ and $\boldsymbol{\phi}_2 = [\phi_{12}, \phi_{22}]^{\mathrm{T}}$, where $\phi_{11} = -\frac{\sqrt{2}}{2}$, $\phi_{21} = -\frac{\sqrt{2}}{2}$, $\phi_{12} = \frac{\sqrt{2}}{2}$ and $\phi_{22} = -\frac{\sqrt{2}}{2}$.

(a) If two observations in the original data set are $(x_{11} = 1, x_{12} = 1)$ and $(x_{21} = -1, x_{22} = -1)$, compute the values of the corresponding transformed observations, namely (z_{11}, z_{12}) and (z_{21}, z_{22}), in the transformed data set (whose feature space is characterized by the first two principle components).

(b) If two observations in the transformed data set (whose feature space is characterized by the first two principle components) are $(z_{31} = 1, z_{32} = 1)$ and $(z_{41} = -1, z_{42} = -1)$, compute the values of the corresponding original observations, namely (x_{31}, x_{32}) and (x_{41}, x_{42}), in the original data set .

Problem 3 (1.5 points)

Consider the "iris" data set in library "datasets". The iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for different species of iris. Perform PCA analysis on the first 4 columns of the data set. Data should be standardized before implementing PCA.

(a) Output the first and second principle component loading vectors.

(b) Draw the biplot.

(c) Output the proportion of variance explained (PVE) by each principle component and draw the scree plot.

Note: **To get full points, include R codes in the appendix sections**