

Applied Data Analysis
Winter 2022
Homework 4
Due February 12, 2022, 9:00am

THIS IS AN INDIVIDUAL ASSIGNMENT
ALL WORK YOU SUBMIT MUST BE YOURS AND YOURS ALONE

Homework Instructions

- This is an individual assignment. All work you submit must be yours and yours alone.
- It consists of one prediction problem, similar to the work we have done over the past two weeks.
- You will need to submit a set of predictions, code file, and writeup.
- Files needed for the homework can be downloaded from Canvas.
- Submit your files to Gradescope.

Question

You are provided with data about superconductors that has been collected. In this problem, you will use this data to predict the critical temperature of superconductors. The critical temperature is the temperature at which the electrical resistivity drops to zero.

The data contains a list of 81 chemical properties of the of the superconductor, along with it's critical temperature. A second data file contains a list of chemical formulas, including the atomic count of 86 different elements, and the chemical formula of each superconductor.

You have been provided with four files. The training data is in `training.csv` and `formula_training.csv`. These files contain the 81 chemical properties, critical temperature, 86 elements with atomic counts, and chemical formulas, for 20,263 superconductors. The files `test.csv` and `formula_test.csv` contain the 81 chemical properties, 86 elements with atomic counts, and chemical formulas, for an additional 1,000 superconductors. Your task is to build an accurate model for predicting `critical_temp`.

Predictions will be evaluated using root mean squared error (RMSE) on the predicted value and the actual value of `critical_temp` for the 1,000 test superconductors:

$$\sqrt{\frac{1}{1000} \sum_{i=1}^{1000} \left(\text{critical_temp}_i - \widehat{\text{critical_temp}}_i \right)^2}$$

Deliverables

You are to turn in:

- A write-up describing what you did, how you did it, and why you did it. How did you arrive at the model you ended up using? (40 points)
- Your code file. Is it commented clearly? Could I read through your file and understand what you've done? Have you used the tools we have covered so far in a coherent manner? (40 points)
- A set of 1,000 predictions for `critical_temp`. These will be evaluated on RMSE and your relative ranking in the class. (20 points)