Question 1: (20 points)

Part 1: Consider the Lasso objective function:

$$\frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2+\lambda\|\boldsymbol{w}\|_1$$

Drive an expression for λ_{max} such that for any $\lambda \geq \lambda_{max}$, the estimated weight vector $\widehat{w}$ is entirely zero.

Part 2: Show that the elastic net regression estimates can be obtained by lasso regression on an augmented data set defined as:

$$\left\|\widehat{\boldsymbol{y}}-\widehat{\boldsymbol{X}}\widehat{\boldsymbol{w}}\right\|_2^2+\gamma\|\widehat{\boldsymbol{w}}\|_1$$

Where $\widehat{\boldsymbol{y}}=\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}_{p\times 1}\end{bmatrix}$; $\widehat{\boldsymbol{X}}=\frac{1}{\sqrt{1+\lambda_2}}\begin{bmatrix}\boldsymbol{X}\\ \sqrt{\lambda_2}\,\boldsymbol{I}_{p\times p}\end{bmatrix}$; $\gamma=\frac{\lambda_1}{\sqrt{1+\lambda_2}}$; and $\widehat{\boldsymbol{w}}=\sqrt{1+\lambda_2}\,\boldsymbol{w}$.

Question 2: Multiple response multivariate regression (40 points)

Forecasts of maximum and minimum air temperatures are essential to mitigate the damage of extreme weather events such as heat waves and tropical nights. The study area is a metropolitan city, Seoul (~605 km^2) with over 10 million people. Seoul is geographically surrounded by four distinct mountains and is divided into the northern and southern parts by the Han River. It is hot and humid due to the East Asian monsoon in summer, so there are a lot of days with hot weather above 30 °C, and precipitation is concentrated in summer (i.e., seasonal rainfall is 892.1 mm in summer and 67.3 mm in winter). The data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the Local Data Assimilation and Prediction System (LDAPS) model operated by the Korea Meteorological Administration. This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. "Question2.csv" contains the data set. Use the first 5000 samples for training and the rest for the test. There are two outputs: next-day maximum and minimum air temperatures which are the last two columns of the data set. To predict the next-day maximum and minimum air temperatures, we are going to use the following models:

1. Ridge Regression
2. Lasso Regression
3. Adaptive Lasso Regression
4. Elastic Net Regression

- Report optimal tuning parameters obtained using cross-validation.
- Report the coefficients obtained with the optimal parameters.

- Report the Mean Square Prediction Error on the test set.

Note that you should standardized the data (zero mean and standard deviation at 1).

Question 3: Group lasso (40 points)

Part 1: In this part, we want to use group lasso to predict the rise time of a servomechanism in terms of two gain settings and two (discrete) choices of mechanical linkages.
"Question3Part1.csv" contains 167 data samples. Use the first 100 samples to build a group lasso regression model; and the rest to evaluate the accuracy of your model.

- Report λ_{min}
- Report your coefficients.
- Report the most informative feature; and uninformative feature(s).
- Report MSE on the test set.

Attribute Information:
1. motor: A, B, C, D, E
2. screw: A, B, C, D, E
3. pgain: 3, 4, 5, 6
4. vgain: 1, 2, 3, 4, 5
5. rise time: 0.13 to 7.10

Part 2: In this part, we want to build a logistic group lasso classifier to predict colon cancer.
"Question3Part2.csv" data set contains 72 samples with 100 predictors (expanded from 20 genes using 5 basis B-splines). Use the first 50 samples to build a logistic group lasso classifier and the rest to measure the accuracy of your classifier.

- Report λ_{min}
- Report your coefficients.
- Report the accuracy of your classifier on the test samples.