SCIF30006: Advanced Data Science for Scientific Computing

Capstone Project

This project will count for 40% of your unit mark. It consists of data analysis tasks and a report on the outcomes. Both components will have the same weighting in this assessment.

This project is intended to assess your understanding and engagement with the different data analysis approaches you have encountered throughout this unit. While this will assess all of the intended learning outcomes for this unit, ILO1, ILO4 and ILO5 will be of greater importance to the assessment.

1. **Explain the basic steps involved in preparing and curating data and assess data using standard statistical descriptors.**
2. Explain different techniques for extracting information from data and select suitable regression models.
3. Describe the basic principles of machine learning, including choice of models and tuning of parameters.
4. **Apply some of the more common learning and clustering algorithms used in machine learning.**
5. **Describe and implement advanced data visualisation techniques for multi-dimensional data sets.**

Instructions

In this exercise, we ask you to showcase the following skills: -

- Analysing data in a scientific context
- Applying exploratory statistics to assess an unfamiliar dataset.
- Using and evaluating learning and clustering algorithms in a scientific context.
- Presenting data analysis results in a professional report, including appropriate figures and data tables to illustrate the work done and support the conclusions presented.

You should aim to spend no more than 20 hours on data analysis/coding for this exercise and we suggest that you aim to complete this project, including any background research and report writing, in around 40 hours.

There are two options, please choose one option to work on. In both cases, please focus on the information given in this briefing document. Both datasets have come from the academic literature, but we have made changes to them, so we want you to take a fresh look, rather than be drawn into old debates. In your report introduction, please focus on presenting the relevant options for data analysis and visualisation in your relevant field, i.e. Chemistry or Physics, rather than on the specific data.

If you are not sure about any element of this project, please get in touch with Natalie (Natalie.Fey@bristol.ac.uk) before spending a lot of time on it.

Submission

Please submit all information needed to check and reproduce your work to the submission point on Blackboard. This would normally involve working code in a Jupyter notebook or a Python IDE, exhibiting best practice, but you could choose to use R or indeed Excel for all or some of your

analyses. Provided all work is fully explained and can be repeated, you can choose the platform you find most appropriate. Your report should be submitted as a word-processed document and submitted in doc/docx or pdf format.

Make sure that you showcase your skills and that you fully justify any choices made in your data analysis. Your report should give an overview of the relevant data analysis options and then present your results, providing the reader with a critical assessment of both data and models included. Make sure you follow the project brief given below and make a clear recommendation. Marks will be awarded for the selection of approaches, clarity of analysis, quality of presentation and clear links with the project brief, as well as the quality of code and analysis. The marking criteria are available on Blackboard; they are the same as used for the computing exercise you completed in TB1.

We have also shared, with permission, the work of a student from an earlier cohort that attracted a very high mark. You can find this on Blackboard. Note that the project brief was different for this work, and this has been shared with you.

Clearly indicate which option you have chosen and submit your work by the deadline of Wednesday at noon in week 23 (26th March 2025) on Blackboard. You can find information about extensions and exceptional circumstances on the assessment page for this unit.

AI Tools

You may not use AI tools in the preparation of assessed work for this unit. Please make sure that you are familiar with current UoB guidance; a good starting point is this page: https://www.bristol.ac.uk/students/support/academic-advice/academic-integrity/, and we strongly recommend that you complete the University's AI study skills module.

Formally, this assessment is classified as:

Category 2: Minimal – You may only use tools such as spelling and grammar checkers in this assignment, and their use should be limited to corrections of your own work rather than substantial re-writes or extended contributions.

This is the UoB default for most assessments.

Data and context

Please choose one of the options for your work.

Option 1: Identifying Pulsars from Broadband Radio Emissions

Pulsars can be detected by radio telescopes on Earth, and they are of interest to astrophysicists. However, most of the candidate signals are due to radio frequency interference or noise, and it can be difficult to label them appropriately, requiring human intervention. You are asked to consult for a research group who have a labelled database in-house but are not sure what the best approach to using this is. They are not interested in expanding or changing their database and have asked you to focus on assessing whether the database can be used to predict labels for a larger dataset.

You need to explore the dataset with standard exploratory statistics and unsupervised learning approaches and then present models which can classify (label) the signals as either pulsars or spurious signals. These models should be evaluated fully with appropriate measures for model fit

and predictive ability, allowing you to make a recommendation to the research group about whether and how this database can be used.

Appendix 1 includes a list of variables and some additional notes, adapted from the published dataset. We are providing you with a small subset of that dataset here and have thus edited the published details.

Your report should follow a standard report format, i.e. Abstract, Introduction, Analysis and Discussion, Conclusion and References, and it should be accompanied by working code to perform the data analyses described.

Option 2: Classifying Molecules into Families

A local contract research organisation (CRO) have collated a dataset of variables related to boiling points. They found that subsets of compounds grouped as hydrocarbons, amines and alcohols are easier to predict than the entire dataset and have manually labelled their dataset accordingly. They are not interested in collecting additional data and are instead looking for an assessment of whether they can use this dataset to predict labels from such simple parameters for a larger database.

You need to explore the dataset with standard exploratory statistics and unsupervised learning approaches and then present models which can classify (label) the compounds as hydrocarbon, amine or alcohol; you may find it easier to divide the dataset further and try out some binary classifications. Your models should be evaluated fully with appropriate measures for model fit and predictive ability, allowing you to make a recommendation to the CRO about whether this approach can be used to assign labels reliably.

Appendix 2 includes a list of variables and some additional notes, adapted from the published dataset.

Your report should follow a standard report format, i.e. Abstract, Introduction, Analysis and Discussion, Conclusion and References, and it should be accompanied by working code to perform the data analyses described.

Appendix 1 – Data for Option 1 (Pulsars)

Note that we do not expect you to do further research on this dataset, but would like you to focus on the data analysis instead.

Adapted from the original data description: This dataset describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South).[1] As pulsars rotate, their emission beam sweeps across the sky, and when this crosses our line of sight, produces a detectable pattern of broadband radio emission. As pulsars rotate rapidly, this pattern repeats periodically. Thus pulsar search involves looking for periodic radio signals with large radio telescopes. Each pulsar produces a slightly different emission pattern, which varies slightly with each rotation. Thus a potential signal detection known as a 'candidate', is averaged over many rotations of the pulsar, as determined by the length of an observation. In the absence of additional info, each candidate could potentially describe a real pulsar. However in practice almost all detections are caused by radio frequency interference (RFI) and noise, making legitimate signals hard to find. Here the legitimate pulsar examples are a minority positive class, and spurious examples the majority negative class.

Each candidate is described by 8 continuous variables. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the Dispersion-Measure-Signal-to-Noise Ratio (DM-SNR) curve. For the purposes of this project, please assume that these are appropriate variables for this task and that the data have been cleaned and annotated appropriately. You can use all variables or a subset in the analysis you present. Note that we have reduced the number of entries substantially compared to the published database, taking a randomly selected subset of the whole database.

Column	**Label**
A	Mean of the integrated profile
B	Standard deviation of the integrated profile
C	Excess kurtosis of the integrated profile
D	Skewness of the integrated profile
E	Mean of the DM-SNR curve
F	Standard deviation of the DM-SNR curve
G	Excess kurtosis of the DM-SNR curve
H	Skewness of the DM-SNR curve
I	Class label (0 = negative, 1 = positive)

Appendix 2: Data for Option 2 (Molecules)

Note that we do not expect you to do further research on this dataset, but would like you to focus on the data analysis instead.

The initial dataset has been collected from the PubChem database,[2] and then augmented with boiling point information from the CAS databases.[3] Unnecessary characteristics and compounds were deleted from the PubChem dataset and new characteristics such as numbers of C, N, O, and side chain were added based on the molecular structures. Compounds have been labelled as "Hydrocarbon", "Alcohol" and "Amine" and then boiling points from CAS have been added to the data table. It was originally constructed to predict boiling points from molecular structures, but here you can use boiling points as a variable for classification. Note that to further simplify the dataset for this project, we have removed the molecular formula and SMILES string of the molecules considered. For the purposes of this project, please assume that these are appropriate variables for this task and that the data have been cleaned and annotated appropriately. You can use all variables or a subset in the analysis you present.

Variable name	**Description**
cmpdname	Compound Name
BoilingPoint	Measured Boiling Point
mw	Molecular Weight
polararea	area of the polar area
heavycnt	number of non-hydrogen atoms
hbondacc	number of hydrogen bond acceptors
C number	number of C atoms
N number	number of N atoms
O number	number of O atoms
Side chain number	number of side chains
Double bond number	number of double bonds
Triple bond number	number of triple bonds
Classify1	classification of structure ("Hydrocarbon", "Alcohol", "Amine")

References

1. M. J. Keith, A. Jameson, W. van Straten, M. Bailes, S. Johnston, M. Kramer, A. Possenti, S. D. Bates, N. D. R. Bhat, M. Burgay, S. Burke-Spolaor, N. D'Amico, L. Levin, P. L. McMahon, S. Milia and B. W. Stappers, *Monthly Notices of the Royal Astronomical Society*, 2010, **409**, 619-627. DOI: 10.1111/j.1365-2966.2010.17325.x.
2. PubChem, https://pubchem.ncbi.nlm.nih.gov/, (accessed 20/03/2024).
3. ACS, https://www.cas.org/support/documentation/cas-databases, (accessed 20/03/2024).