Ryan Herrin, Luke Stodgel

DS7333

01/14/2023

**Modeling Chemical Compound Data using LASSO and Ridge Regression**

**Introduction**

The purpose of this regression task is to predict “critical temperature” for different chemical compounds found in our data set.

**Methods**

The data set is composed of different chemical properties such as atomic mass, atomic radius and density, as well as chemical name abbreviations as they are found on the periodic table. The data set has 168 features and 21,263 rows.

We created LASSO (L1) and Ridge (L2) regression models using this data set.

Here is an ***overview*** of our script execution order:

1. Merge the two given data sets into one
2. Scale the data
3. Use GridsearchCV to find the best alpha for our LASSO model, then graph the coefficients at each alpha
4. Use best alpha from the LASSO gridsearch to find the average MSE over 100 fits, and graph those MSEs to obtain a distribution
5. Use GridsearchCV to find the best alpha for our Ridge model, then graph the coefficients at each alpha
6. Use best alpha from the Ridge gridsearch to find the average MSE over 100 fits, and graph those MSEs to obtain a distribution
7. Display the results for LASSO
8. Display the results for Ridge

***Specifics based on our overview***

***Scaling***

We scaled our data using the StandardScaler package and we used the scaled X data to build our models.

***GridsearchCV***

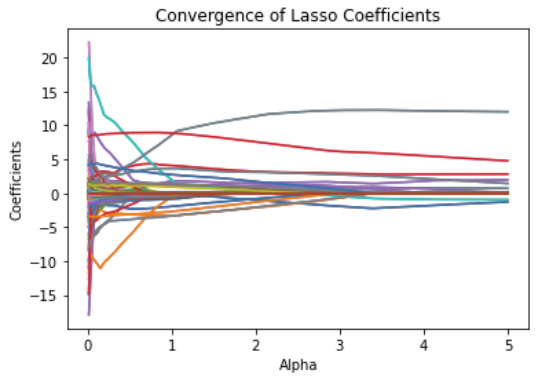
For our LASSO gridsearchCV algorithm, we used the parameters *cv=5* and *scoring=’r2’*. We tested alphas from 0.01 to 5.0, incrementing by 0.01 each time. For each alpha we performed a randomly seeded 5-fold cross-validation. The best alpha we got from this was 0.35.

For our Ridge gridsearchCV algorithm, we used the parameters *cv=5* and *scoring=’neg\_mean\_squared\_error’*. We tested alphas from 1 to 1999, incrementing by 10 each time. For each alpha we performed a randomly seeded 5-fold cross-validation. The best alpha we got from this was 1891.

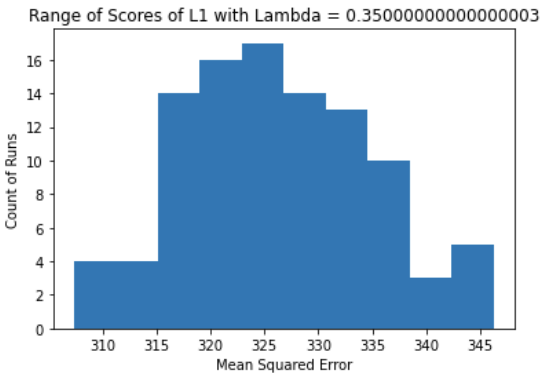
For both the LASSO and Ridge gridsearch algorithms, we picked their respective scoring metrics based on the optimal alpha value they were producing, and how well our models performed using those alphas. For example, we picked scoring=’r2’ instead of scoring=’neg\_mean\_absolute\_error’ for our LASSO gridsearch because the optimal alpha we got when using scoring=’r2’ gave us significantly more consistent model predictions and MSE scores compared to when we used scoring=’neg\_mean\_absolute\_error’.

**Results**

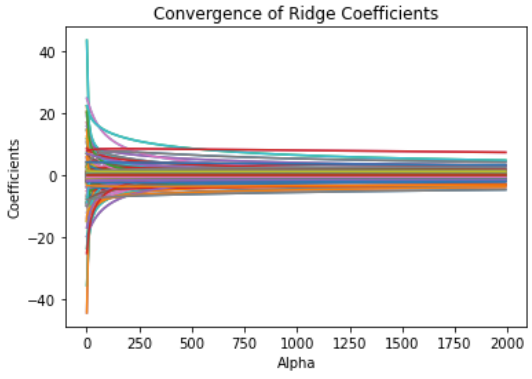
Below is a graph showing the convergence of the lasso coefficients during the gridsearch runs. As you can see, at around alpha=0.35 most of the 168 features have had their coefficients reduced to 0.

****

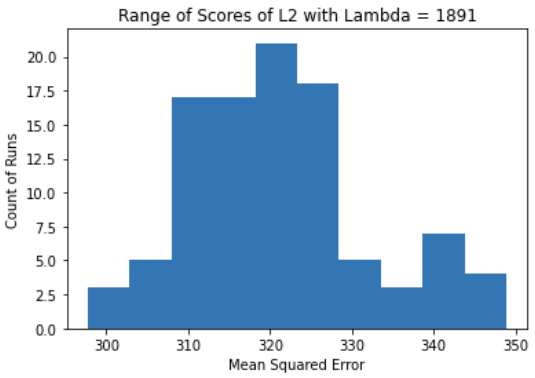
Below is a graph showing the distribution of MSEs obtained from running 100 LASSO models with lambda = 0.35. Our average MSE was 326.18 and a 95% confidence interval for these is [324.22, 328.14]. The confidence interval was calculated using the norm.interval function from the python scipy package.

****

Below is a graph showing the convergence of the ridge coefficients during the gridsearch runs. As you can see, at around alpha=1891 most of the 168 features have had their coefficients significantly reduced and the variable noise has been reduced as well.

****

Below is a graph showing the distribution of MSEs obtained from running 100 LASSO models with lambda = 0.35. Our average MSE was 323.67 and a 95% confidence interval for these is [321.48, 325.85]. The confidence interval was calculated using the norm.interval function from the python scipy package.

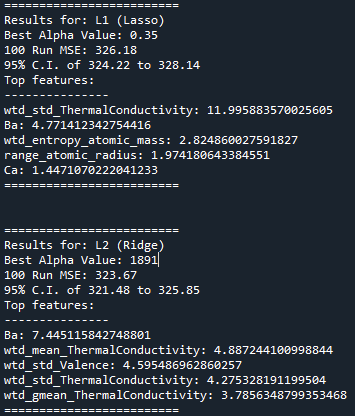
****

**Conclusion**

For L1 regularization we obtained an optimal alpha of 0.35, average MSE of 326.18 (CI: [324.22, 328.14]) and the most important feature was wtd\_std\_ThermalConductivity with a coefficient of 11.99. The next runner up for most important feature was Ba with a coefficient of 4.77.

For L2 regularization we obtained an optimal alpha of 1891, average MSE of 323.67 (CI: [321.48, 325.85]) and the most important feature was Ba: 7.45 with a coefficient of 7.44. The next runner up for most important feature was wtd\_mean\_ThermalConductivity with a coefficient of 4.89.

If we had to pick one *most important* variable, we would pick wtd\_std\_ThermalConductivity. With 168 features, we think that using L1 regularization along with the scoring=’r2’ is better able to capture the relationship between the independent and dependent variables in this data set than L2 regularization.

****

Appendix A

Link to main script: <https://github.com/Abillelatus/ML2-Case-Studies/blob/main/CaseStudy_1/MSDS-7333-CaseStudy-1.py>

Link to full CaseStudy\_1 full contents:

<https://github.com/Abillelatus/ML2-Case-Studies/tree/main/CaseStudy_1>