Matlab Assignment: Predicting the Biodegradability of Chemicals from QSAR Data

Morgan Jones*

Understanding the biodegradability of chemicals is crucial in preventing their harmful accumulation and environmental dispersion. Quantitative Structure-Activity Relationships (QSAR) is a chemoinformatics modeling technique used in the field of chemistry and pharmacology to relate the chemical structure of a molecule to its biological or chemical activity. QSAR models attempt to encapsulate essential aspects of chemistry by establishing mathematical relationships between the structural features of molecules and their corresponding biological or chemical properties.

A dataset containing QSAR information along with the biodegradability status of 1,055 chemicals has been curated in [1]. You don't need any prior knowledge of chemistry or QSAR to work with this data set. You'll find the dataset in the coursework briefing files under the file named `QSAR_data.mat`. Each row in this dataset corresponds to a specific chemical, and the 41 features associated with each chemical is found in the first 41 columns of `QSAR_data.mat`. The final column in the dataset serves as a data label, where '1' indicates biodegradability and '0' signifies non-biodegradability.

Task: The task is to build a fully-validated, predictive model for biodegradability using any technique or model that you wish to choose. To complete this task you can use Matlab or Python.

Required Output: You're required to upload **two files** named using dm (data modelling) and your nine digit registration number. For example if you used Matlab and had a reg number of 123456789 you would upload the files dm123456789.pdf and dm123456789.m. Note, you are not required to upload the original data set file `QSAR_data.mat`.

1. The pdf file will contain your report that will be a **maximum of 4 pages**. The report will be formatted using the IEEE Transactions, Journals, and Letters template. This report will be structured into the following sections: Abstract, Introduction, Data Processing, Methodology, Model Analysis, Conclusion and Recommendation, References. Further details about what to include within each section are explained on the next page of this document.

2. The other file will be a master script of your code. This code should be well commented and run. Upon inspecting this file it should plausible that all numerical outputs presented in the report can be generated by running the code.

Additional Information: Only upload two files, a pdf of the report and the master script of the code file, that includes all custom code functions. Include in the comments of the code master script what are the required toolboxes to run the code. Assume that I have installed all the required toolboxes. You can use any inbuilt function from a toolbox however the highest marks will be given to those who demonstrate some level of creativity/novelty when applying standard machine learning techniques, this may require you to code some functions from scratch to allow for these customization. Any modifications/customizations should be explained and highlighted in the report.

References

[1] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, "Quantitative structure–activity relationship models for ready biodegradability of chemicals," *Journal of chemical information and modeling*, vol. 53, no. 4, pp. 867–878, 2013.

*M. Jones is the module leader of ACS6427 - Data Modelling and Machine Intelligence from the Department of Automatic Control and Systems Engineering, The University of Sheffield. E-mail: `morgan.jones@sheffield.ac.uk`

Matlab Assignment: Format

Morgan Jones

***Abstract*—For this coursework you are required to format your report using the IEEE Transactions, Journals, and Letters template. This template is provided in both LaTeX and Microsoft Word formats within the coursework briefing files. You should not edit the template margin sizes or text sizes. Your report should visually look like this document. Moreover, you are required to structure your report into the following sections:**

1) **Abstract**
2) **Introduction**
3) **Data Processing**
4) **Methodology**
5) **Model Analysis**
6) **Conclusion and Recommendation**
7) **References**

The maximum length of the report is four pages in which all of the above sections should be included. The purpose of this document is to explain what you're expected to include within each section. The abstract should provide a brief overview of the entire report: The problem you're trying to solve, what machine learning methods are applied and a summary of what was discovered numerically.

I. Introduction

In this section you will introduce the problem you're trying to solve and its importance. You will also provide a literature review, including the original data source [1] and related papers, as well as relevant papers on machine learning methods. You will discuss ethics associated with this practical problem and/or machine learning in general.

II. Data Processing

In this section you have the opportunity to go beyond what is covered in labs and lectures. You will discuss what you have done to the data before applying any machine learning algorithm and why you have done this. There is potential to talk about some (not all since the report is limited to 4 pages) of the following topics: data visualization methods like histograms and violin plots, mean-centering, variance scaling, missing data, duplicate data, outliers, and data compression (clustering or PCA).

III. Methodology

You may break this section down into smaller subsections/paragraphs in which you will mathematically explain each chosen machine learning model. Remember the total length of the report should be no more than 4-pages so you must be concise and think carefully about what information relating to the model/technique that is most important. Details about the mechanism/algorithm used to train the model and prevent overfitting should also be provided. The goal of this section is for you to demonstrate a technical understanding of the underlying theory for some machine learning methods.

M. Jones is the module leader of ACS6427 - Data Modelling and Machine Intelligence from the Department of Automatic Control and Systems Engineering, The University of Sheffield. E-mail: morgan.jones@sheffield.ac.uk

IV. Model Analysis

In this section you will mathematically present performance metrics. You will clearly explain if any splits have been made to the data. You will demonstrate your ability to communicate complex data in figures. You will analyse each figure/plot.

V. Conclusion and Recommendation

In this section you will discuss the advantages and disadvantages of different machine learning methods. In your conclusion you will recommend a learning algorithm and justify this recommendation.

References

[1] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, "Quantitative structure–activity relationship models for ready biodegradability of chemicals," *Journal of chemical information and modeling*, vol. 53, no. 4, pp. 867–878, 2013.

Matlab Assignment: Predicting the Biodegradability of Chemicals from Quantitative Structure-Activity Relations (QSAR)[1] Data

Marking Scheme

It is recommended that you study these criteria before completing the assignment.

Criteria	Indication of Marks Assigned
Code • Did the code run? • Is it plausible that the code could produce the results reported? • Is the code well commented demonstrating a good level of understanding?	/10
Abstract, Introduction & Referencing • Was it clear what was the practical importance of the problem? • Were ethical considerations discussed? • Was a high-quality literature review conducted?	/10
Data Processing & Methodology • Has some type of data processing been applied? • Has the data processing been explained and justified well? • Has the student demonstrated a level of understanding of machine learning techniques rather than just blindly applying inbuilt toolbox functions? • Has the student attempted to creatively modify well-known machine learning models or attempted methods beyond the scope of this course?	/10

Model Analysis, conclusion and recommendation
• Have techniques been attempted? Have the techniques succeeded?
• Were the reported results sufficient and appropriate to determine the quality of the modelling process and the final model? Could the conclusions be justified?
• Has a critical appraisal, including both numerical and theoretical aspects, taken place of the technique?
• Have the numerical results been analyzed and discussed in any depth?
• Has the work gone beyond the treatments that have been undertaken in lab sessions?
• Have comparison methods and benchmarks been explored?
• Have the following factors been considered:
  ○ Modelling choices, especially in terms of type structure and complexity
  ○ Has complexity control and validation been considered?
  ○ Have choices made been defended?
	/20

Total	/50
Penalties	
Failed to use the IEEE Transactions, Journals, and Letters template format	-5%
Wrong file type (pdf file for report and single master script for code)	-5%
Exceeded 4 page report limit	-5%
late submission (www.shef.ac.uk/ssid/exams/policies)	Variable

Unfair Means

The assignment should be completed individually. You should not discuss the assignment with other students and should not work together in completing the assignment. The assignment must be wholly your own work. Any suspicions of the use of unfair means will be investigated and may lead to penalties. See http://www.shef.ac.uk/ssid/exams/plagiarism for more information.

[1] Mansouri et al, (2013) Quantitative Structure−Activity Relationship Models for Ready Biodegradability of Chemicals, J Chemical Information & Modelling, 53, 867-878.