Assignment 2

PHYS30053

Advanced Computational Physics & Machine Learning

In this assignment, you will develop Python code to identify patterns in data using machine learning methods. You will train machine learning models for a well-defined task, and evaluate the performance of the model in executing this task.

1 Problem

In this assignment, you can choose a dataset and corresponding problem statement from the list below. Having chosen a dataset and problem, you will need to implement an appropriate machine learning model that is capable of solving the problem. You should use suitable techniques for training the model, including an appropriate choice of loss function. You should use appropriate methods to characterise and evaluate the performance of your model, both during training and with unseen data.

1.1 Galaxy Classification

The Sloan Digital Sky Survey (SDSS) was an ambitious and highly influential astronomical survey that operated between 2000 and 20008 [1]. It used a 2.5m optical telescope at Apache Point Observatory, New Mexico. The survey imaged over 1 million galaxies, which required cataloging for further research. The Galaxy Zoo project [2] was an effort to use citizen scientists to characterise and catalogue the galaxy images. In principle, a machine learning approach to characterising such images could speed up the process of cataloguing by orders of magnitude.

1.1.1 Dataset

https://www.kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge/data

This dataset is taken from the Galaxy Zoo 2 project [2], using data from the Sloan Digital Sky Survey [1].

1.1.2 Problem

Your task is to develop a machine learning solution which can classify each galaxy image in the dataset according to the labels described in Table 2 of the Galaxy Zoo 2 paper [2]. You should employ appropriate techniques to validate your solution and evaluate its performance.

1.2 Brain Tumour Identification

This dataset contains MRI images of human brains, from a combination of sources. Such scans are typically interpreted by an expert radiographer, who will identify abnormalities including malignant tumors. A machine learning approach may benefit a radiographer in terms of the speed with which they can interpret scans, or as a backup mechanism to reduce the probability of false negatives.

1.2.1 Dataset

https://www.kaggle.com/datasets/denizkavi1/brain-tumor

This dataset was recorded at Nanfang Hospital, Guangzhou, China, and General Hospital, Tianjing Medical University, China, from 2005 to 2010. It comprises 3064 slices from 233 patients, containing 708 meningiomas, 1426 gliomas, and 930 pituitary tumors [3].

1.2.2 Problem

Your task is to develop a machine learning solution which can classify brain scans as healthy, or indicating one of three types of tumor; meningioma, glioma or pituitary tumor. You should employ appropriate techniques to validate your solution and evaluate its performance.

1.3 Solar Power Forecasting

This dataset describes the power output and meteorological measurements as time-series data from rooftop photovoltaic cells. This data can be used to make predictions of future power output from such installations.

1.3.1 Dataset

https://datadryad.org/dataset/doi:10.5061/dryad.m37pvmd99

This dataset comprises measured photovoltaic (PV) power generation data and on-site weather data collected from 60 grid-connected rooftop PV stations in Hong Kong over a three-year period (2021-2023) [4].

1.3.2 Problem

Your task is to develop a machine learning solution that can predict the power output of a photovoltaic installation over time. Note that you can choose to :

- produce a model which can predict the output of a particular installation, based on the past data from that installation

- produce a model which generalises to all installations, which will predict future output based on metadata about that installation. (You may wish to complete the above task before proceeding to this one).

1.4 Choose Your Own Problem

You are also free to identify an open source dataset and define your own problem. However, you **must** obtain approval from the unit director, Dr Jim Brooke.

2 Assignment Instructions

2.1 Code

You may develop your code however you wish, but you should submit your final results as a single Jupyter notebook (ie. one .ipynb file). Make sure you check the notebook runs correctly from a fresh kernel before submitting.

Your code must produce all results presented in your report. It should also include all code required to validate your solution, including any quantitative assessment of the output or comparison with experimental data or analytic solutions.

You MAY use AI to generate code for this assignment. However, you must properly attribute any generated code using comments, to make it completely clear what lines of code have been generated, including prompts and information about the AI/generator used.

Some things you will need to think about while developing your solution :

- How to validate your approach, including intermediate steps
- How to present your results appropriately
- How to quantitatively assess your results
- How to factorise your code (eg. what functions to use)
- Whether and how you can make your code re-usable

Please note that there are many open source solutions to the problems set here on kaggle and other repositories. You should feel free to search for ideas, however, the work you submit **must be your own**. Any work submitted which is excessively similar to publically available solutions will be referred to the academic misconduct process as potential plagiarism.

2.2 Report

Your report should describe how your code solves the problem set, why you made the choices you did, and what you have learnt from the exercise.

Do not re-state the problem. Assume the reader has read this assignment brief before they read your report - you can refer to sections and equations in this brief.

The final section of your report must be titled "Use of Generative AI". In this section you should describe how you used AI for the assignment. Make sure to include a list of the AI tools used, and a description of your approach. You do not need to repeat prompts that are included in the code comments. If you did not use AI at all, include the section but simply state "AI was not used in this assignment".

The report should be no longer than 2500 words.

Below are some specific points you should consider including in your report :

- What class of machine learning model did you use, and why ?
- What training approach / loss function did you use, and why ?
- How did you select the other hyperparameters of the model ?
- What methods did you use to evaluate the performance of the model, and why ?

2.3 Submission

You should submit BOTH the following files at the Blackboard submission point **Advanced Computational Physics and Machine Learning Coursework 2** :

- Jupyter notebook : file **Name_CW2.ipynb**
- Report : file **Name_CW2.pdf**

Where Name should be replaced with your name.

Bibliography

[1] Donald G. York et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, sep 2000.

[2] Kyle W. Willett et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, September 2013.

[3] Jun Cheng et al. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLOS One*, 120(3), oct 2015.

[4] Z Lin et al. A high-resolution three-year dataset supporting rooftop photovoltaics (pv) generation analytics. *Scientific Data*, 12(63), 2025.