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 December 12, 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: (25 points)

Solve the following optimization problem using the scaled form of alternating direction method of multipliers (ADMM).

$$\min \frac{1}{2}\|Ax-b\|_2^2+\lambda_1\|z\|_1+\frac{\lambda_2}{2}\|z\|_2^2 \qquad \text{s.t.} \qquad x-z=0$$

a) Write the augmented Lagrangian function (the scaled form) and drive the ADMM updates (Show your work).

b) In this part, you implement your own regression algorithm using your solution in part (a) to predict the performance decay over time of the Gas Turbine (GT) compressor. The range of decay of compressor has been sampled with a uniform grid of precision 0.001. The compressor decay coefficient is in the range of [0.95,1]. The dataset is provided as "Question1.csv". The last column of the datasets corresponds to the output we want to predict.

- The 13 features are:
  Lever position (lp)
  Ship speed (v) [knots]
  Gas Turbine (GT) shaft torque (GTT) [kN m]
  GT rate of revolutions (GTn) [rpm]
  Gas Generator rate of revolutions (GGn) [rpm]
  Port Propeller Torque (Tp) [kN]
  Hight Pressure (HP) Turbine exit temperature (T48) [C]
  GT Compressor outlet air temperature (T2) [C]
  HP Turbine exit pressure (P48) [bar]
  GT Compressor outlet air pressure (P2) [bar]
  GT exhaust gas pressure (Pexh) [bar]
  Turbine Injection Control (TIC) [%]
  Fuel flow (mf) [kg/s]
  GT Compressor decay state coefficient
  Consider $\rho = 1$, $\lambda_1 = 0.1$ and $\lambda_2 = 0.9$.
- Use the first 2000 samples to learn x. Report your coefficients (x). Plot objective function $\left(\frac{1}{2}\|Ax-b\|_2^2+\lambda_1\|z\|_1+\frac{\lambda_2}{2}\|z\|_2^2\right)$ versus iterations.

- Report your Sum of Absolute Errors on the test set (data samples 2001-2387).

Question 2: (25 points)

In this problem, you build a set of different models for predicting liver disease by analyzing laboratory values of blood donors. "Question2.csv" contains the dataset. Use the first 250 samples for training and the rest as test set. The last column of the dataset represents the labels. There are two categories: "1" indicates the presence of liver disease, "0" indicates normal.

a) Build a logistic regression model. Present the coefficients.
   - Present the confusion matrix for the test set.

b) Build a Ridge logistic regression model.
   - Present the optimal tuning parameter, and the coefficients.
   - Present the confusion matrix for the test set.

c) Build a lasso logistic regression model. Present the optimal tuning parameter and your model coefficients.
   - Present the confusion matrix for the test set.

d) Build an adaptive lasso logistic regression model. Present the optimal tuning parameter and your model coefficients.
   - Present the confusion matrix for the test set.

Question 3: (25 points)

Solve the following optimization problem using proximal gradient method:

$$\min_{X} \ \lambda \|X\|_* + \frac{1}{2}\|Y_\Omega - P_\Omega(X)\|_2^2$$

1- Drive a closed form solution for matrix X. Write a pseudo code for your algorithm.

Complete Algorithm: .

Input: observation samples $Y_{ij}, (i,j) \in \Omega$ of matrix Y$\in \mathbb{R}^{m\times n}$

Initialize: $X_0 = 0, \lambda = 10,$ gradient step size $t = 1$

Output: $X \in \mathbb{R}^{m\times n}$

2- "Question3.jpg" contains image Y and "Question3.mat" contains Ω. Code your algorithm and display output image X. Plot objective function versus iteration.

Question 4: (25 points)

In this question, you use kernel ridge regression to perform energy efficiency analysis. "Question4.csv" contains 768 data samples. The first 8 columns are the features, and the last column is the response variable. The first 600 data samples are for training and the remaining data samples are for the test. In this question, we use Gaussian kernel $k(x,y) = e^{\left(-\frac{\|x-y\|_2^2}{h}\right)}$.

a) First, set the parameter $\lambda = 0.01$ and find your optimal kernel bandwidth (h) using your training data - you should not allow the test dataset to influence your choice of h. Randomly

split your training data into 80% training and 20% validation to tune h. Report the optimal kernel bandwidth (h) and the corresponding MSE.

b) You need to set the free parameter λ using your training data, given the optimal value of h obtained in part (a). Report your optimal λ and the corresponding MSE.

c) Report the performance of your algorithms in terms of mean-squared error on the test set.