MATH637, Fall 2023

Homework 5

Due Monday, December 1st, 11:59pm

Notes: There will be no Colab template for this assignment. You are supposed to create and submit the Colab notebook (which contains the corresponding codes, figures, and conclusions).

(2%) Step 1. Generating dataset

Write Python code to generate a dataset that contains 20000 examples.

In this dataset, each 2-dimensional input $\mathbf{X} = (x_1, x_2)$ is drawn uniformly random from a multivariate normal distribution with

$$\mu = (0, 0) \quad \text{and} \quad \Sigma = \begin{pmatrix} 1 & 1.75 \\ 1.75 & 4 \end{pmatrix}$$

and the response y is computed by

$$y = 2x_1 + \epsilon$$

where ϵ is Gaussian noise with mean zero and standard deviation 0.1.

(4%) Step 2: Linear feature selection.

- Set up a Lasso regression model with regularization parameter λ
- Use 5-fold cross-validation to choose an optimal value of λ, denoted by λ^*
- Perform Lasso(λ^*)
- Conclusion: Does the procedure recover the correct significant/non-significant features?

(4%) Step 3: Lasso with standardized data.

- Using Min Max Scaler with feature range $(-1, 1)$ to standardize the feature $\mathbf{X}$ to obtain the transformed matrix $\mathbf{X}'$
- Repeat Step 2 with data $(\mathbf{X}', y)$.
- Conclusion: Does the procedure recover the correct significant/non-significant features?