ISyE 8803 – Topics on High Dimensional Data Analytics

Exam II

- For all questions, you are required to clearly state all assumptions you make and show all necessary details of your solutions.
- You are not allowed to discuss the exam content with your fellow students or receive aid on this exam.
- You are expected to observe the Georgia Tech Honor Code throughout the exam.
- Exam is due on Aug 1 at 11:59 pm. Late submission is NOT accepted. Please submit your solutions via Canvas.

Question 1. Regularization (25 points)

In this problem, you build a set of different models to classify 3 different beans. A total of 16 features were obtained from the grains. "Question1.csv" contains the dataset. There are 3871 observations. Use the first 2800 data samples as your training set; and the rest (from 2801 to 3871) as test set.

Attribute Information:

1) Area (A): The area of a bean zone and the number of pixels within its boundaries.
2) Perimeter (P): Bean circumference is defined as the length of its border.
3) Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.
4) Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
5) Aspect ratio (K): Defines the relationship between L and l.
6) Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.
7) Convex area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
8) Equivalent diameter (Ed): The diameter of a circle having the same area as a bean seed area.
9) Extent (Ex): The ratio of the pixels in the bounding box to the bean area.
10) Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
11) Roundness (R): Calculated with the following formula: (4piA)/(P^2)
12) Compactness (CO): Measures the roundness of an object: Ed/L
13) ShapeFactor1 (SF1)
14) ShapeFactor2 (SF2)
15) ShapeFactor3 (SF3)
16) ShapeFactor4 (SF4)

The last column represents the classes:

17) Class 1 to 3 which correspond to 'Barbunya', 'Bombay', 'Seker'.

1- Run a multinomial logistic regression to classify different beans. Present the coefficients obtained. Present the confusion matrix on the test set.
2- Use ridge multinomial logistic regression to classify different beans. Explain why we use ridge regression. Present the optimal tuning parameter obtained using cross-validation, the coefficients for this parameter and the confusion matrix on the test set.
3- Use lasso multinomial logistic regression to classify different beans. Explain why we use lasso. Present the optimal tuning parameter obtained using cross-validation, the coefficients for this parameter and the confusion matrix on the test set.

4- Use adaptive lasso multinomial logistic regression to classify different beans. Consider $\widehat{w}_j = \frac{1}{\left(\widehat{\beta}_j^{ridge}\right)^{0.5}}$. Explain why we use adaptive lasso regression. Present the optimal tuning parameter obtained using cross-validation, the coefficients for this parameter and the confusion matrix on the test set.
5- Which model leads to the sparsest coefficients? Which model has the highest classification accuracy?

Question 2. Group Lasso (25 points)

Predicting students' final grade in mathematics – Dataset student.csv consists of various information about 347 high school students. Description of the features can be found in student.txt

(a) Create dummy variables for all of the features except "absences", "G1", "G2", and "G3". Remove original variables for which you created dummy variables. Split data into 80% for training and 20% for testing.

(b) Fit lasso regression using training set to predict "G3" based on the other features. Use cross validation with 10 folds to select optimal λ value. Report MSE value on the test set.

(c) Group dummy variables obtained from the same original variable. For example, original variable "guardian" can have three values "mother", "father", and "other". We can code it using variables:

$$\text{"guardian\_mother"} = \begin{cases} 1 & \text{"guardian"} = \text{"mother"} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{"guardian\_father"} = \begin{cases} 1 & \text{"guardian"} = \text{"father"} \\ 0 & \text{otherwise} \end{cases}$$

Those two variables belong to the same group. Fit group lasso regression using training set to predict "G3" based on the other variables. Use cross validation with 10 folds to select optimal λ value. Report MSE value on the test set.

(d) Try grouping variables in a different manner. Fit group lasso regression with grouping of your choice using training set to predict "G3" based on the other variables. Use cross validation with 10 folds to select optimal λ value. Report MSE value on the test set.

(e) Comment on differences (or similarities) between groups of variables which are selected by models from (b), (c), and (d)

Question 3. Robust PCA (25 points)

In this problem, you will be implementing Robust PCA to study decomposition of abstract art.

a) (5 pts) Write the mathematical formulations of Robust PCA (both before and after convex relaxation).
b) (5 pts) Write the update steps yielded from implementing Augmented Lagrangian Multipliers Method.
c) (10 pts) Implement Robust PCA on *image.jpg* and plot the images created from the low-rank matrix and the matrix of sparse outliers.
d) (5 pts) Qualitatively describe your final image outputs. What features are captured in the low-rank matrix and the matrix of sparse outliers?

Question 4. Matrix completion (25 points)

Matrix completion is the task of filling in the missing entries of a partially observed matrix. In this question, we will practice matrix completion to fill the missing values in two datasets, namely, "NonRandomMiss.xlsx' and "RandomMiss.xlsx". In both datasets, there are 1500 missing values, which are replaced with zeros. In "NonRandomMiss.xlsx', the values of one of every six columns are missing. In "RandomMiss.xlsx", the missing values are randomly distributed.

a) Write the PFBS algorithm (see Slide p21-22 in Module 5) to fill the missing values both datasets. Report the mean, standard deviation, maximum and minimum of the filled values.
b) The original data are stored in "Original.xlsx". Please use it as a reference to compute the recovery errors for both datasets, report the recovery errors for the missing values and the whole datasets. What do you observe from these errors?
c) Plot the datasets (the original and the two recovered images). Which recovered image looks closer to the original?