ACS341 – Machine Learning Coursework Assignment

Dr Morgan Jones & Dr Giuliano Punzo

This Assignment is worth 35% of the final module mark.

The challenge

Accurately predicting household energy consumption allows local power distribution companies to better forecast energy trends and perform demand management[1]. Power system demand management has gained heightened importance as the world transitions towards renewable energy[2]. The rhetoric of the UK aiming to become "the Saudi Arabia of wind"[3] with the emergence of wind farms in the North Sea[4] has seen the nation pivot away from conventional fossil fuels towards cleaner, more sustainable sources. The North Sea's wind farms furnish a bountiful but highly variable power supply for UK households, providing a path towards national energy independence by reducing reliance on the importation of fossil fuels. Nevertheless, the primary technical hurdles hindering the increased adoption of wind energy in the UK revolve around efficiently transmitting power over long distances from the North Sea to urban centres[5], coupled with the challenge of seamlessly meeting demand during periods of low wind energy production or increased household energy use. In this project, we aim to address a component of these challenges by constructing a predictive model for household energy demand. Our client, the national grid, may then use our model to help forecast when alternative energy production facilities need to be ramped up to meet household energy demands.

This coursework aims to create an effective machine-learning workflow for predicting household energy data. Your assigned tasks, detailed on the following page, require you to devise solutions independently. Alongside demonstrating your data modelling abilities, this assignment evaluates your professional engineering skills, including adherence to specifications, delivering tested and commented code, meeting client requirements, and justifying your approach.

Deliverables

1. A report as a single PDF file;
2. Code submitted as a single .zip file.

Data available

You have been granted access to the 'household_energy_data.csv' dataset, comprising 50,392 entries. The first row contains the names of each feature variable, while the subsequent 50,391 rows contain the corresponding data points associated with each household snapshot. These data snapshots capture household energy demands, smart meter readings of diverse household appliances, and concurrent weather conditions. The dataset consists of 30 columns, each representing distinct features. The first column is entitled "**EnergyRequestedFromGrid_kW_**" and this is the variable we are trying to predict.

---

[1] Ndiaye, Demba. et al. "Principal component analysis of the electricity consumption in residential dwellings." *Energy and buildings* 43.2-3 (2011): 446-453.

[2] Jones, Morgan. et al. "Solving dynamic programming with supremum terms in the objective and application to optimal battery scheduling for electricity consumers subject to demand charges." *2017*

[3] Bamisile, Olusola, et al. "Enabling the UK to become the Saudi Arabia of wind? The cost of green hydrogen from offshore wind."

[4] Potisomporn, Panit, and Christopher R. Vogel. "Spatial and temporal variability characteristics of offshore wind energy in the United Kingdom." *Wind Energy* 25.3 (2022): 537-552.

[5] Cullinane, Margaret, et al. "Subsea superconductors: The future of offshore renewable energy transmission?." *Renewable and Sustainable Energy Reviews* 156 (2022): 111943.

Task/Assessment Description and Marks Available

Task	Marks available
Task 0: Provide well-commented code that could plausibly reproduce all results shown in the report. The code should have a main run file within the zip folder (see the following page for more details) with comments on what the code does and which toolboxes are required for the code to run.	10
Task 1: Conduct data cleaning. This could involve deciding which features to drop and which relevant features to keep, how to scale, pre-process, bound the data, etc. It could also involve a discussion about which features are most important to this specific prediction task, taking into consideration information and domain-specific knowledge other than the provided data set**.** Clearly discuss in the report what data cleaning was done and the reasons for doing this.	30
Task 2. Build a linear regression model to predict household energy consumption based on your processed data set from Task 1. Discuss implementation and technical issues such as collinearity in the report. Provide plots and metrics to assess the quality of your model.	20
Task 3. Build a second model (for example a high-order polynomial, an ANN or even a technique we have not seen in class). Detail how overfitting to the data set was mitigated. Discuss implementation and technical issues in the report. Compare the results with the linear regression model from Task 2 and justify which model is the better model. Summarize the report by articulating the motivation, ethical issues and future challenges in machine learning and AI technologies in the context of this project.	40
Penalties	
Incorrect report/code layout (for layout see following page)	-5%
Wrong file type	-5%
Exceeded page limit	-5%
Late submission (See University policy at https://www.sheffield.ac.uk/mltc/courses/learning/validation)	Variable

Technical Report and Code.

Report

- You are permitted a **maximum** of five A4 sides of 11 point type and 25mm margins. Any references, plots and figures must be included within these five pages. Don't waste space on cover pages or tables of contents. If you exceed the limit you will be penalised and content not within the 5-page limit will not be marked.
- You must save your document as a **pdf** file *only* - **no other** format is acceptable.
- Your report should consist of three sections corresponding to Tasks 1,2 & 3.

Code

Your code must run standalone, in other words, when testing we will clear the workspace and load your code. Any function you created should be included in the .ZIP file. Do not include the data in your submission. Your code should work with the dataset provided, in the shape and format it was provided, which is available to the staff marking your work. Should the data require any pre-processing, this should be done within your code. Already pre-processed data or any dataset different from the one provided will be discarded if found in your submission.

Within the .ZIP file there should be a script named "main_run", this is the file we will run, and it should generate all the results from the report. At the beginning of the "main_run", you should follow standard programming conventions and provide comments concerning the implementation details including details of any external toolboxes required.

This assignment is designed to be done in MATLAB, however, should you find yourself more comfortable using Python, you are free to use it. You are also free to use toolboxes/libraries but must detail their use in the comments in the "main_run" file.

Extenuating Circumstances: If you have any extenuating circumstances (medical or other special circumstances) that might have affected your performance on the assignment, please get in touch with the student support office (lecturers are righteously kept outside the process) and complete an extenuating circumstances form. Late submission rules apply with a reduction in 5% for every additional late day and a score of zero after 5 days.

Unfair means: All work must be **completed as individuals**. References should be used to support your domain analysis research. Suspected unfair means will be investigated and will lead to penalties. For more information on the university unfair means' guidance, please check: http://www.shef.ac.uk/ssid/exams/plagiarism.