CS7CS4/CSU44061 – Machine Learning
Academic Year 2023-2024
Trinity College Dublin

Final assignment 2023-24

- Unlike the rest of the module coursework you must do this assignment entirely yourself - you must not discuss or collaborate on the assignment with other students in any way, you must write answers in your own words and write code entirely yourself. If you use any online or other external content in your report you should take care to cite the source. It is mandatory to complete the declaration that the work is entirely your own and you have not collaborated with anyone - the declaration form is available on Blackboard. All submissions will be checked for plagiarism.
- Reports must be typed and submitted as a separate pdf on Blackboard (not as part of a zip fie).
- Include the source of code written for the assignment as an appendix in your submitted pdf report. Also include a separate zip file containing the executable code and any data files needed. Programs should be running code written in Python, and should load data etc when run so that we can unzip your submission and just directly run it to check that it works. Keep code brief and clean with meaningful variable names etc.
- Important: For each problem, your primary aim is to articulate that you understand what you're doing - not just running a program and quoting numbers it outputs. Generally most of the credit is given for the explanation/analysis as opposed to the code/numerical answer.
- If you use machine learning models not covered in the course then you must take care to show that you understand them and are not just running code in a "black box" fashion (so explain how predictions are generated from an input, what the cost function is, what the model parameters and hyperparameters are and how they affect the predictions etc).
- Reports should typically be about 5 pages, with 8 pages the upper limit (excluding appendix with code).

Part 1 [75 marks]

<u>General instructions</u>: Download the *Dublinbikes* dataset at https://data.gov.ie/dataset/dublinbikes-api. The dataset is organised in party quarterly (i.e., four data-files per year) and in part monthly, meaning that the project might involve concatenating different portions of the dataset. Note that there might be many possible answers to each of the questions that follow, so you will need to be creative and think about this on your own.

<u>Scenario</u>: You are working for FUTURE-DATA a local company specialised in data science. Dublin City Council hired your company to study the impact of COVID-19 on the city-bikes usage as they are planning to optimise the city-bike system. Dublin City Council had originally structured the city-bike network based on the forecasts of bike usage up to 2030. However, they think that the usage may not match the initial prediction because of the impact of the pandemic on our mobility. FUTURE-DATA decided that their first step should be to investigate the impact of the pandemic on the usage of the city bike network.

<u>Task</u>: The company agreed with our manager on two goals:

1. To assess the impact of the pandemic on the city-bike usage for the pandemic period.
2. To assess the impact of the pandemic on the city-bike usage for the post-pandemic period.

Hint 1: Note that there are many ways to do this and angles to explore. Temporal vs. spatial dynamics might have changed. A first approach might be to use descriptive statistics only. But it is also required to use machine learning to estimate how the city-bike usage would have been if the pandemic had not happened. Predictions can be augmented by including information that is not available in the Dublin-bikes dataset (e.g., weather data).

Hint 2: The original features tell us about bike and bike stand availability. However, that is a different concept from “bike usage” i.e., how many bikes have been taken from (or brought to) that station. We suggest deriving a “bike usage” features for the analyses. Other ways of tackling the tasks are also accepted, but remember to justify your choices and discuss your results clearly. Please also remember to report clear, compact figures.

[indicative breakdown of mark: (i) data preprocessing and feature engineering 20 marks, (ii) machine learning methodology 20 marks, (iii) evaluation 25 marks, (iv) report presentation 10 marks]

Part 2 [25 marks]

(i) What is a ROC curve? How can it be used to evaluate the performance of a classifier compared with a baseline classifier? Why would you use an ROC curve instead of a classification accuracy metric? [5 marks]

(ii) Give two examples of situations where a linear regression would give inaccurate predictions. Explain your reasoning and what possible solutions you would adopt in each situation. [5 marks]

(iii) The term ‘kernel’ has different meanings in SVM and CNN models. Explain the two different meanings. Discuss why and when the use of SVM kernels and CNN kernels is useful, as well as mentioning different types of kernels. [10 marks]

(iv) In k-fold cross-validation, a dataset is resampled multiple times. What is the idea behind this resampling i.e. why does resampling allow us to evaluate the generalisation performance of a machine learning model. Give a small example to illustrate. Discuss when it is and it is not appropriate to use k-fold cross-validation. [5 marks]