Topics on High-Dimensional Data Analytics

ISYE 8803 - HW4

Problem 1. Closed-form solutions with orthonormal predictors (30 points)

Given the following objective functions in matrix form:

$$OLS: \quad \min_{\beta} ||y - X\beta||_2^2 \tag{1}$$

$$Ridge: \quad \min_{\beta} ||y - x\beta||_2^2 + \lambda||\beta||_2^2 \tag{2}$$

$$Lasso: \quad \min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1 \tag{3}$$

Where,

- $y \in R^n$ are the observed outputs, n is the number of observations
- $X \in R^{n \times p}$ are the observed inputs, p is the number of inputs
- $\beta \in R^p$ are the parameters to the estimated
- $\lambda \geq 0$ is tuning parameter that controls the amount of shrinkage

Let X be orthonormal.

(a) (6 points) Show that the Ordinary Least Square regression problem has the following closed form solution: $\hat{\beta}^{ols} = X^T y$

(b) (6 points) Show that the Ridge regression problem has the following closed form solution:

$$\hat{\beta}^{ridge} = (1+\lambda)^{-1}\hat{\beta}^{ols}$$

(c) (18 points) Show that the Lasso regression problem has the following closed form solution:

$$\hat{\beta}_j^{lasso} \begin{cases} \hat{\beta}_j^{ols} - \frac{\lambda}{2} & \text{if } \beta_j^{ols} > \frac{\lambda}{2} \\ 0 & \text{if } |\beta_j^{ols}| \leq \frac{\lambda}{2} \\ \hat{\beta}_j^{ols} + \frac{\lambda}{2} & \text{if } \beta_j^{ols} < -\frac{\lambda}{2} \end{cases}$$

Hints:

- Show that the minimization problem in equation (3) can be reduced to:

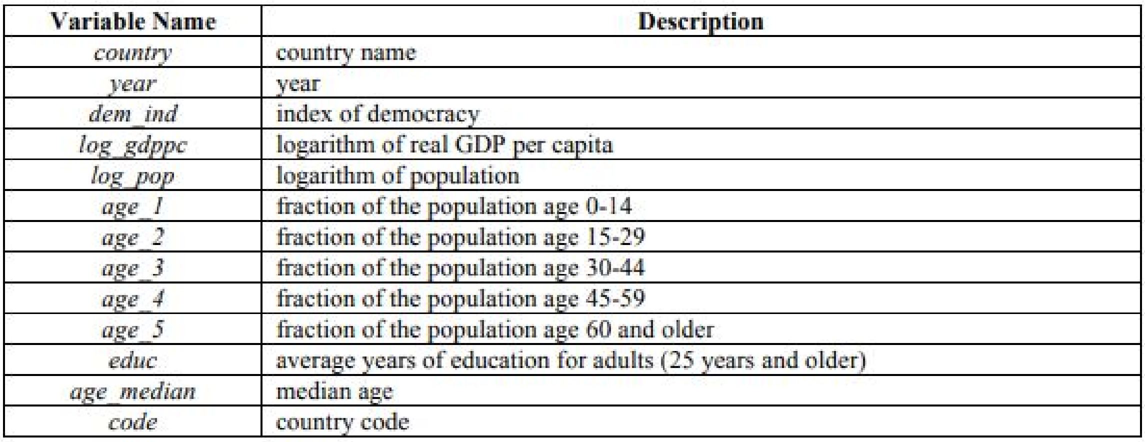
$$\min_{\beta_j} \beta_j^2 - 2\hat{\beta}_j^{ols}\beta_j + \lambda|\beta_j|$$

- Find the optimal value for $\hat{\beta}_j^{lasso}$, by considering the following cases:

  1. $\hat{\beta}_j^{ols} > 0$
     (a) $\hat{\beta}_j^{ols} \leq \frac{\lambda}{2}$
     (b) $\hat{\beta}_j^{ols} > \frac{\lambda}{2}$
  2. $\hat{\beta}_j^{ols} = 0$
  3. $\hat{\beta}_j^{ols} < 0$
     (a) $\hat{\beta}_j^{ols} \geq -\frac{\lambda}{2}$
     (b) $\hat{\beta}_j^{ols} < -\frac{\lambda}{2}$

Problem 2. Regularized Regression (45 points)

One of the most extensively researched question in political economy is the relationship between income per capita and democracy. File "income_democracy.csv" contains a panel data set for 195 countries for the years 1960, 1965, ... 2000. The data were supplied by Professor Daron Acemoglu and are a subset of the data used in his paper with Simon Johnson, James Robinson, and Pierre Yared, “Income and Democracy” American Economic Review. Dependent variable is democracy index and covariates include logarithm of real GDP per capita and some other demographic properties (columns 5 to 12). You must split the data set into training (80%) and test sets (20%) sets. Consider the pooled data, it means you are not accounting for each countrie’s fixed effect and remove the missing data.



Variable Name	Description
country	country name
year	year
dem_ind	index of democracy
log_gdppc	logarithm of real GDP per capita
log_pop	logarithm of population
age_1	fraction of the population age 0-14
age_2	fraction of the population age 15-29
age_3	fraction of the population age 30-44
age_4	fraction of the population age 45-59
age_5	fraction of the population age 60 and older
educ	average years of education for adults (25 years and older)
age_median	median age
code	country code

Your job is to predict democracy index by real GDP per capita and other demographic features. In order to predict the democracy index we are going to use the following models:

1. Ridge Regression (10 points)

2. Lasso Regression (10 points)

3. Adaptive Lasso Regression (10 points)

4. Elastic Net Regression (10 points)

For each of the models please do the following:

1. Fit the model on the training dataset.

2. 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.

3. Report the coefficients obtained with the optimal parameters.

4. Report the Mean Square Prediction Error for the test set.

Note that you should standardized the data. Conclusion: (5 points) Which model will you select to predict the democracy index? Why?

Problem 3. Functional Linear Regression - (25 points)

A combustion engine produces gas with polluting substances such as nitrogen oxides (NO_x). Gas emission control regulations have been set up to protect the environment. The NO_x Storage Catalyst (NSC) is an emission control system by which the exhaust gas is treated after the combustion process in two phases: adsorption and regeneration. During the regeneration phase, the engine control unit is programmed to maintain the combustion process in a rich air-to-fuel status. The average relative air/fuel ratio is the indicator of a correct regeneration phase. Our goal is to predict this value, using the information from ten sensors (Table 1). To do so, we are going to use group lasso regression. The data for this problem can be found as NSC.mat. Please proceed as follow:

(a) (4 points) Plot and present the observations for each sensor in the training data set.

(b) (6 points) Use B-splines with 8 knots to reduce the dimensionality of the problem.

(c) (4 points) Write the problem that we want to solve in mathematical notation. Clearly explain what your notation represents.

(d) (6 points) Use group lasso to learn the B-spline coecients. Which sensors are correlated with the air/fuel ratio?

(e) (5 points) Predict the air/fuel ratio for the observations in the test dataset, it can be found as NCS.test.mat. Present the Mean Square Prediction Error.

#	Description
1	air aspirated per cylinder
2	engine rotational speed
3	total quantity of fuel injected
4	low presure EGR valve
5	inner torque
6	accelerator pedal position
7	aperture ratio of inlet valve
8	downstreem intercooler preasure
9	fuel in the 2nd pre-injection
10	vehicle velocity