Topics on High-Dimensional Data Analytics
ISYE 8803 -
Homework 6

Problem 1. Closed form solution for Ridge Regression (5 points)
The Ridge objective function in matrix form is:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\beta}||_2^2 \tag{1}$$

where

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

Show that the Ridge regression problem has the following closed for solution:

$$\hat{\beta} = (\boldsymbol{X^T X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2}$$

where $\boldsymbol{I}_p$ is the identity matrix of dimension $p \times p$.

Hint: Differentiate the objective function with respect to $\boldsymbol{\beta}$ and solve the equation:

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left[\frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\beta}||_2^2\right] = 0. \tag{3}$$

Problem 2. Single predictor soft-thresholding derivation (25 points)
Consider a single predictor setting, based on samples $\{(x_i, y_i)\}_{i=1}^n$. Assume that the data has been standardized $\left(\text{i.e. } \frac{||\boldsymbol{x}||_2^2}{n} = 1\right)$. The problem is to minimize with respect to β the function:

$$\frac{1}{2n}\sum_{i=1}^n (y_i - x_i\beta)^2 + \lambda|\beta| \tag{4}$$

where $\lambda \geq 0$. The standard approach to this univariate minimization problem would be to take the gradient with respect to β, and set it to zero. There is a complication, however, because the absolute value of the function $|\beta|$ does not have a derivative at $\beta = 0$. However, we can proceed by direct inspection of the function in equation (4) and find

$$\hat{\beta} = S_\lambda\left(\frac{1}{n}\sum_{i=1}^n y_i x_i\right) = \begin{cases} \frac{1}{n}\sum_{i=1}^n y_i x_i - \lambda & \text{if } \frac{1}{n}\sum_{i=1}^n y_i x_i > \lambda \\ 0 & \text{if } \frac{1}{n}|\sum_{i=1}^n y_i x_i| \leq \lambda \\ \frac{1}{n}\sum_{i=1}^n y_i x_i + \lambda & \text{if } \frac{1}{n}\sum_{i=1}^n y_i x_i < -\lambda \end{cases}. \tag{5}$$

Please follow the steps:

a. (5 points) Show that the minimization problem in equation (4) can be written as:

$$\min_{\beta \in \mathbb{R}} \frac{1}{2}\beta^2 - \frac{1}{n}\beta \sum_{i=1}^{n} y_i x_i + \lambda|\beta|. \tag{6}$$

b. (5 points) Case 1: $\frac{1}{n}\sum_{i=1}^{n} y_i x_i > 0$. Under this scenario which one of the following options must be true:

- $\beta = 0$
- $\beta \leq 0$
- $\beta \geq 0$
- $\beta < 0$
- $\beta > 0$

Explain your reasoning.

(a) Case 1a: $\frac{1}{n}\sum_{i=1}^{n} y_i x_i \leq \lambda$. Under this scenario, find the optimal value for β. Explain your reasoning.

(b) Case 1b: $\frac{1}{n}\sum_{i=1}^{n} y_i x_i > \lambda$. Under this scenario, find the optimal value for β (Hint: think of the solution for the problem $\min_{x \in \mathbb{R}} ax^2 + bx + c$ where $a, b, c \in \mathbb{R}$ are constants).

c. (5 points) Case 2: $\frac{1}{n}\sum_{i=1}^{n} y_i x_i = 0$. Under this scenario which one of the following options must be true:

- $\beta = 0$
- $\beta \leq 0$
- $\beta \geq 0$
- $\beta < 0$
- $\beta > 0$

Explain your reasoning.

d. (5 points) Case 3: $\frac{1}{n}\sum_{i=1}^{n} y_i x_i < 0$. Under this scenario which one of the following options must be true:

- $\beta = 0$
- $\beta \leq 0$
- $\beta \geq 0$
- $\beta < 0$
- $\beta > 0$

Explain your reasoning.

(a) Case 3a: $\frac{1}{n}\sum_{i=1}^{n} y_i x_i \geq -\lambda$. Under this scenario, find the optimal value for β. Explain your reasoning.

(b) Case 3b: $\frac{1}{n}\sum_{i=1}^{n} y_i x_i < -\lambda$. Under this scenario, find the optimal value for β (Hint: think of the solution for the problem $\min_{x\in\mathbb{R}} ax^2 + bx + c$ where $a, b, c \in \mathbb{R}$ are constants).

e. (5 points) What is the optimal solution for the minimization problem in equation (4)?

Problem 3. Sparse linear regression (45 points)
In this problem the goal is to predict the concentration of carbon oxide (CO) in mg/m^3. For this purpose, we have the following information provided by air quality sensors:

- Benzene (C6H6) concentration in μg/m^3
- Non Metanic HydroCarbons (NMHC) concentration in μg/m^3
- Nitrogen Oxides (NOx)concentration in ppb
- Nitrogen Dioxide (NO2) concentration in μg/m^3
- Ozone (O3) concentration in μg/m^3
- Temperature (T) in Celsius degrees
- Relative Humidity (RH)
- Absolute Humidity (AH)

The training dataset is provided as train.air.csv and the test dataset is provided as test.air.csv. In both cases, the first column corresponds to the output we want to predict (CO).

To predict the Carbon Oxide concentration we are going to use the following models:

- Ridge Regression (10 points)
- Lasso Regression (10 points)
- Adaptive Lasso Regression (10 points)
- Elastic Net Regression (10 points)

In this problem you are allowed to use built in functions. For each of the models do the following:

- Fit the model
- Present optimal tuning parameters
- Present the coefficients obtained with the optimal parameters
- Present the Mean Square Prediction Error for the test dataset

Note that you should standardize the data.

Conclusion: (5 points) Which model will you select to predict the concentration of Carbon Oxide? Why?

Problem 4. 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 eleven 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 follows:

a. (5 points) Plot and present the observations for each sensor in the training data set.

b. (5 points) Use B-splines with 8 knots to reduce the dimensionality of the problem.

c. (5 points) Write the problem that we want to solve in mathematical notation. Clearly explain what your notation represents.

d. (5 points) Use group lasso to learn the B-spline coefficients. (For this part you can use built in functions, you could also use the code that you provided for problem 2 in last homework). Which sensors are correlated with the air/fuel ratio? Please plot them.

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

Table 1: List of on-board sensors