FIT5149 S2 2019 Assessment 1
Predicting the Critical Temperature of a Superconductor

Aug-2019

Marks	15% of all marks for the unit
Due Date	17:00 Friday 13 Spetember 2019
Extension	An extension could be granted for circumstances. A special consideration application form must be submitted. Please refer to the university webpage on special consideration.
Lateness	For all assessment items handed in after the official due date, and without an agreed extension, a 10% penalty applies to the student's mark for each day after the due date (including weekends, and public holidays) for up to 5 days. Assessment items handed in after 5 days will not be considered.
Authorship	This assignment is **an individual assignment** and the final submission must be identifiable your own work. Breaches of this requirement will result in an assignment not being accepted for assessment and many result in disciplinary action.
Submission	You are required to submit two files, one is either a Jupyter notebook or a R Markdown file, another is the PDF file generated by them. The two files must be submitted via Moodle. Students are required to accepted the terms and conditions in the Moodle submission page. A draft submission won't be marked.
Programming language	R in Jupyter Notebook or R Markdown

Introduction

Superconductivity is a phenomenon of exactly zero electrical resistance and expulsion of magnetic flux fields occurring in certain materials, called superconductors, when cooled below a characteristic critical temperature. Superconductors are widely used in many industry fields, e.g. the Magnetic Resonance Imaging (MRI) in health care, electricity transportation in energy industry and magnetic separation, etc.

Predicting the critical temperature (T_c) of a superconductor is still an open problem in the scientific community. In the past, simple empirical rules based on experiments have guided researchers in synthesizing superconducting materials for many years. Nowadays, features (or predictors) based on the superconductor's elemental properties can be generated and used to predict T_c.

In this task, we are going to analyze superconductor data from the Superconducting Material Database maintained by Japan's National Institute for Materials Science (NIMS). The aim is to build statistical models that can predict T_c based on the material's chemical properties.

Specifically, you are going to analyse a superconductor data set, which is based on real world material science data. The problem you are going to solve is: Can you

- **predict** the critical temperature T_c given some chemical properties of a material?
- **explain** your prediction and the associated findings? For example, describe the key properties associated with the response variable.

Data set

The data set was originally from from the Superconducting Material Database maintained by Japan's National Institute for Materials Science(NIMS) and prepossessed by Kam []. It contains 21,263 material records, each of which have 82 columns: 81 columns corresponding to the features extracted and the last 1 column of the observed T_c values. Among those 81 columns, the first column is the number of elements in the material, the rest 80 columns are features extracted from 8 properties (each property has 10 features). Detailed data preparation process can be found in [].

The data set files are stored in UCI's website below (click the hyper-line to download the data)

superconduct.zip : After you unzip the file, there are two data sets: **train.csv** can be used to train and validate prediction models and build a description (21,263 material records). Each record consists of 82 columns, containing number of elements (column 1), features extracted from 8 properties (columns 2-81) and the critical temperature (column 82). **unique_m.csv** tells you the chemical formula of each corresponding material.

In order to finish the analyse task, you should split the provided **train.csv** into your own training and testing sets before building the models.

Task description

In this assessment, you will focus on the following two tasks.

Prediction task

For the prediction task, the underlying problem is to estimate the critical temperature given a new conductor's properties. There are eight properties that can be used: Atomic Mass, First Ionization Energy, Atomic Radius, Density, Electron Affinity, Fusion Heat, Thermal Conductivity, Valence. For each property, ten features are extracted: Mean, Weighted mean, Geometric mean, Weighted geometric mean, Entropy, Weighted entropy, Range, Weighted range. Standard deviation, Weighted standard deviation. The provided data sets are well organised, you do not need to wrangle the data. But make sure you understand the intuition of these attributes.

To measure the performance of your model(s), you firstly split the original data into training and testing set, fit the model using the training set, do the predictions on the test set and compute the Mean Squared Error (MSE).

In this task, you are required to develop models that can accurately predict a superconductor's critical temperature. To finish the task, you should

1. develop and compare 2 to 3 models;
2. describe and justify the choice of your models;
3. analyze and interpret your results

Please note that testing set cannot be used to train your models.

Description task

The purpose of the description task is identify the key properties for a superconductor. In other words, which property contributes the most to your model's performance? Descriptions can be based on variable correlation analysis, regression equations, linguistic descriptions, or any other form. The descriptions and the accompanying interpretation must be comprehensible, useful. To finish this task, you should use proper data analysis techniques (e.g., EDA, statistics) to

1. identify a subset of attributes that have a significant impact on the prediction of the critical temperature;
2. and give statistical reasons of your finding.

Files to be submitted

There are two files required to be submitted, which are

- The **R** implementation of the two tasks in one file.
  - The file **must be** either a **Jupyter notebook** or **an R Markdown file**. Besides the R code, all the discussions must also be included in the file.
  - The name of the file **must be** in one of the following formats:
    * XXXXXXXX_FIT5149_Ass1.ipynb
    * XXXXXXXX_FIT5149_Ass1.Rmd

    You should replace "XXXXXXXX" with your student ID.
- A PDF file generated by the Jupyter notebook or R Markdown. The name of the PDF file must be in the following format
  - XXXXXXXX_FIT5149_Ass1.pdf

Please refer to the Assessment 1's Moodle page for how to submit the two files. Please note that If you do not follow the instruction to name your files, **a penalty will be applied**.

Additional learning resources

This assessment is based on the paper *A Data Driven Statistical Model for Predicting the Critical Temperature of a Superconductor* at https://arxiv.org/pdf/1803.10260.pdf

- Raw data is available at http://supercon.nims.go.jp/supercon/material_menu

Warning: Monash University takes academic misconduct very seriously. You can learn from the above materials and understand the principle of how the analysis was done. However, you must finish this assessment with your own work.