TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Final Assignment 2022-23 CS7CS4/CSU44061 Machine Learning

- **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 file).**
- **Include the source of code written for the assignment as an appendix in your submitted pdf report (as text, not as screenshots). 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 good care to demonstrate 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 10 pages the upper limit (excluding appendix with code).**

Downloading Dataset

- Go to the Inside Airbnb web site at `http://insideairbnb.com/get-the-data/`, scroll down to "Dublin, Leinster, Ireland" and download the `listings.cvs.gz` and `reviews.csv.gz` files. The first file contains details of Airbnb listings in Dublin, the second contains timestamped reviews. The listings data contains summary review scores for each listing (i.e. each individual property), broken down by location, cleanliness, value, etc (the entries are labelled `review_scores_rating,review_scores_accuracy,review_scores_cleanliness,review_scores_checkin,review_scores_communication,review_scores_location,review_scores_value`).

Assignment

1. Write a short report evaluating the feasibility of predicting for a listing the individual ratings for accuracy, cleanliness, checkin, communication, location and value and also the overall review rating. Note that the rating values all tend to be high values (greater than 4) so some care is needed when evaluating prediction accuracy. Also not all reviews are in English and may contain non-text content such as emojis, so you will need to decide how best to accommodate that. What features tend to be associated with a high rating e.g. do superhosts tend to have higher ratings, does the neighbourhood affect the location rating, do more highly rated listings tend to have more reviews?

   Appropriate feature selection is likely to be important so give this due attention. Select two distinct machine learning approaches (justify your choice), apply them to the dataset and critically evaluate their prediction performance.

   Remember its v important to clearly explain/justify any design choices that you make and present data in support of any conclusions you arrive at. When presenting results you must include enough information to allow the reader to understand how the values presented were calculated and how they should be interpreted (that includes model parameters, plotted values etc). Include any code you use in an appendix.

   Take get full marks for the report you will need to some time to organise your report in an efficient way: avoid long rambling text and "brain dumps" (more is not always better), give some thought to presenting your results and analysis without using large numbers of figures (as a guideline you might have 1 figure for initial data visualisation and feature selection, 1 figure for the machine learning model fit and evaluation, and for the comparison between the two approaches, 1 figure for the additional questions; each of these figures can use subplots to save space), avoid lots of wasted whitespace (using two column format can save a lot of space).

   [75 marks: indicative breakdown (i) feature engineering 20 marks, (ii) machine learning methodology 20 marks, (iii) evaluation 20 marks, (iv) visualisation and report writing 15 marks]

2. Don't just Google for the answers to the questions below and do not use jargon you don't fully understand. Read the lecture notes, think about your answers and make sure to explain them in your own words.

   (i) Give two examples of situations when logistic regression would give inaccurate predictions. Explain your reasoning. [5 marks]

   (ii) Discuss some advantages and disadvantages of a kNN classifier vs an MLP neural net classifier. Explain your reasoning. [5 marks]

   (iii) 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. Why are $k = 5$ or $k = 10$ often suggested as good choices? [10 marks]

   (iv) Discuss how lagged output values can be used to construct features for time series data. Illustrate with a small example. [5 marks]