Overview

In this work, you will need to present your own code that performs nested cross-validation and the k-nearest neighbour algorithm, build confusion matrices, and estimate distances between data samples.

The assignment consists of a single Jupyter notebook. A template is provided which provides further details on the task, and which you should adapt to include your solution. **Please refer to the notebook linked below for full instructions for this coursework.** You can import it into your own notebook server running on your own machine.

The purpose of the coursework is to help you:

- Get familiar with **common python modules / functions** used for ML in python

- Get practical experience **implementing** ML methods in python

- Get practical experience regarding **parameter selection** for ML methods

- Get practical experience on **evaluating** ML methods and applying cross-validation

Notes:

- don't use libraries that implement kNN or cross-validation. We want to see your code!

- Remember to comment all of your code (see here for tips: https://stackabuse.com/commenting-python-code/). You can also make use of Jupyter Markdown, where appropriate, to improve the layout of your code and documentation.

- Please add docstrings to all of your functions (so that users can get information on inputs/outputs and what each function does by typing SHIFT+TAB over the function name. For more detail on python docstrings, see here: https://numpydoc.readthedocs.io/en/latest/format.html)

- When a question allows a free-form answer (e.g. what do you observe?), create a new markdown cell below and answer the question in the notebook.

Plagiarism: please make sure that the material you submit has been created by you. Any sources you use for code should be properly referenced. Your code will be checked for plagiarism using appropriate software.

Submitting your work

Your completed Jupyter notebook should be run fully, saved, and then downloaded to your computer as an **.html** file (File-> Download As-> HTML). This is the file you should then upload below for marking.

As you will be graded based on this non-executable HTML, it is important that any outputs your code generates are clear and presentable. Please double check your html file in a web browser before submission.

Review Criteria

1. Data exploration (+ 2 questions) [10 marks]

2. Code, docu. & comments (KNN + Evaluation + NCV) [10 marks]

3. Results (KNN folds + Summary + Confusion matrices) [10 marks]

4. Final questions [10 marks]

5. Overall presentation quality (code and markdown) [10 marks]

We will typically grade your work into the following categories:

0,1,2,3: clearly a fail 4,5: acceptable but with obvious limitations 6,7: solid submission but does not go beyond what was taught 8,9,10: very impressive work that goes beyond what was taught