TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Supplemental Assignment 2020-21 CS7CS4/CSU44061 Machine Learning

- **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.**

Downloading Dataset

- Go to the Dublin Bikes Open Data web site at `https://data.gov.ie/dataset/dublinbikes-api`. Download the "Dublinbikes 2020 Q1 usage data". This should be a large-ish (250MB) csv file. The file contains the number of bikes at each of the bike stations in Dublin, updated every 5 mins from 1st Jan to 24th Feb 2020. Pick two bike stations to study, choosing two stations with different patterns of behaviour e.g. one located in the city centre and one in the suburbs. In your submitted report please state which two stations you chose and why.

Assignment

1. Write a short report evaluating the feasibility of predicting bike station occupancy 10mins, 30mins and 1 hour in the future. Appropriate feature selection is likely to be important so give this due attention. Select two 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 any conclusions you arrive at. Include any code you use in an appendix. [75 marks: indicative breakdown (i) feature engineering 20 marks, (ii) machine learning methodology 20 marks, (iii) evaluation 25 marks, (iv) report presentation 10 marks]

2. (i) What is a ROC curve. How can it be used to evaluate the performance of a classifier. [5 marks]

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

   (iii) Discuss three pros/cons of an SVM classifier vs a neural net classifier. [5 marks]

   (iv) Describe the operation of a convolutional layer in a convNet. Give a small example to illustrate. [5 marks]

   v) 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. [5 marks]