Homework 1

Statistical Machine Learning • STAT GR 5241 • Spring 2025

Assigned: January 27

Due: February 21

Instructions:

- *You may work with others on this homework assignment but all solutions must be written up and submitted individually.*
- *All homework assignments must be submitted in pdf format and should be at most* **6 pages** *in length. Any material beyond 6 pages will not be graded.*
- *You must submit all code used to complete this homework assignment as an appendix (this can go beyond the 6 pages). Failure to submit code will result in 20% reduction.*
- *You are permitted 4 total late days on homework assignments throughout the semester. Any late days taken beyond these 4 will incur a 20% reduction per day.*
- *This homework assignment has 110 total points available and any points earned above 100 will count as extra credit.*

1. (60 Points) Data Analysis. For this problem, please use the Communities and Crime data set available at `https://archive.ics.uci.edu/dataset/183/communities+and+crime`. The goal is to predict the violent crime rate based upon features of a community. If you process the data beyond what is given (which you will likely need to!), please document all steps.

   (a) (30 points) What are the most important features?

   (i) Compare and contrast the top features as determined by:
   - Statistical significance via Least Squares.
   - Best Subsets.
   - Step-wise approaches (and/or Recursive Feature Elimination).
   - Lasso.
   - Elastic Net.

   (ii) Fit and visualize regularization paths for the following methods:
   - Lasso
   - Elastic Net (at two separate α's).
   - Ridge.

   (iii) Reflect on these results. Are the top features different for each method? Why or why not? If there are tuning parameters, how did you determine these? Do different tuning parameters yield different important features? Are there any features consistently selected by all methods? What are the most important features and how did you determine this? Explain and expand on your responses.

   (b) (30 points) Which linear method is best for prediction? For this part of the problem, you will need to randomly split the data into 60% training, 20% validation and 20% test sets. If methods have any tuning parameters, you should fit models on the training data, choose the tuning parameters by minimizing the prediction error on the validation set, and then report the final prediction error on the test set. Repeat this procedure 10 times and average the results.

   (i) Compare the average prediction MSE on the test set for the following methods:

- Least Squares.
- Ridge Regression.
- Best Subsets.
- Step-wise approaches (and/or Recursive Feature Elimination).
- Lasso.
- Elastic Net.

(ii) Visualize the results of your comparisons. You can choose to only show the results with the best tuning parameter values.

(iii) Reflection. Which types of methods give the best prediction error? Why do these methods perform well? Do any methods seem to overfit to the training set? If so, why? Do all the methods that give similar predictions choose the same subset of variables? Which is the overall best method for prediction on this dataset? Explain and expand on your responses.

2. (50 points) Empirical or Mathematical Demonstrations of Regression Properties. For these problems, you can empirically demonstrate the desired properties by using the data mentioned above or designing your own linear regression simulation. Alternatively, you may mathematically prove these properties or demonstrate these via a mathematical example.

   (a) (10 points) Empirically demonstrate or mathematically show that that fitting linear regression with an intercept term is equivalent to (i) fitting linear regression when centering Y and centering the columns of $\mathbf{X}$, and (ii) fitting linear regression when adding a column of ones to $\mathbf{X}$.

   (b) (10 points) Empirically demonstrate or mathematically show that the least squares solution has zero training error when $p > n$.

   (c) (30 points) Empirically demonstrate or mathematically show properties of methods with correlated features.

      i. Empirically demonstrate or mathematically show that the least squares estimate has high variance for correlated features.

      ii. Empirically demonstrate or mathematically show that ridge regression groups highly correlated features for sufficiently large λ.

      iii. Empirically demonstrate or mathematically show that lasso regression tends to select only one out of a group of features that are highly correlated.