IML 2023: Predicting Saturation Vapour Pressure From Molecular Properties

Helmi Karesti, Maija Säisä

2023-11-03

Instructions

In this term project, you will train a regression model on a data set of atmospheric measurements. To complete the project, you should deliver:

- Sun 10 December: Predictions (to be submitted on Kaggle - we will provide the link later) for the test set, and a preliminary version of your project report as a single PDF file in Moodle.
- Fri 15 December: Term project presentation for some of you.
- Sun 23 December: The final report as a single PDF file in Moodle.

Your task

Since **saturation vapour pressure** is a continuous variable, your task is to build a regression-based machine-learning model that uses the aforementioned interpretable features or topographical fingerprints of the molecules.

NOTE: This is a non-trivial regression task. It is possible to do it in many ways. The most straightforward regression model that you could build is a linear regressor, but that is inefficient for the task since it assumes a linear relationship between input features and the predicted values. Therefore, you should undertake thorough data exploration, pre-processing, feature selection, model selection, R2 score estimation, etc., appropriately since you will report and analyse your choices and results in the project deliverable.

The project's purpose is not to (even try to!) replicate any methods in the literature, make a super-complex best-performing classifier that beats everything else or attempt to use other data sources, etc., to obtain the best possible R2 score . Do not use any method that you do not understand yourself! Accuracy of the predictions on the test data is not a grading criterion by itself, even though a terrible accuracy may indicate something else fishy in your approach (which could affect grading).

The final report should contain, among other things, the following:

- The names of the group members.
- The name of the team you used to submit the predictions on Kaggle.
- The stages of your data analysis, including how you looked at the data to understand it (visualisations, unsupervised learning methods, etc.).
- Description of considered machine learning approaches and pros and cons of the chosen approach for this application.
- Steps you took to select good features and model parameters.
- Summary of your results, insights learned, and how the regression model performed.
- As a final section, please include a self-grading report (at most 1 page) that suggests a grade for yourself (integer 0-5) by using the attached grading instructions (see below).

It is enough to use one of the basic algorithms, do the feature and model selection parts as instructed (you should probably use cross-validation!), and prepare a well-written report to pass the project.

Practical instructions for writing the report:

- Your report should read like a self-contained blog post or scientific article that is understandable and without any task description. You should explain what you have done and why you have done it so that a person familiar with machine learning can understand what you have done and could, in principle, reproduce what you have done based on your report alone. Put some emphasis on presentation and readability (one of the grading criteria): imagine that the report's reader would be your future boss, who appreciates a clear and concise presentation.
- You are not required to hand in any program code. Your report, thus, should look different from a code listing! Your report may contain code snippets if you explain what the reader is supposed to conclude from your code. We may look at them, but we won't go fishing for results and missing details from your code. In other words, all relevant parts of your report should be understandable without going through any code. If you need to include more significant chunks of code, please put them in an appendix so we can easily skip them when grading your report.
- Your report may include tables or figures. Always explain in detail what the tables or figures show and what the reader expects to conclude from them. If you have a figure or table, the text should refer to it at least once.
- You can use suitable typesetting software that produces legible PDF output (LaTeX, Word, R Markdown, etc.). There is no strict page limit so you can use a readable font (e.g., 12 pt serif font), margins, and appropriately sized figures. Note that Jupyter Notebooks often lead to poorly formatted pdf. Out of curiosity, I took a random sample of 16 similar final reports that got total points from other courses I lectured. The task was identical to this one but without self-grading (which may add a page). The page counts of these final reports were 7, 7, 9, 10, 10, 10, 12, 12, 13, 13, 14, 14, 14, 14, 14, and 14. The reports had between 7 and 14 pages, the median being 12.5.

Even though you can modify your approach and adjust your algorithms for the final report, you are not required to (and probably should not) make significant changes. The idea is to polish the report and complete whatever steps you have planned.

Introduction

Data analysis

Methods

Results

Discussion