1001 Data Analysis Project 2 Report - gjd9961

January 27, 2022

1 Introduction

The purpose of this report is to explain the results obtained in our statistical analysis of the "Movie Ratings Data-Set". The statistical prediction methods used in this report include Linear Regression, Ridge Regression, and LASSO regression. Additional descriptive statistics were provided using Python Data-Frame manipulation to calculate co-variance and thus correlation.

2 Descriptive Statistics

Employing the techniques we've learned about PCA, I created a helper function to calculate a Co-Variance matrix using the users ratings as features. To accomplish this, first I selected only the necessary data, the first 400 columns and the first 1097 rows, transposed the data such that the observations of user ratings would become the features, and then I standardized the data by subtracting the mean and dividing by the standard deviation of each column.

The resulting co-variance was a square symmetric matrix with dimensions 1097x1097. I then calculated the maximum of each column of the co-variance matrix, making sure to first subtract the Identity Matrix as I didn't want the max function to grab the diagonal values, which should all be approximately 1, as it just expresses the variance of our standardized features. After accomplishing this, I looped through the Co-Variance matrix once again to identify the user pair that expressed the highest correlation for each user. The results are as follows and also in Figure 1 at the end of this report.

3 Regression Analysis

The second part of this report focused on predicting users personal question responses, (the dependent variable), based on their respective ratings of the movies, (the features, or dependent variable). After diving up the data into its respective Data-Frames, we employed three methods, Ordinary Least Squares regression, Ridge-Regression, and LASSO-Regression. Ridge-Regression and LASSO-Regression increased training MSE while decreasing testing MSE, as they each used the L2 and L1 norm respectively to regularize the procedure and minimize over-fitting our regression models.

User1	User2	Correlation
0	583	0.551171
1	831	0.725494
2	896	0.784047
3	364	0.640055
4	896	0.528441
5	99	0.612641
6	239	0.602601
7	896	0.514100
8	896	0.706144
9	1004	0.752591

Table 1: Uesrs $1, \ldots, 9$ Correlation

4 Correlation Q&A

1. What is the pair of the most correlated users in the data?
   User: 831 had the highest correlation with user: 896

2. What is the value of this highest correlation?
   Highest Correlation value: 0.99954

3. For users 0, 1, 2, ..., 9, print their most correlated users.
   See Figure 1 above

5 Regression Q&A

1. Model $df\_pers = function(df\_rate)$ by using the linear regression
   Training MSE for Linear Regression model: 0.5983
   Testing MSE for Linear Regression model: 3.28386

2. Model $df\_pers = function(df\_rate)$ by using the ridge regression with hyper-parameter values alpha from [0.0, 1e-8, 1e-5, 0.1, 1, 10]
   The best choice for alpha is 10, as the testing error for Ridge regression is minimized. See figure 2 below.

3. Model df_pers = function(df_rate) by using the lasso regression with hyper-parameter values alpha from [1e-3, 1e-2, 1e-1, 1]. What is the best choice for alpha?
   The best choice for alpha is 1 or 10, as the training and testing error for LASSO regression those alpha values is both minimized and the same. See figure 3 below.

Iteration	Alpha	Training Error	Testing Error
0	0	0.5983	3.28386
1	1e-08	0.5983	3.28386
2	1e-05	0.5983	3.28385
3	0.1	0.59841	3.18705
4	1	0.60337	2.69378
5	10	0.65533	1.79504

Table 2: Ridge Regression Training and Testing Error by Alpha

Iteration	Alpha	Training Error	Testing Error
0	0	0.5983	3.28686
1	1e-08	0.5983	3.28685
2	1e-05	0.5983	3.26904
3	0.1	1.18138	1.16724
4	1	1.19952	1.17425
5	10	1.19952	1.17425

Table 3: LASSO Regression Training and Testing Error by Alpha