Programming Project 2

March 2024

The second "Programming Project" asks you to investigate data, train machine learning algorithms with the data, and model underlying trends in the data. You should use the skills you developed in the Special Topics in Artificial Intelligence and Deep Learning module (SCIFM0002) to achieve this. Details of the scientific problem that you will investigate are given below. You are expected to consider and justify your analysis based on what you have learned in this unit.

The project should consist of a code component and a written report. The deadline for this work is **1st May 2024**; make sure you get started as soon as possible. This project will be worth 40 % of your total unit mark, weighted as 80 % on the code and 20 % on the report. This assessment will consider the code you write, including any documentation, code clarity, programming style, and the discussion in the written report. While the report's subject focuses on a particular area of science, **the assessment of this work covers the material in the SCIFM0002 module**; therefore, directions for background discussion are given in the details below. Similar to the previous project assessment, this assessment should take around 40 hours to complete.

The code component should include all of the Python code used to perform your analysis, including any plots produced (either within a Jupyter Notebook or referenced as external files). The code should be self-contained and clear which parts of the analysis are performed at each stage. The code should be well commented and additional details can be added using separate Markdown cells (if using a Jupyter Notebook). **The code should be able to run as a whole independently (from top to bottom)**.

For the report component, you should write it using a word processor of your choice and submit it alongside the code as a Word document (.doc or .docx) or a PDF (.pdf). The report should include background details of any scientific context, a description of your analysis, discussion, and conclusions. You should also include a short abstract summarising your findings and details of any references used. See Section 2 for these details as bullet points. The report itself does not need to include explicit details of the code used. However, the overall approach and any relevant outputs should be related to the submitted code document. The separate code document can be referred to if needed.

When completed, use the "Project 2" Blackboard submission point to upload your files (on the "Assessment, submission and feedback" course content area for the "Special Topics in Artificial Intelligence and Deep Learning" unit).

You will submit multiple files for this project (at least one code file and one report). All code files should be submitted in a runnable Python format, e.g. a Jupyter notebook (".ipynb") or Python file (".py"). Other formats may not be accepted (such as ".html" and ".pdf" documents). **Please check the format for all files before submitting**.

All submissions for this assessment must have been created by the submitter and should not be created or copied using alternative resources (including copying from your peers or using large language model (LLM) tools to generate code or text). Please also see the University policy on plagiarism.

1 Dunking Biscuits in Tea

You have been employed by McVitie's biscuit company in data science. In this role, you have been tasked with analysing data related to the structural properties of different types of biscuits that the company produces. In particular, the company is interested in how the biscuits interact with tea when dunked (Figure 1).

Figure 1: A chocolate digestive biscuit about to be dunked into a cup of tea.

In this project, you will be given access to a series of datasets. The expectation is that you will use both machine learning and model optimisation to investigate and draw conclusions from the data. Additionally, you should compare the different approaches taken to understand the data.

1.1 Scientific Background

The structure of biscuits consists of gluten fibres that interlock with the help of sugars, milk, and other flavourings. This interlocking nature means that biscuits are porous in nature, with microscopic pores going through the structure. A colleague with a background in materials science suggested that the biscuits' porous structure may mean that liquids are soaked up by capillary flow action. The Washburn equation governs this capillary flow:

$$L = \sqrt{\frac{\gamma r t \cos{(\phi)}}{2\eta}}, \tag{1}$$

where L is the distance that the fluid travels into the solid, γ is the surface tension of the liquid, r is the radius of the capillary pore through which the liquid is travelling, t is the length of time for the capillary flow to occur, and ϕ is the contact angle between the solid and the liquid and η is the dynamic viscosity of the liquid. However, the colleague notes that this is a relatively simple model. The complex cross-linked structure of a biscuit may limit the utility of the Washburn model.

1.2 Available Data

Three different types of experimental data have been collected. You can use all of these datasets to complete the project.

1.2.1 "Big Data" Collection

Over the past year, one of the experimental team members has repeatedly dunked different biscuits into cups of tea. In these experiments, inspired by the Washburn equation, there has been a range of experimental parameters. This data has been stored in a comma-separated values file called `dunking-data.csv`, which has six columns, one for each of the experimental parameters measured (the column headings are given with the columns in the order that they appear in the file):

1. `gamma`: the tea surface tension, in $\mathrm{N\,m^{-1}}$.
2. `phi`: the contact angle between the biscuit and the tea surface, in rad.
3. `eta`: the tea dynamic viscosity, in $\mathrm{Pa\,s}$
4. `L`: the distance up the biscuit that the tea was visible, in m.
5. `t`: the time after initial dunking that the measurement was made, in s.
6. `biscuit`: the type of biscuit that was dunked, which is Rich Tea, Hobnob or Digestive.

The experimental setup used is given in Figure 2.

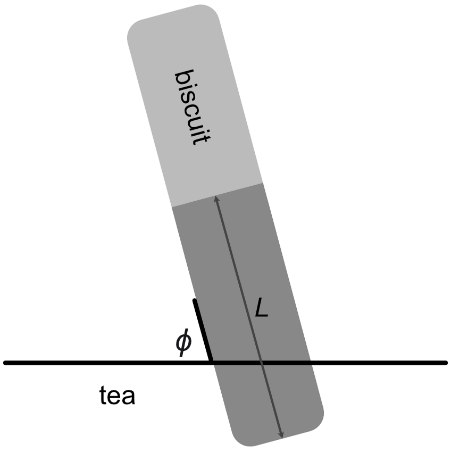

Figure 2: A diagram of the experimental setup used in the biscuit-dunking measurements.

1.2.2 Time-Resolved Measurements

Another dataset has been provided, which investigates the capillary flow rate of the tea in the biscuits. This involved taking a blind sample of each of the biscuits and measuring the length that the tea soaked up the biscuit in a time range from 30 s to 300 s. The same experimental setup was used for all three measurements:

- the tea surface tension was measured at $6.78 \times 10^{-2}\,\mathrm{N\,m^{-1}}$.
- the contact angle was 1.45 rad.
- the tea dynamic viscosity was $9.93 \times 10^{-4}\,\mathrm{Pa\,s}$.

The biscuit used for each measurement is unknown, and the data files are titled `tr-1.csv`, `tr-2.csv`, and `tr-3.csv`. Each data file contains the same information, three columns of experimental data:

1. `t`: the time elapsed in the measurement, in s, the dependent variable.
2. `L`: the length the tea has soaked up the biscuit, the independent variable, in m.
3. `dL`: an estimate of the uncertainty in length, also in m.

1.2.3 Microscopy Measurements

The final dataset was generated by taking a subset, one-sixth, of the `dunking-data.csv` samples and finding the pore radius by microscopy. This data file has similar information to the `dunking-data.csv` but without the biscuit type and with the pore radius. The columns, in order, are:

The final dataset is a subset of the `training_data.csv`, where the biscuit type is no longer identified, but this has been replaced with the pore size as measured with microscopy. Therefore, this data also has six columns:

1. `gamma`: the tea surface tension, in $\mathrm{N\,m^{-1}}$.
2. `phi`: the contact angle between the biscuit and the tea surface, in rad.
3. `eta`: the tea dynamic viscosity, in Pa s
4. `L`: the distance up the biscuit that the tea was visible, in m.
5. `t`: the time after initial dunking that the measurement was made, in s.
6. `r`: the radius of the pore, in m

This is available as `microscopy-data.csv`.

Assessment: You have been tasked with investigating the available data and outlining the analysis that you have performed in a written report. **The aim is not to do the most analysis possible but to find a coherent story in the analysis and present this.** Some examples of questions that the team are interested in include:

- Can a machine learning algorithm be used to identify the different types of biscuits?
- How is the pore radius different between the three types of biscuits?
- How accurate is the Washburn equation for biscuits, and can a machine learning regressor perform better?

You have been asked to write a report on the available data, perform an analysis, and outline the data results. Furthermore, highlight and describe how the team might use these data in future data-driven investigations.

1.3 Optional: Beyond the Brief

Within the separate mark descriptor document, you will see that completing the outline above to a good standard will allow you to achieve a very good mark. To start to move beyond this and achieve the very highest marks (excellent work), as well as fully completing the outline, your submission would need to offer additional quantification or comparison of the data (described as "beyond the brief"). Can you use this information to comment further on your findings and offer more quantitative results?

2 Report

The report should focus on the analysis of the results, so you should aim for your report to be approximately 1000 to 1500 words or 3 to 4 sides of A4 paper (including plots). Note that this is not a hard limit for the report, where possible you should aim to provide adequate detail, but also be concise and direct.

The report itself should be clearly split into the following sections:

- Abstract: Short overview and summary of the key results of your analysis.
- Introduction: A brief introduction to the problem you are solving.
- Analysis and Discussion: Details of the approach and analysis performed for the different parts of the investigation. This section should include plots or summary tables as appropriate.
- Conclusions: Any overall conclusions that can be drawn from your analysis.
- References: Full details of references used in the construction of this report (using the Royal Society of Chemistry citation format, see edu.rsc.org/resources/how-to-reference-using-the-rsc-style/1664.article for details).