**ISyE 8803: Topics on High Dimensional Data Analytics**

**Homework 3**

**Question 1 (30 points)**

Given the following optimization problems in matrix form:

*Ordinary Least Squares*:

*Ridge*:

*Lasso*:

*Elastic Net*:

where

* is the vector of observed outputs, n is the number of observations
* is the matrix of features, p is the number of features
* is the vector of parameters to be estimated
* , where is a hyperparameter controlling the shrinkage penalty

Assuming X is *orthonormal*,

1. Show that is a closed form solution for the Ordinary Least Squares regression problem.
2. Show that is a closed form solution for the Ridge regression problem.
3. Show that…

… is a closed form solution for the Lasso regression problem.

1. Using the derivations in parts a, b, and c, derive a closed form solution for the Elastic Net regression problem, where is the Lasso shrinkage parameter and is the Ridge shrinkage parameter.

Hints:

* When is orthonormal, (where is a identity matrix)
* For part (c), show that the optimization problem for Lasso can be reduced to :
* For part (c) Lasso, consider the cases  , , and as well as those in the given solution. They may help show why the inequalities are what they are.

**Question 2 (35 points)**

The ovarian cancer data set (“grp.csv” and “obs.csv” files) consists of gene data for 216 patients, 121 of whom have ovarian cancer, and 95 of whom do not. For each patient, there is a vector of data containing the expression of 4000 genes. The gene data is highly correlated, so that many patients have significant overlap in their gene expression. Please follow the below instructions to reduce the dimensionality of this problem:

a) Dimensionality reduction with Bspline and Lasso

* First, split the data set randomly (80% training set and 20% test set).
* Use B-splines with 8 knots to reduce the dimensionality of the problem.
* Use lasso to learn the B-spline coefficients and report the values of coefficients. Note that Y values are “Cancer” or “Normal”, so the response type is binomial. Plot the coefficient values against lambdas.
* Predict the probability of having cancer for the observations in the test dataset. Consider the threshold for the probability of having cancer and then compute the Mean Square Prediction Error.

b) Dimensionality reduction with PCA and Lasso

* Now find the principal component of 4000 genes for both training and test set. Do not forget to scale data before PCA computation. Consider the 10 eigenvectors correspond to the 10 largest eigenvalues.
* Use lasso to learn the PCA coefficients and report the values of coefficients. Plot the coefficient values against lambdas.
* Predict the probability of having cancer for the observations in the test dataset. Consider the same threshold for the probability similar to previous part and then compute the Mean Square Prediction Error. Compare the accuracy of part A and B.

**Question 3 (35 points)**

The goal of this problem is to predict the performance decay over time of the Gas Turbine (GT) compressor. The dataset contains 11934 observations. The range of decay of compressor has been sampled with a uniform grid of precision 0.001 so to have a good granularity of representation. For the compressor decay state discretization the kMc coefficient has been investigated in the domain [0.95,1]. Ship speed has been investigated sampling the range of feasible speed from 3 knots to 27 knots with a granularity of representation equal to tree knots. A series of measures (13 features) which indirectly represents of the state of the system subject to performance decay has been acquired and stored in the dataset over the parameter’s space.

For each record it is provided:

- A 13-feature vector containing the GT measures at steady state of the physical asset:

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 Injecton Control (TIC) [%]

Fuel flow (mf) [kg/s]

- GT Compressor decay state coefficient

The dataset is provided as "Shiptrain.csv" and "Shiptest.csv". The last column of the datasets correspond to the output we want to predict. We have split the dataset into training (80%) "Shiptrain.csv" and test (20%) "Shiptest.csv" sets. In order to predict the performance decay over time of the GT compressor and turbines, we are going to use the following models:

1. Ridge Regression

2. Lasso Regression

3. Adaptive Lasso Regression

4. Elastic Net Regression

For each of the models please do the following:

* Fit the model on the training dataset.
* Report optimal tuning parameters obtained using cross-validation.

Note: You must tune the lambda parameter for all models and the alpha parameter

for the elastic net regression model.

* Report the coefficients obtained with the optimal parameters.
* Report the Mean Square Prediction Error for the test set.

Note that you should standardized the data.

Conclusion: Which model will you select to predict the performance decay over time of the GT compressor and turbines? Why?