Spring 2023, 958:588 Financial Data Mining Homework 1.

DUE: Midnight, Friday, February 10th

INSTRUCTIONS: Submit all files on Canvas. For the free response parts, type your solution into an electronic document or take a picture of your solution and make sure that all hand-writing is legible. For the code parts, submit the completed files on Canvas. **Email and physical submissions will NOT be accepted.**

For code that imports csv files, make sure that all the imported files are in the same directory as the code.

NOTE: This homework is available in both R and Python. You may choose the programming language most familiar to you.

NOTE: Your code must run without error. Points may be deducted if there is any error that causes the code to crash.

NOTE: You may work in groups with other students but you must write down the names of all the members of your group.

1 Linear Regression Basics

This problem is designed to be a gentle introduction to the coding questions in Problem 2 and Problem 4.

1.1 (code and free response)

Download the `hw1_problem1` R or Python notebook file. The code should run without error. Describe how the samples Y_i, X_i's are generated in this file. What is the distribution and variance of the noise ϵ_i?

1.2 (code and free response)

Modify λ to vary between $0, 5, 10, 20, 40$. What do you observe about the resulting bias, variance, and MSE?

2 Ridge Regression with Cross-Validation

Download the hw1_problem2 R or or Python notebook file. Download also movies_hw1.csv. Our dataset movies consists of 2951 movies along with their average rating (vote_average) on the website www.themoviedb.org/. Each movie also has features such as runtime in minutes, log of budget in dollars, age in number of days since release, number of votes, genre indicators, etc. We will use these additional features to predict the rating of a movie. We use a training set of 300 movies and a test set of 2651 movies.

2.1 (code and free response)

Complete the code in hw1_problem2 R or Python file by filling in all the parts with the label FILL IN.

What is the final test error? What is the lambda chosen by CV?

2.2 (code and free response)

An advantage of using ridge regression with cross-validation is that we can safely construct additional features which may help us better predict the response Y. In the part of hw1_problem2 R or Python file labeled For part (b), construct the following features. An example is given showing how to construct log_age which is the log of (age + 1).

1. log_budget_sq: the square of the log_budget.
2. log_revenue_sq: the square of the log_revenue.
3. log_vote_count: the log of (vote_count + 1). We add 1 in case the vote count is 0.
4. Action.Adven: indicator that is 1 if the movie is both Action and Adventure. 0 otherwise.
5. Rom.Com: indicator that is 1 if the movie is both Romance and Comedy. 0 otherwise.
6. vote_budget: the product of log_vote_count and log_budget.
7. long: indicator that is 1 if the movie runtime is greater than 120 minutes. 0 otherwise.

What is the final test error? What is the lambda chosen by CV?

3 Variable Selection

3.1 (free response)

Let s be a fixed integer. Consider the following variable selection procedure for selecting the s most relevant variables. For every column $X_{\cdot,j}$, compute the correlation $c_j = Cor(X_{\cdot,j}, Y)$. Output the s variables whose correlations are the largest in absolute value, i.e., output $\hat{S}$ such that $|\hat{S}| = s$ and such that $|c_j| \geq |c_{j'}|$ for any $j \in \hat{S}, j' \notin \hat{S}$.

Is this a good procedure? Why or why not? Justify in a few sentences.

3.2 (free response)

Let u be a p-dimensional vector and $\lambda \in [0, \infty)$. What is $\arg\min_{v \in \mathbb{R}^p} \|u - v\|_2^2 + \lambda \|v\|_2^2$ in terms of u and λ?

Use your answer to argue informally that if $v^* = \arg\min_{v \in \mathbb{R}^p} \|u - v\|_2^2 + \lambda \|v\|_2^2$ and if $u_j \neq 0$ for every $j \in \{1, \ldots, p\}$, then $v_j^* \neq 0$ for every j as well.

4 Refitted Lasso with Cross-Validation

4.1 (code)

Download the file `hw1_problem4` R or Python file, `cars.csv`, and `ISTA.R`. We use the same `cars` dataset as from the lecture demos. Our goal is to predict the fuel efficiency of a car in its mpg (miles per gallon) using features such as weight, horsepower, acceleration, etc.

Complete `hw1_problem4` R or Python file by filling in all the incomplete parts of the file labeled `FILL IN`.

4.2 (free response)

Why do we compute the cross-validation error `errs[il]` using `beta_refit` instead of `beta_lasso`? Justify in a sentence or two.

5 Linear Algebra Basics

5.1 (free response)

Suppose $p < n$. Is it always possible, for any $n \times p$ matrix X and any n-dimensional vector Y, to find a p-dimensional vector β such that $X\beta = Y$ (i.e., $\|X\beta - Y\|_2^2 = 0$)? Justify your answer intuitively with a sentence or two.

5.2 (free response)

Let $X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} \in \mathbb{R}^{3 \times 2}$ matrix. Perform matrix multiplication to show that

$$X^\top X = X_{1\cdot} X_{1\cdot}^\top + X_{2\cdot} X_{2\cdot}^\top + X_{3\cdot} X_{3\cdot}^\top.$$

5.3 (free response)

Recall that the trace of a square matrix $M \in \mathbb{R}^p$ is defined as the sum of the diagonal entries of M, that is, $\text{tr}(M) = \sum_{i=1}^p M_{ii}$.

Let $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$ and $B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$. Perform matrix multiplication to show that $\text{tr}(AB) = \text{tr}(BA)$.

Extra credit (3 points)

5.4 (free response)

Suppose $X \in \mathbb{R}^{n \times p}$ where $p \geq n$. Suppose that the rows of X are linearly independent so that $XX^\top$ is invertible. Let $\tilde{\beta} := X^\top(XX^\top)^{-1}Y$. Prove that $\tilde{\beta}$ is the solution to the following:

$$\begin{aligned}&\min_{\beta \in \mathbb{R}^p} \|\beta\|_2^2\\ &\text{such that } \; X\beta = Y.\end{aligned}$$

Hint: first show that for any vector $\alpha \in \mathbb{R}^p$ satisfying $X\alpha = 0$, we must have that $\alpha^\top \tilde{\beta} = 0$.