SCIF30003 – Advanced Data Science for Scientific Computing

Exercise 2 – Assessment of Topic 2: Models and Predictions

This exercise will count for 15% of the marks for this unit. It is intended to assess your technical skills in applying unsupervised learning and Bayesian approaches to investigate a scientific data set and relates mostly to the 6th and 7th intended learning outcome (ILO) for this unit:

> Apply some of the more common learning and clustering algorithms used in machine learning
>
> Describe and implement advanced data visualisation techniques for multi-dimensional data sets.

It also overlaps with the 4th ILO to some extent: "explain different techniques for extracting information from data and select suitable regression models", although we will not ask you to fit a regression model in this case.

Instructions

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

- Principal component analysis (PCA)
- K-means clustering
- Bayesian approaches
- Analysing data in a scientific context

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 (although Excel for PCA is pretty hideous, you've been warned). Provided all work is fully explained and can be repeated, you can choose the platform you find most appropriate.

Make sure that all questions are answered and that you showcase your skills, including good practice in coding. Marks will be awarded for the clarity of analysis, quality of presentation and clear links with the scientific content, as well as the quality of code and analysis. The marking criteria are broadly similar to those used for exercise 1, and they are available on Blackboard.

Submit your work by the deadline of Wednesday at noon in week 17 (22nd February 2023) on Blackboard. You can find information about extensions and extenuating circumstances on the assessment page for this unit, but please make sure you use the extension request form provided there, rather than the central route.

Data and Context

The periodic table of the elements is related to the electronic configuration and while this is governed by relatively simple principles, there are many exceptions as well. In this exercise, you will be investigating some of the properties of the lanthanides. This work is based on a published paper and you should read this carefully before you start: O. Horovitz, C. Sârbu, *J. Chem. Ed.* **2005,** *82*, 473.

For part 1, we would like you to analyse the data included in this work and provided in the file Lanthanides.csv with Principal Component Analysis (PCA) and k-means clustering. You will need to decide how to deal with missing data and identify and explain any clusters observed. Since the original paper includes PCA, we will focus on how you execute this in a suitable coding language and how you display the results.

For part 2, please address the following:

The 1st ionisation energy is defined as the energy required to remove one electron from the outer shell of an atom (labelled as I1 in your data table). Considering the change in atomic nucleus and electronic configuration, how would you expect the 1st ionisation energy to change on moving across the f-block? Hence, considering also the results of your PCA analysis, compute posterior distribution(s) for the average 1st ionisation energy for the Lanthanides. You should justify your choice of prior and likelihood distributions and sampling of data points.

While this exercise does not require a formal report, please make sure you cite any references used and make clear links between your results and the wider scientific context.