BINF 610 Applied Machine Learning (Spring 2023)

HW1

Due: March 7

Regression

In this assignment, you are asked to analyze gene expression data using regression models. The data set (downloadable at Canvas site) contains transcriptomic data in tab delimited format for five genes (columns) under 500 different conditions (rows). The hypothesis is that gene5's expression is controlled by gene1, gene2, gene3, and gene4, but the actual relationship is unknown.

A) Train a linear model, in three modes: i) no regularization, ii) ridge regularization, iii) LASSO regularization

B) Train a polynomial model (up to degree 10), in three modes: i) no regularization, ii) ridge regularization, iii) LASSO regularization

For each case, you are required to report the following:

1. The trained models, namely, the best values of the model parameters.

2. Plot the learning curves for training errors and validation errors, with a 400 (train)-100(validation) split of the dataset and varied training size from 20 to 400 with an increment of 20.

3. Discuss each case (i.e., linear model vs polynomial mode) in terms of underfitting or overfitting.

4. Suggest the overall best model for fitting the underlying relationship between the regulate and regulators.