STA160 Final Project Report: Critical Temperature Prediction for Superconductors Based on Extracted Features

Siyuan Li
University of California, Davis
sssli@ucdavis.edu

Abstract: Within the report, we will apply a detailed data analysis for one real multivariate dataset. The data explored contain continuous variables. The report will extend the exploration of various regression methodologies to predict superconductors critical temperature. The analysis also intends to apply the prediction by reducing the data dimension with the extracted features that have the most contribution to represent the fitted model.

DATA ANALYSIS: SUPERCONDUCTIVITY DATA SET

I. INTRODUCTION

The analysis of the project is concentrated on Superconductor data, which contains two important parts for the feature description. The resource originally came from National Institute for Materials Science (NIMS), and provided by Kam Hamidieh from University of Pennsylvania, Department of Statistics. The first part of data collected information of elemental properties of superconductors, and the second part of data summarized the information of superconductors' chemical formulas. Specifically, the first part of data is constructed with 81 different features of 21263 superconductors, and all of parameters are real-valued continuous variables.

II. DESCRIPTIVE DATA EXPLORATION

Following the first step of data analysis, we carried on an exploration of basic test statistics of the dataset. For further convenience to utilize the data resources, we define the train.csv as *feature dataset* and the unique_m.csv as *chemical dataset.* We will apply clustering model and regression analysis for the feature dataset to predict critical temperatures. It is necessary for us to understand the its information of how the dataset is set up. Meanwhile, chemical dataset is also helped for the analysis, and it served as a subsidiary to develop a deeper understanding of element distribution inside superconductors.

As a crucial step for cleaning the original dataset, we find that neither of two data resources has missing values. In feature dataset, we see that the data collected 8 different properties for every superconductor. As for every property, it calculated 10 related mathematical features for statistical application. Those variables are numerical and continuous. We summarized the information in the following table.

Properties of superconductors:	Property-related features:
1. Atomic Mass 2. First Ionization Energy 3. Atomic Radius 4. Density 5. Electron Affinity 6. Fusion Heat 7. Thermal Conductivity 8. Valence	1. Mean 2. Weighted mean 3. Geometric mean 4. Weighted geometric mean 5. Entropy 6. Weighted entropy 7. Range 8. Weighted range 9. Standard deviation 10. Weighted standard deviation

Table 1

In feature dataset, there are two unique columns of data. One is the prediction label, critical temperatures; and the other is about the number of elements contained in one superconductor. They will serve an important role in our model prediction; we therefore summarized the distribution of two variables.

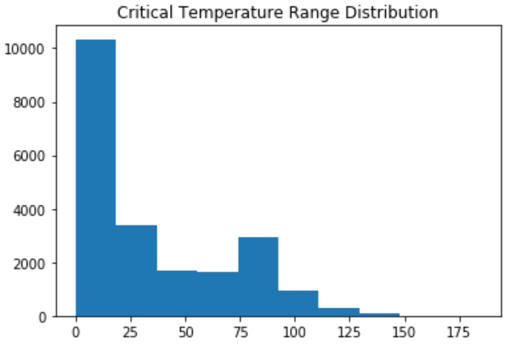

Figure 2

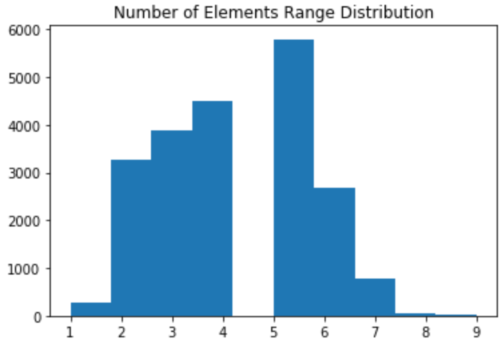

Figure 1

	Mean	Median	Min	Max	Std
1. Critical Temperature	34.4212	20.00	0.000210	185.00	34.2543
2. Element numbers	4.1152	4	1	9	1.4392

Table 2

By the above summary plots and table, we find that the critical temperature for most superconductors is clustering in the interval of zero and 25 (K). As for the number of elements inside those superconductors, the distribution if quite even, which inferred that there is no typical pattern for our analysis objects. In this way, we explored the chemical dataset to find more information of superconductor attributes. Generally, there are 86 chemical elements recorded. We intended to find some similarities among 21263 superconductors. The plot and statistical results are presented as below.

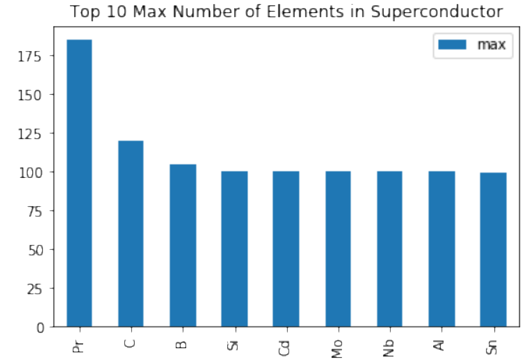

Figure 3

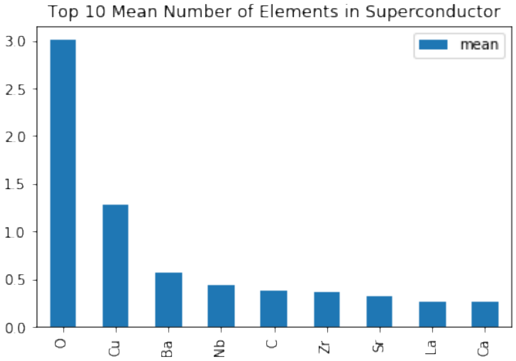

Figure 4

From the chemical dataset, we explored the distribution of 86 elements inside all researched superconductors. The most key feature for us to understand is to find which elements appear most inside those subjects. Through a general review, we can observe by Figure 4 that Oxygen has the most frequent appearance in superconductors, it followed Cu, Ba, Nb and Carbon. While the observation of Figure 3, the max number elements bar plot, showed us that Pr have the largest number in one superconductor's chemical formula broke-up. It followed by C, B, Si and Cd. We have found some elements both appeared in the above plots, for instance, Carbon. It implies that this element has a very determined function in consisting superconductors. For further prediction analysis, we would apply the former knowledge, and discuss advantages and setbacks for each methodology applied.

III. Regression Model Analysis

To reiterate our aim in this report is to construct a robust model for the extracted features of superconductor to predict its critical temperature. As the basis exploration of feature dataset demonstrated us the prediction label is numerical variable, and the other variables are continuous, we would largely focus on linear regression model techniques. The methodologies we applied at the first step included four trials, and they are Multiple Linear Regression (Ordinary Least Squares), Ridge Regression, Lasso Regression and Bayesian Ridge Regression. To compare the model efficiency and robustness, we calculated the mean square errors (MSE) and coefficient of determination (R squared) for estimation. The statistical results based on 30 random testing are summarized as below table.

Linear Model	MSE	R^2
OLS	310.7876	0.7349
Ridge	312.8587	0.7335
Lasso	323.6733	0.7238
Bayesian Ridge	312.4642	0.7329

Table 3

By the above table, we can see that out of four utilized model fits, the Lasso Regression provided the most unsatisfying result. It has the largest mean square error and also the least R squared value. It clearly infers that Lasso regression model cannot explain the prediction outcomes in a good way. Meanwhile, we compared the other three model fits and find that those regression methods have the similar outcomes. Their MSE values is located in the interval of between 310 and 312. At the same time, their R squared value is about 73.20%. specifically, we find the best model fit is by multiple linear regression, which has the least mean square error (310.7876) and the largest R squared value (0.7349). For further regression analysis, we would have a deeper exploration of the model fit and change its parameter values to produce the better prediction outcome.

However, since we are dedicated to find the best model fit for predicting critical temperature. We again looked back to our feature dataset. Previously, we have already found that the label's distribution in Figure 3, and it is not evenly distributed. By observing its range and frequency, we found the critical temperature is heavily right-skewed, which means that most superconductors have the lower critical temperature. The most frequency is contained in the interval between zero(K) and 25(K). We hope to categorize this group as "Lower Temperature". The maximum critical temperature is 185(K). To better fit the models, we define the left data points into two additional groups as "Medium Temperature" (25-100K), and "Higher Temperature" (100-185K). For later exploration, we will fit same methodology for each group and compare their predicting performance, which can better help us to improve the model accuracy.

Also, as the dataset contained many information related one single property of superconductors. We would like to reduce the data dimension through Principal Component Analysis (PCA). The result is summarized in following plots and table.

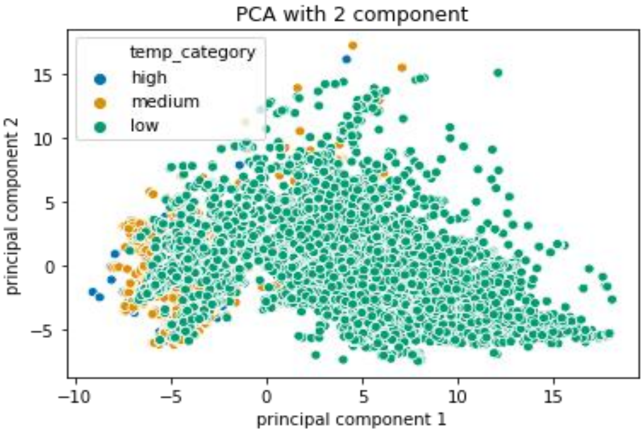

Figure 5

As the data has been categorized into three groups, we performed the principal component analysis on the feature dataset. In our analysis, we choose to set the number of PCA equals 2. Then we calculated the mean squared distance (MSE) between the new transformed data and the basic form. The value equals to 0.5058, which a good standard result for our data dimension reduction. Therefore, we would use the PCA transformed data to run three model fit. We hope to compare the improvements with the previous exploration.

As for a better comparison between PCA techniques and the unchanged form, we first refit the linear models on the grouped feature dataset. Again, the MSE value is estimated for evaluating model performance. Based on the previous analysis, we found that Lasso regression is not a good model fit, and therefore we no longer apply it. The statistical results for the other three model are listed as follow.

Groups (non-PCA)	Linear Model	MSE	R^2
Low Temperature	OLS	22.9496	0.4767
	Ridge	22.9740	0.4756
	Bayesian Ridge	23.6498	0.4606
Medium Temperature	OLS	259.4006	0.5166
	Ridge	264.2284	0.5076
	Bayesian Ridge	270.7527	0.4956
High Temperature	OLS	156.6744	0.5435
	Ridge	189.1305	0.8694
	Bayesian Ridge	73.2976	0.3061

Table 4

According to the above summary table, we find that after the splitting of original data essentially help the predication model to have better result. In all three temperature groups, we found that mean squared errors are reduced by quite a large amount. It indicated that the linear models should be separated to different conditions in order to have additional improvements. Specifically, we explored each group’s result. It demonstrates that for lower temperature group, all three models produced a great prediction result. Their MSE value is reduced to around one tenth then before. Multiple linear regression still be the best model fit with least MSE value and largest R Squared value. However, its advantage is not significant. In medium temperature group, we found that MSE value is not reduced greatly. The reason might be smaller number of samples of superconductor are recorded in this interval.

Again, multiple linear regression model had the best performance. Under this group condition, OLS model’s advantage is obvious than the other two. When we are analyzing the high temperature group, the result is quite opposite to our previous two result. We found that multiple linear regression model for this group condition produced the unsatisfying prediction result, as it has a large MSE value. Compared to the ridge regression and OLS, the result infers that there is no best model fit because Bayesian Ridge Regression model the smallest MSE with a relatively small R squared value. While Ridge function has a good result in explaining the prediction variables, its MSE value is the largest among three techniques. From a general perspective, we still believe the Multiple linear regression model could be the suitable option for superconductors’ critical temperature prediction.

In the next step analysis, we understand that outliers might be a serious problem for our linear models. We would utilize the PCA to reduce the data dimension and refit the model fit to compare its changes.

Groups (PCA-2 component)	Linear Model	MSE
Low Temperature	OLS	36.3303
	Ridge	36.3304
	Bayesian Ridge	36.3305
Medium Temperature	OLS	502.3364
	Ridge	502.3364
	Bayesian Ridge	502.3585
High Temperature	OLS	78.2216
	Ridge	78.2216
	Bayesian Ridge	78.2783

Table 5

By the previous PCA analysis, we find the transformation for both the training and testing of feature datasets. The three linear models are refit by the transformed data, and mean squared values are calculated for each group. This time as we have standardized the dataset, MSE value doesn't have a significant difference within group. What we find meaningful is related to the comparison against the previous model fit without principal component application. We find that for lower temperature group, the MSE value is increased. It infers that performing PCA under this group is not a good option. This is might due to the samples in lower group is already approximately normal distributed and without any disturbing effect of outliers. The reduce of dimension cause the model to lose capturing all important features, thus leading the prediction error increased. In the second group, we find that the result is quite opposite. The reduce of data dimension brought a great information loss in the prediction model fit. We observed that MSE values for all three methods are greatly increased. Also, PCA might not be a good option medium temperature group. However, in the high temperature group, the result is rather pleasing. We found that in general PCA helped us to reduce mean squared error for all models. In general, PCA can be considered as a good technique to boost our computation timing, and it certainly have trade-offs in capture all necessary features for prediction. As in our analysis, we still choose the original dataset to perform further exploration.

Based on our above regression analysis, we think the multiple linear regression model has the overall best performance. We therefore summarized the 10 most important features for our predication model. They are the variables has the most significant effect when changing a unit.

Top ten key features for OLS model fit	
1. First Ionization Energy - Weighted geometric mean 2. Atomic Radius - Weighted geometric mean 3. Valence - Weighted geometric mean 4. Valence - Entropy 5. Atomic Radius - Entropy	6. Atomic Mass - Entropy 7. First Ionization Energy - Weighted geometric mean 8. Valence - Geometric mean 9. Electron Affinity - Entropy 10. Valence - Standard deviation

Table 6

IV. CLASSIFICATION ANALYSIS (K NEAREST-NEIGHBORS)

The regression analysis has provided a good view to understand the superconductors' prediction task. In this part of exploration, we would extend our research in K-Nearest Neighbor algorithms to help us find the critical temperature based on extracted features. As traditionally conceived, KNN method is frequent be used for classification problems. However, its intuition can be also applied in regression problems. Our aim is to find the best combination K value for prediction model and check if the method is a better fit than multiple linear regression (OLS). The below summary statistics demonstrated the KNN model performance at four values of neighbors based on 30 random testing samples.

Groups (KNN model)	Number of neighbors	MSE
Low Temperature	1	16.8416
	2	18.6504
	5	25.1176
	10	31.2900
Medium Temperature	1	204.6548
	2	240.2766
	5	412.5602
	10	540.5724
High Temperature	1	76.3096
	2	79.5035
	5	122.4112
	10	124.8643
Total Original Data	1	158.5096
	2	195.1448
	5	319.8631
	10	424.0413

Table 7

As we computed the mean squared errors for KNN model at each different circumstance, we compared those obtained results with previous value. In general, we run the algorithm with the total feature dataset, and with every divided temperature groups. We find that the MSE value for total data points is much smaller than the Multiple linear regression model. The smallest MSE value of KNN model in the superconductors' critical temperature is 158.5096, whereas the OLS model is much larger than this value (310.7876). Clearly, we saw a 48.9974% reduction in mean squared errors. Specially, we further our analysis in comparing three individual groups statistical result. We found that the conclusion remains the same as the whole dataset, every temperature group's test MSE value is smaller for K nearest-neighbor algorithm than for linear regression. Moreover, looking at KNN algorithm itself, we also hope to find the best parameters value for the model. We modified the k value from 1, 2, 5 to 10. The trend in above table shows us that as the number of neighbor increases, our prediction error also increased. The best model fit is selected at k =1. In total, we could consider that for predicting superconductors' critical temperature, KNN model is a good option to applied. Its intuition focused on finding the most related data point to our new testing sample. Therefore, we can largely reduce the MSE error as we picked the most similar training data point when utilizing the K Nearest-neighbor algorithm.

CONCLUSION

With the prediction model analysis for superconductors to obtain the critical temperatures, we carried on two-part exploration. The major concentration of first part is on regression model. As the predicting label is a real-valued and continuous variable, we researched on different linear model techniques.

In summary, we applied Multiple Linear Regression, Ridge Regression, Lasso Regression and Bayesian Ridge Regression. We calculated the MSE value and R squared value to determine the performance of each model fit. The second part of our exploration focused on a unique methodology, K Nearest-Neighbor algorithm. The technique can be applied both for solving classification and regression issues. According to the statistical analysis, in regression model fits, we determined the OLS model is the best model to predict critical temperature. When we are approaching the predicting labels in our feature dataset, we find that the distribution of the label is not quite normal distributed, which is not perfectly satisfied the basic assumptions of linear model. However, we categorized the data points into three major groups to reduce the interference of sampling distribution. The calculation result of MSE value of multiple linear regression model still provided the overall best performance for individual group. Additionally, we find the most contributing coefficients features related to "First Ionization Energy," "Valence," and "Atom Radius." Therefore, we pick the OLS model as the best model in regression analysis.

In the next step, we utilized the KNN method to find the most similar data point to predict the temperature, which finally proved to be a better solution than OLS model. The model has both smaller MSE value when using in total dataset scope and single temperature group. In summary, we preferred the KNN algorithms for superconductivity data to predict the critical temperature.

As for our reflection in the exploration, there are also some research directions we hope to continue in the future study. due to the uneven distribution of prediction labels, we categorized the data group into three groups. However, the action might leaded us to lose some key extracted feature. We always saw a good model fit performance in the lower temperature group. While in the medium and higher temperature group, none of the above models can produced a great prediction result. Later, we hope to explore on the chemical constructions of superconductors and carry on an analysis of the effect brought by various elements. Meanwhile, we believe there is also a much improvement for our regression analysis. As the label is continuous instead of categorical, we are dedicated to find a more robust linear model based on mathematical calculation. Last, we also hope to find the correlations between all the input variables. As our data dimension is larger to perform computational algorithms, we might be under the circumstances of getting overfitting models. we therefore need to continue the research on the coefficients of linear regression models to detect their improvements and drawbacks.

References

[1] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

[2] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science &amp; Engineering*, *9*(3), 90–95.

[3] Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.

[4] McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

[5] Waskom, M., Botvinnik, Olga, O&#39; Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, *Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017).* Zenodo. *https://doi.org/10.5281/zenodo.883859*