Predicting the Bulk Modulus of Inorganic Crystals

The provided data stems from Matminer[1] and is extracted from a data set of inorganic crystals.[2] The starting point for the data is the crystal composition which is then expanded with the ”Magpie” feature generator. We limit ourselves here to just predict the bulk modulus. The script get_data.py is the one with which the data was generated. It is technically not of interest to you, but if you are interested in how it was done or you do not understand some feature names, you can look there and compare with the matminer documentation.

Expectations for the Report

Write this report in the style of a scientific paper similar to a short communication. That means the content list should be something like this:

- Introduction
- Background and Theory
- Methods
- Discussion and Results
- Conclusion

Proper English and the basics of scientific writing (citations, reference to figures, etc.) are expected. Write a brief, but proper Introduction/Motivation (max. 1 page text) for the report. In the Theory section explain the Tree regressor of your choice (see Task 2.3).

Expectations for the Code

Use the functions and packages that are provided to you by us. Do not import other functions or packages. Your code should be commented and your variable names should make sense as otherwise it is hard to help you if you get stuck. A variable that makes ”sense” has name that indicates its stored values. When you submit your report, submit the code as well.

Task 1: PCA

Task 1.1: Implementation

Download the script pca.py and implement the PCA with the class PrincipalComponentAnalysis. Use the provided class variables (e. G. self.means) to store intermediate results. This way the validity of your solutions is checked by our StudOn program. You should check your solution on StudOn with the provided test there (Only the PCA implementation is checked on StudOn). If your test succeeds and the other tasks have been done too, you may hand in your report.

Task 1.2: Spectrum of Principal Components

Perform PCA on the features and plot the (sorted) cumulative spectrum of principal components versus the number of principal components. Discuss what this implies for the data set.

Task 2: Finding the Optimal Model

In this section, you shall try to find the model that makes optimal predictions. Measure the prediction performance with the coefficient of determination R^2. Please read this entire task before starting working on this Task as otherwise you may write stupid code!

Task 2.1: Derivation Linear Regression

Derive univariate and multivariate linear regression for a one dimensional output. Put it in a separate file (not in the report) and call it Supplementary Information. It does not need to be commented much, but show the calculation as the end result is in the lecture slides.

Task 2.2: The Optimal Linear Model

Fit a Linear Least Squares model, a Lasso and a Ridge model to the provided data. Find the optimal regularization parameter for Ridge and Lasso via Gridsearch. Save the ten most important features (largest norm of weights) for later.

Task 2.3: Polynomial Expansion

Perform a polynomial feature expansion with cross terms (e. g. x1 * x2) up to a order of your choice and train the previously mentioned linear models to it (least squares, ridge and lasso with optimal regularization parameter). Please give a reason for your choice of order (e. g. computational expense). Save the ten most important features/terms (largest norm of weights) for later.

Task 2.4: Tree Regressors

Train a decision tree regressor and choose the optimal tree depth. Additionally select one of the methods to modify a decision tree (Adaboost, Gradient-Boost, Hist-Boost). Store the ten most important features. Explain the modificiation method that you chose in the theory section. Decision trees have an feature importance feature. Explain it (roughly) in the theory section. Do not fit the tree regressors to the polynomially expanded features as it will take forever and regression trees are already nonlinear models.

Task 2.5: Kernel Ridge Regression

Take two kernels of your choice and optimize the kernel parameters and the regularization parameter by grid search. Give a short explanation why kernel ridge regression is not particularly suited (except with the linear kernel) for feature selection.

Task 2.6: Compare performance across different models

During the entire task 2 (unless explicitly specified otherwise) fit each model once to standardized data[1] and once to non-standardized data. Change the size of the training set. Discuss how both things affect the model performance (measured by the regression metric(s) of your choice) and how it changes the emerging most important features.

Task 3: Feature Selection

In this task, compare the previously found most important feature i) across the models ii) with methods specifically designed for feature selection.

[1] the polynomial features should be standardized as well

Task 3.1 Least Angle Regression

Apply LASSO with LARS to find the ten most important features. Explain (roughly) how LARS works in the theory section. Do this both for the initial features and the polynomial+interaction expansion.

Task 3.2 Recursive Feature Elimination

Take a tree model (either the un-modified tree or some modified version) and one linear model (e. g. LASSO) and apply recursive feature elimination (RFE) with both to find the ten most important features. Explain RFE in the theory section.

Task 3.3 Comparison

Compare the features identified in Task 2 with the ones in Task 3. Do the methods agree with one another? Can these features be physically motivated? Are the features consistent for different training set sizes?

References

[1] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla *et al.*, “Matminer: An open source toolkit for materials data mining,” *Computational Materials Science*, vol. 152, pp. 60–69, 2018.

[2] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. Van Der Zwaag, J. J. Plata *et al.*, “Charting the complete elastic properties of inorganic crystalline compounds,” *Scientific data*, vol. 2, no. 1, pp. 1–13, 2015.