NYC Data Science Bootcamp

Regularization and Cross Validation

Section 1: Ridge Regression

Purpose: Demonstrating understanding of how to run model

1. **Load:** Load in the `[06]Prostate.txt` dataset into your workspace. This dataset comes from a study by Stamey et a. (1989) of prostate cancer, measuring the correlation between the level of a prostate-speci c antigen and some covariates. The included variables are the log-cancer volume, log-prostate weight, age of patient, log-amount of benign hyperplasia, seminal vesicle invasion, log-capsular penetration, Gleason score, and percent of Gleason scores 4 or 5; the response variable is the log-psa.
2. **Train test split:** Create an 80% - 20% train-test split with your data. Please use set.seed(0) so the results will be reproducible.
3. **Fit a model:** Use library `glmnet` to fit a ridge regression model on your training data by setting up a grid of lambda values `10^seq(5, -2, length = 100)`. Save the coefficients of these models in an object.
4. **Visualization:** Plot the coefficients of these models and comment on the shrinkage.
5. **Cross Validation:** Perform 10-fold cross validation and use set.seed(0) on your training data with the grid of lambda values defined in part 2. Save the output as an object.
6. **Visualization:** Create and interpret a plot associated with the 10-fold cross validation completed in part 4.
7. **Results:** What is the best lambda?
8. **Fit a model:** Fit a ridge regression model using the best lambda on the test dataset. What is the test MSE associated with the best lambda value?
9. **Refit a model & Results:** Refit the ridge regression using the best lambda in your original dataset. Briefly comment on the coefficient estimates and MSE. Why is this MSE smaller than the test MSE you found in part 7?

Section 2: Machine Learning Theory

Purpose: Demonstrate theory of lecture material

1. **Lambda:** What is lambda in Ridge, Lasso and Elastic Net regression? What is its function?
2. **Cross Validation:** What is the purpose of doing cross validation? Explain the k-fold cross validation process.
3. **Cross Validation:** How should we choose cross validation folds typically? What does it mean when we choose a larger number of k-fold?

Section 3: Challenge Questions - Lasso Regression

Purpose: Push yourself for more advanced topics

Continue using the `[06]Prostate.txt` dataset.

1. **Lasso Regression:** Repeat the entire analysis performed in question #1, but use the lasso regression method this time.
2. **Compare:** Compare your final ridge and lasso models. Which one would you choose to use? Why?