MATH 3190 Homework 6

Regularization, Cross-validation, Dimension reduction

Due 4/11/2022

In this homework you will practice using cross-validation to fit data using LASSO and K-nearest neighbor models. Please upload to your GitHub an R Markdown document answering the following:

1. (20 points) A researcher wants to determine how employee salaries at a certain company are related to the length of employment, previous experience, and education. The researcher selects eight employees from the company and obtains the data shown below (the dataset is available as a tibble in the .Rmd).

Salary	Employment	Experience	Education
$57,310	10	2	16
$57,380	5	6	16
$54,135	3	1	12
$56,985	6	5	14
$58,715	8	8	16
$60,620	20	0	12
$59,200	8	4	18
$60,320	14	6	17

   (a) Fit a standard least squares regression model to these data and interpret the results. After looking at the statistical significance of the βs, which covariates would you include in a final model?

   (b) Use `glmnet` to fit a LASSO model to these covariates. Try λ=1000, 800, 500, and 1. How do the results compare to each other and the least squares model?

   (c) Which LASSO model (i.e. λ) would you select? (note you are not just restricted to λ values of 1000, 800, 500, and 1). Justify your answer.

   (d) Use `glmnet` to fit a Ridge regression model to these data. Try λ=1000, 800, 500, and 1. How do these results differ from the least squares and LASSO models?

2. (20 points) The `cereal.csv` dataset provides nutritional information on nearly 80 common breakfast cereals. The 'rating' column provides an overall rating for each cereal (possibly from Consumer Reports?). Use a LASSO regression model to identify the best predictors of cereal rating. Evaluate the model for λ values of 8, 5, 3, and 1 (among others). Which λ would you choose and why? Which covariates best explain the rating?

3. (20 points) An automobile consulting company wants to understand the factors on which the pricing of cars depends. Use an Elastic Net model and the `car_price_prediction.csv` dataset to determine which variables are significant in predicting the price of a car. Use cross-validation to find an optimal value for λ. Interpret your final model.