ISyE 8803 –Topics on High Dimensional Data Analytics

Exam II

- You are not allowed to discuss the exam content with your fellow students nor receive aid on this exam.
- You are expected to observe the Georgia Tech Honor Code throughout the exam.
- Exam is due on July 31 at 11:59pm (U.S. Eastern Time). Late submission is NOT accepted. Submit your solutions via Canvas.
- Submit your exam answers in PDF format. For problems that require programming, supply your codes in separate files.

Question 1

In this problem, you build a set of different models to predict liver disease by analyzing laboratory values of blood donors. You can find the training and test datasets in "Q1train.csv" and "Q1test.csv". The last column of the datasets contains the response variable which can take the values "1" (liver disease) or "0" (normal). In order to predict liver disease, build the following models:

- Logistic regression
- Ridge logistic Regression
- Lasso logistic Regression
- Adaptive Lasso logistic Regression (Gamma=1)

For each of the models perform the following tasks:

- Fit the model on the training dataset.
- Report optimal tuning parameters obtained using cross-validation.
- Report the coefficients obtained with the optimal parameters.
- Report the classification accuracy for the test set.
- Report the confusion matrix for the test set.

Question 2

The dataset "ratings.csv" contains ratings, from 1 to 5, for 300 different restaurants. In this question, you will be performing experiments on this data set to determine how missingness affects the performance of matrix completion.

Part 1 – Missing Completely at Random

Randomly select values in the ratings matrix to label as missing. Do this by generating random variates from a binomial distribution where the probability of "success" is set to 5%. At the selected locations set the entries to a value of zero. Then solve the following optimization problem:

$$\min_{\mathrm{M}} \|M\|_*$$

$$\text{subject to} \quad \mathrm{M(i,j)} = \mathrm{M}_0(i,j) \quad \forall (i,j) \in \{known\ set\}$$

Make sure to calculate the relative reconstruction error:

$$\frac{\|M - M_0\|_F}{\|M_0\|_F}$$

Repeat the optimization problem by increasing the percentage of missing data by 5% up to 50%. Plot the reconstruction error vs the percentage of missing data. Additionally, for the case where 50% of the data is missing, report the frequency of each value in the matrix with missing values, the recovered matrix, and the original full dataset.

Part 2 – Missing Not at Random

To simulate a case where customers avoid going to poorly rated restaurants, generate missing entries where ratings of value 1 or 2 are three times more likely to be missing than ratings of 3 or 4. Ratings of 5 will have a missing rate of 0%. To do so, use random variates from a binomial distribution, where the probability of "success" is adjusted for each rating value such that the overall missing rate is 5%.

Solve the same optimization problem as in Part 1 and repeat the experiment by increasing the percentage of missing data by 5% up to 50%. For the last case, report the frequency of each value in the matrix with missing values and the recovered matrix.

Finally, generate a plot of reconstruction error vs percentage of missing data. Compare with the results in Part 1.

Question 3

Remote sensing images provide useful information about land cover. However, these images can be contaminated by clouds, shadow, or other noise. Robust principal component analysis (PCA) has obtained stunning performance in removing noise from images. It assumes such data are composed of a low-rank component and a sparse component. Mathematically, given an image matrix, M, $M \in R^{m \times n}$, M can be decomposed into a low-rank matrix L and a sparse matrix S, which is estimated by minimizing the following constrained optimization problem:

$$min_{(L,S)} \ \|L\|_* + \lambda||S||_1, \text{ s.t. } M = L + S$$

In this question, we want to remove the noise from a land cover image. The image called 'image.jpg' is provided. The image has some noise, such as English words and numbers, which should be removed before further processing.

1. Read the satellite image and convert it into a grayscale image. Present the color image and the grayscale image.
2. Implement Robust PCA on the grayscale image to obtain the background image (low-rank matrix), and the noise image (sparse matrix). Present both images. The noise should be separated from in the original image. In the background image, the noise should be

blurred or invisible, while the edges of lands are still visible. Please do not use built-in Robust PCA function directly.

3. Do edge detection on the grayscale image and the background image, then calculate the mean squared errors (MSE) and correlation between two edge detected images. Present the two edge-detected images and report the MSE and correlation coefficient (the coefficient should be above 0.5). You may use built-in functions for Canny algorithm or other algorithms to do edge detection.