MLDS Supervised Learning

Final Assessment 2021-22

DEADLINE: Monday 28 March 2022, at 08:59 (UK).

1 Introduction

Hydrocarbons are chemical compounds with molecules consisting of hydrogen and carbon atoms. Hydrocracking is an industrial process in which hydrocarbons with molecules made of long chains of atoms are broken down into smaller molecules. This is done by the addition of hydrogen at high pressures and temperatures in the presence of a chemical catalyst. This process is used in the petrochemical industry to increase the proportion of extracted hydrocarbons that have shorter molecules, which are typically more useful to consumers and attract a higher profit on sale.

The process of hydrocracking can be controlled by adjusting the temperature of the reactor (or equivalently its pressure), the catalyst that is used and the time for which the hydrocarbon mixture is within the reactor. The composition of the effluent hydrocarbon mixture will depend on these settings of the reactor and on the composition of the feed hydrocarbon mixture that entered the reactor.

The most useful and valuable hydrocarbons are not the heaviest (most dense, with longest molecules) or the lightest (least dense, with smallest molecules), but those within a density in a specific, intermediate range. This is known as the target range.

The chemometrics team at an oil refinery want to take a data-driven approach to understanding which of their control variables are most influential on the yield of hydrocarbon in the target range of densities. They would also like to be able to accurately predict the yield in this range for a given feed composition and reactor calibration. This would allow them to not only maximise profits, but to minimise waste by calibrating reactor output to meet but not exceed demand. You have been hired as a consultant to help the team achieve these aims.

2 Data Description

The chemometrics team has provided you with the data in the csv file `mass_fractions.csv`. This file contains contains 497 observations of 47 variables. These detail the reactor settings on each of 497 days: which of 3 catalysts were used, the reactor temperature in degrees Fahrenheit, and the reactor residence time of the mixture in hours. Along with the reactor settings for each day, you are given the the composition of the feed and effluent hydrocarbon mixtures. These are given as the proportion of the overall mass in each of 20 density intervals. Interval 1 corresponds to the longest, heaviest molecules and interval 20 to the shortest, lightest molecules. The team are targeting a high combined yield in density ranges 13, 14, and 15.

3 Report Guidelines

Your report should compare at least two supervised learning approaches to this modelling task. It should include:

- A title page which includes an executive summary of your main findings (at most 200 words);
- An introduction giving the aims and motivations of the project and any relevant background information;
- An exploratory analysis of the data provided;
- A methods section describing to the chemometrics team the models you will consider and how you will compare these;
- A results section describing the outcomes of applying these models and comparisons to the data provided;
- A summary and detailed interpretation of your main findings for the chemometrics team. You should highlight any limitations to your work and how some of these might be addressed in future work.

4 Submission Guidelines

Your report should be at most **8 pages in length**, including the title page and any figures or appendices. Your report should be a pdf document titled `YOUR-CID_chemometrics_report.pdf`

You should also provide a clearly commented code file that allows your analysis, modelling and figures to be recreated by the chemometrics team, who have a basic understanding of R programming. This code file should be titled `YOUR-CID_chemometrics_analysis.R`.

Please ensure that you upload these documents to the correct sections of the virtual learning environment, doing so before the deadline stated at the top of this document. You will be marked based on the documents as presented, accommodations will not be made for incorrect uploading of documents.

Plagiarism: The report and code that you submit should be your own work and properly attribute any sources or references that you use. Note that software tools are used for plagiarism detection.