COURSEWORK 2:
DATA SCIENCE PROJECT

Intention is to have you preform an end-to-end data science project.

This will be versioned properly and you will explore an aspect of ML techniques.

The project will be on a public Github and marked via this GitHub and an oral presentation.

This project should demonstrate your data science skills. It also provides you a data science project to link to a CV and a good project will enhance your employability.

COURSEWORK SUBMISSION DETAILS

Dates and deadlines:
Hand in Friday 19th December at 1pm.

Document to hand in:
Document to hand in on Moodle will be a .txt file with these details:

UP number:
Link to public github:
Dataset chosen (one of three given):
Research question 'Q3' chosen:
Date of last commit to the Github:
Last commit message to Github:

Notes about using other code:

As ever you may use any other resources that you wish, but your sources MUST be referenced properly - any copied code found from kaggle competitions, medium articles etc not referenced will result in deducted marks.

COURSEWORK DETAILS

You will answer three questions below using one of the test datasets of your choice.

Q1: Pick a traditional - non neural network approach to your problem. Why did you pick this approach and how well does a traditional approach do?

Q2: Repeat the answer with a Neural Network approach - compare the two approaches.

Q3: Pick <u>one</u> of the 'research questions' below and explore it with your dataset and neural network.

Research questions:

1. How does data augmentation affect the performance of a neural network?
2. How do parameters of a convolutional (or autoencoder) neural network affect performance of a network?
3. How does choice of activation function affect the performance of a neural network?
4. How does pre-training (and epochs and batches) affect the performance of a neural network?
5. How do choices about data such as i) amount of training data and ii) balance of classes in a classification problem affect the performance of a neural network?

Your code must be displayed in your github - <u>Github must be public</u>:

COURSEWORK DETAILS

What should the Github look like:

README - this explains your project idea and outlines the structure of the document and how to run it including what python dependencies you will need
dependencies.txt - this lists your python dependencies
py - this is a folder and your codes are structured under here
functions.py - a piece of code which includes helper functions you have written such as 'def get_data()' or 'def augment_data()' or def datasplit_train_test()' etc'
Q1_folder - code for answering Q1 lives here. The document to be marked is a single self contained ipynb (which can call your functions.py file) - tidy your notebooks like a 'medium' or 'toward data science' article which provides a tutorial of your method and explains choices made. This notebook should be aimed at a 'beginner ML audience'.
Q2_folder - code for answering Q2 lives here. The document to be marked is a single self contained ipynb (which can call your functions.py file) - tidy your notebooks like a 'medium' or 'toward data science' article which provides a tutorial of your method and explains choices made. This notebook should be aimed at a 'beginner ML audience'.
Q3_folder - code for answering Q3 lives here. The document to be marked is a single self contained ipynb (which can call your functions.py file) - tidy your notebooks like a 'medium' or 'toward data science' article but this time for an 'intermediate audience' - i.e. you can assume your audience has read notebooks Q1 and Q2.

COURSEWORK ASSESSMENT

Written (GitHub) part = 80%
Oral presentation = 20% (given in last week of term and timetabled)

Oral Rubric

Audience: Becky, Chris and Dan (no other audience)

8 mins total assessment

4 min presentation on Q3: 3 slides maximum - 50%

1. What was motivation for the dataset and question chosen?
2. What methods are used and why?
3. What are the conclusions?

4 min questions from examiners - 50%

1 min feedback

You will be marked only on substance NOT style, however, if you would like feedback on style to help you pre-masters presentation next term you will be given the choice.

COURSEWORK ASSESSMENT

Written Rubric

Quality of head README file documentation and dependencies document - 20%

- Github README files presentation - 5%
- Description of dataset (not just what but also where and why) - 5%
- Description of question / motivation - 5%
- All dependencies working (showing version of software used) - 5%

Github usage - 10%

- Use of branches for versioning - 5%
- Commit messages - 5%

Feature engineering - 10%

- Reasoning behind features chosen - 5%
- Cleaning and normalising of data and reasoning for this - 5%

Readability and usability of tutorial in Q1 and Q2 - 40% (20% per Q1, Q2)

- Structure of tutorial - 5% (enough but not too much information, appropriate for audience)
- Markdown and commenting - 5% (good code practices)
- Use and labelling of figures / images and markdown text content in notebook and ease of use for a beginner - 10%

Readability and usability of tutorial in Q3 - 20%

- Structure of notebook - 2.5%
- Markdown and commenting - 2.5%
- Depth of exploration of research question - 15%

GITHUB TIPS

Commit regularly with useful messages

Create different branches, e.g. dev, main

Play around and experiment in dev

Push to main when things work and are stable

For your final submission, make sure it is on main

DATASETS

Explore Kaggle for source datasets for your projects: https://www.kaggle.com

Tip: Find a dataset you are interested in, e.g. historical data, sports data, economic data

(You'll have more fun if you are interested in the result)

DO NOT COPY CODE FROM OTHER PARTICIPANTS ON KAGGLE!