ISYE 6740/CSE 6740/CS 7641: Homework 4
80 Points Total v1.0
Due: 11:59am Nov. 17 Friday

Name:

GT Account:

Are you required to do the extra problem? Yes/No

Instruction: Please write a report including answers to the questions and the plotted figures. Please write the code in `MATLAB` and submit your code in a **'zip' file** via T-Square. You are only allowed to use specific existing package/library as requested in the problems. You need to show the iterative procedures in the code. Your code is also supposed to have explanatory comments.

1) Implementation of K Nearest Neighbor (20 points)

(a) Use MATLAB to implement the KNN algorithm, you need to use the ℓ_1 distance and the ℓ_2 distance respectively. Specifically, given two points x and y, the ℓ_1 distance is $\|x-y\|_1$, and the ℓ_2 distance is $\|x-y\|_2$. Please use the data in "train-KNN.mat" as the training set, and use data in "test-KNN.mat" as the test set.

(b) Please try $K = 1$, $K = 2$, $K = 5$ and $K = 20$ respectively, and plot the classification results. What do you observe from these results?

2) Proximal Mapping (20 points)

(a) Please derive the optimal solution to the following optimization problem,

$$\arg\min_v \frac{1}{2}\|u-v\|_2^2 + \lambda\|v\|_1,$$

where u is a d-dimensional vector.

(b) **For PhD students**, you are also required to derive the optimal solution to the following optimization problem,

$$\arg\min_v \frac{1}{2}\|u-v\|_2^2 + \lambda\|v\|_2.$$

3) Greedy Algorithm (40 points) You are given three files: "train-greedy.mat" contains the training set; "valid-greedy.mat" contains the validation set; "test-greedy.mat" contains the test set; "true-beta" contains the true regression coefficient vector β^*. Given a candidate model with β, the validation error is defined as

$$\|\widetilde{y} - \widetilde{X}\beta\|_2^2,$$

where $\widetilde{y}$ is the response vector of the validation set, and $\widetilde{X}$ is the design matrix of the validation set. In your report, you need to give the estimation error $\|\beta - \beta^*\|_2$ and prediction error on the testing set, which is defined as

$$\frac{1}{m}\|\overline{y} - \overline{X}\beta\|_2^2,$$

where $\overline{y}$ is the response vector of the testing set, $\overline{X}$ is the design matrix of the testing set, m is the number of samples in the testing set.

(a) Please use MATLAB to implement the following greedy algorithm:

Input: $X = [X_{*1}, ..., X_{*d}] \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$

Output: $\mathcal{A}^{(k)}$ and $\beta^{(k)}$

Initialize: $\mathcal{A}^{(0)} = \emptyset$ and $\beta^{(0)} = 0$

for $k = 1, 2, ..., K$

$i^{(k)} = \arg\max_i |X_{*i}^\top (X\beta^{(k-1)} - y)|$

$\mathcal{A}^{(k)} = \{i^{(k)}\} \cup \mathcal{A}^{(k-1)}$

$\beta^{(k)} = \arg\min_\beta \|y - X\beta\|_2^2$ subject to $\beta_j = 0$ for all $j \notin \mathcal{A}^{(k)}$.

end

(b) Please implement the ridge regression estimator using MATLAB. The ridge regression estimator is defined as

$$\widehat{\beta}^{\text{Ridge}} = \arg\min_\beta \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

Please select the optimal λ from $\lambda = 0.0125,\ 0.025,\ 0.05,\ 0.1,\ 0.2$.

(c) Please obtain the solution path using the `lasso` function provided in MATLAB. Please select the optimal λ from the default sequence of regularization parameters, provided by the function. Note that the `lasso` function yields regression models with intercepts. You need to take the intercept into consideration when you compute validation and testing errors.

(d) Given a Lasso estimator $\widehat{\beta}^{\text{Lasso}}$, we have obtained a refit OLS estimator by

$$\widehat{\beta}^{\text{refit}} = \arg\min_\beta \|y - X\beta\|_2^2 \text{ subject to } \beta_j = 0 \text{ for all } \widehat{\beta}_j^{\text{Lasso}} = 0.$$

Please get $\widehat{\beta}^{\text{refit}}$ using the Lasso estimator obtained in (c). Is $\|\widehat{\beta}^{\text{refit}} - \beta^*\|_2$ smaller than $\|\widehat{\beta}^{\text{Lasso}} - \beta^*\|_2$? Why?