Module Code
COM00029H

BEng, BSc, MEng and MMath Degree Examinations 2019–20

DEPARTMENT OF COMPUTER SCIENCE

Introduction to Neural Networks (INNS)

Open Individual Assessment

Issued: 1st March 2021, 12:00 noon
Submission due: 26th April 2021, 12:00 noon
Feedback and marks due: 24th May 2021, 12:00 noon

All students should submit their answers through the electronic submission system: http://www.cs.york.ac.uk/student/assessment/submit/ by 26th April 2021, 12:00 noon. An assessment that has been submitted after this deadline will be marked initially as if it had been handed in on time, but the Board of Examiners will normally apply a lateness penalty.

Your attention is drawn to the section about Academic Misconduct in your Departmental Handbook: https://www.cs.york.ac.uk/student/handbook/.

Any queries on this assessment should be addressed by email to Simon O'Keefe at simon.okeefe@york.ac.uk. Answers that apply to all students will be posted on the VLE.

Rubric:

Answers must not exceed 8 A4 pages in total, with a minimum 11pt font and minimum 2cm margins either side. This limit includes any title page, diagrams, references, and so on. Excess pages will not be marked.

Your answer must address all of the questions outlined in the 'Report' section to gain full marks. Each question must be answered in its own dedicated section of the report, in the order given. Material outside the four prescribed sections will not be marked.

Your exam number should be on the front cover of your assessment. You should not be otherwise identified anywhere on your submission.

1 Scenario

The scenario for this assessment is the introduction of a rental bike scheme. It is important to have the right number of rental bikes available and accessible to the public at the right time to reduce waiting time. An essential part of this is the prediction of bike count required at each hour of the day, in order to ensure a stable supply of rental bikes.

The data for this assessment are taken from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand), and may be found on the Assessment page of the module VLE site.

The data are for each hour of each day, and relate to weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), and the number of bikes rented per hour.

2 Task

This assessment requires you to create a neural network classifier to predict the number of bikes required, given the hour of the day and the weather data for that hour.

After any experimentation and testing, your report must present and evaluate a final, trained network that can be used by others to predict the number of bikes required for previously unseen inputs.

You must report on this task in sections 1-3 of your report. You must then consider the related application of a neural network (described below) in section 4.

3 Data

For this assessment, you will use the data in the file `SeoulBikeData.csv`, a file that contains the following variables:

- Date: year-month-day
- Rented Bike count: Count of bikes rented at each hour
- Hour: Hour of the day
- Temperature: Temperature in Celsius
- Humidity: % humidity

- Windspeed: m/s
- Visibility: x10m
- Dew point temperature: Celsius
- Solar radiation: MJ/m2
- Rainfall: mm
- Snowfall: cm
- Seasons: Winter, Spring, Summer, Autumn
- Holiday: Holiday/No holiday
- Functional Day: NoFunc (Non Functional Hours), Fun (Functional hours)

Use of this data should reference these publications:

[1] Sathishkumar V E, Jangwoo Park, and Yongyun Cho. 'Using data mining techniques for bike sharing demand prediction in metropolitan city.' Computer Communications, Vol.153, pp.353-366, March, 2020

[2] Sathishkumar V E and Yongyun Cho. 'A rule-based model for Seoul Bike sharing demand prediction using weather data' European Journal of Remote Sensing, pp. 1-18, Feb, 2020

4 Report

You may use any appropriate software to create a neural network. It is not necessary to submit code for the neural network.

You must give a justification for your choice of network type and structure, and proper evaluation of its performance is more important than the ability of the network to produce "correct" results. You must evaluate thoroughly the performance of the network in terms of accuracy, precision, and so on as appropriate to the problem. If your network does not perform perfectly you will still get credit for attempting to explain the performance.

Where you make reference to appropriate literature to support your design and analysis, your answer should show that you have understood the literature you have selected, are able to explain the material, and can critically compare material from different sources.

Your report must be **no more than 8 A4 pages in length**, including any diagrams and references. You do not need a title page, but if you have one it is included in the page count.

The report must contain the following sections, in the order given below, and material outside the four prescribed sections will not be marked.

1. [20 marks] **Discussion of architectures**

   This section should:

   - describe (briefly) the data you have, and how much there is of it;
   - identify the type of problem;
   - identify which classes of architectures would be suitable;
   - give a brief discussion of the technical features of the architectures, and the advantages and disadvantages of each;
   - state which class of architecture you are going to use and justify your choice, relating the characteristics of the problem to the advantages/disadvantages of the architecture.

   To do this you might need to:

   - do some preliminary experiments with simple versions of the architecture to get a feel for what will work;
   - do some exploratory data analysis to see what the characteristics of the data are;
   - consider the principles involved and relate them to the problem.

2. [40 marks] **Creation and application of neural networks**

   This section should:

   - describe the chosen inputs to (and outputs from) the networks;
   - describe how the data you started with have been preprocessed;
   - give sufficient detail for someone else to process a new batch of data for use with the final trained network;
   - state which training algorithm you selected, and explain how you selected that training algorithm. For this training algorithm, give sufficient detail to enable someone to use the same training algorithm in exactly the same way. This does NOT mean (for example) describing gradient descent in great detail. It DOES mean giving any parameters, initialisation, etc, even if they are the toolbox defaults;
   - explain the process you went through in making the selection of the final architecture, for example, the number of neurons or the number of layers to use.

   To do this you might need to:

   - test one or more networks to demonstrate the effect of different preprocessing choices on the performance of the network;
   - try different training algorithms on one or more networks to compare performance;
   - evaluate a number of networks, and record details of their structures and how they performed. You may summarise repeated tests of the same structure.

3. [20 marks] **Results and evaluation**

   This section should:

   - explain the metric or metrics you have used for comparison between networks;
   - give a synopsis of the results obtained from the *final* selected network;
   - evaluate the results, in relation to the problem posed in the scenario.

   To do this you might need to:

   - consider different metrics for performance, appropriate to the problem. Remember that a Mean Squared Error (MSE) on its own is not always helpful in judging how well something works;
   - identify anything of interest in the results, such as areas of particularly good or

poor performance, or variation between different training runs;

- reflect on the conclusions that you may draw from the results, and whether they are showing that the neural network is useful in this case.

4. [20 marks] **Further application**

For urban transport planning, travel surveys are used to obtain critical information about individuals use of different modes of transport. Smartphone-based GPS travel surveys record travelers' GPS traces automatically. However, GPS records do not tell us what mode of transport was used at any time. In order to obtain this trip information, the GPS data needs to be labelled to indicate modes of transport (e.g. walk, bike, car, bus)

Assume we have a dataset representing a GPS traces for many individuals over several days. The data was captured at a 5s interval, and contains an index, date and time, latitude, longitude, altitude, speed and acceleration. The task now is for a neural network to identify the mode of transport in use at any particular time from an individual's tracking data.

In this section you do not need to create a neural network. You should discuss the issues you would need to consider in relation to:

- use of the data;
- selection of an architecture;
- construction of the network;
- evaluation of the network.

End of examination paper