COLLEGE *of* ENGINEERING, SCIENCE *and* ENVIRONMENT
SCHOOL *of* INFORMATION *and* PHYSICAL SCIENCES
*The* UNIVERSITY *of* NEWCASTLE

# COMP3330/6380, Semester 1, 2023

## Individual Assignment: Introductory Machine Learning Project

| | |
|---|---|
| **Assessment Type** | Written Assignment |
| **Weighting** | 25% |
| **Length** | COMP3330 students: 4-8 pages excluding Appendix<br>COMP6380 students: 6-12 pages excluding Appendix |
| **Due Date** | Week 7 - Wednesday, 5 April 2023 (5 pm) |
| **Submission** | Online via Canvas / Turnitin |

## Description

Students work on classification tasks with ANNs and SVMs by using Python, scikit-learn, PyTorch, and other relevant Python packages. The submission includes:

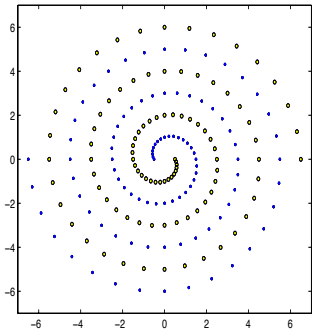
1. A report that addresses the questions specified below with written text, figures, and tables. The report is to be formatted in Springer LNCS format and exported to `pdf` format. The LNCS conference paper template is available at the following link: `https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines`. The report length in pages should fall in the given range. To save space, figures and/or tables can be placed into an Appendix.
2. All Python code used to produce the results discussed in the report as Supplemental Material. Marks may be deducted for any results in the report that cannot be reproduced from submitted code (allowing for some variation due to random model initialisation etc.).

The following pages contain the questions to be addressed in the report, and the marking criteria.

## Question 1: Variations of the Two-Spiral Task

Perform an experimental study on the following variations of the two-spiral task:

a) **ANN**: Start with the "original dataset" of Lang and Witbrock (1988) with 194 training points (see Figure on the right). How fast and how well can you solve this task using a feed-forward NN? The $(x, y)$-coordinates of the points in the dataset will be supplied in Canvas.




b) **ANN**: Generate a variation of the 2-spiral task. Then solve the associated classification task using ANNs and discuss your approach and solution in comparison to a). You may consider, e.g., to generate three spirals or four spirals, or use code such as available at `https://gist.github.com/45deg/e731d9e7f478de134def5668324c44c5`

c) **ANN vs. SVM**: Run experiments that allow to compare how well ANNs and SVMs can solve the two classification tasks above, and discuss the outcomes.

For each subquestion try out different architectures, hyperparameters, and methods. Compare and discuss their performance (e.g., speed, generalisation, convergence). The performance of the solutions can be evaluated by visual inspection of a generalisation test applied to all pixels of a section of the $(x, y)$-plane (that for the 2-spiral data should result in two intertwined spiral shaped regions). You may also think about alternative performance measures, e.g., considering false positives and false negatives, confusion matrices, learning curves etc.

## Question 2: BUPA Liver Disorders Dataset

The BUPA Liver Disorders Data Set was collected by BUPA Medical Research Ltd. in the 1980s for research on liver health. It has since been donated to the UCI Machine Learning Repository and is available from:

`https://archive.ics.uci.edu/ml/datasets/liver+disorders`

The dataset contains 345 samples of five explanatory variables and one target variable. The explanatory variables represent various blood markers and come as integers, but can be regarded as continuous features. The target variable is the self-reported number of alcoholic drinks consumed per day. There is also a seventh variable distributed with the dataset, which represents the train/test split chosen in the original research. This variable has often been mistakenly interpreted as the target class (McDermott and Forsyth, 2016) and can be disregarded here.

a) Prepare the data for a binary classification task by dichotomising the original target variable into two classes using a cut-off value (e.g., 3 drinks/day, such that the two classes are relatively balanced). You will also normalise the explanatory variables, which may help with model convergence later on.

b) Train different classifiers and report the maximum accuracy that you can achieve in a 10-fold metric cross-validation experiment. For solving this you can train an SVM and/or a Neural Network, or some combination. Your experiments should also investigate what impact normalisation or some alternative forms of preprocessing of the explanatory variables has on the results.

## Marking Criteria

| Criterion | Marks | Description |
|---|---|---|
| **Question 1 a)** | 4 | Experimentation with different ANN parameters (e.g., number of layers, batch size, etc.) based on suitable performance metrics (e.g., learning curves, confusion matrices, visualisations, etc.). This process should be documented, including plots and tables. Overall findings should be discussed and interpreted. |
| **Question 1 b)** | 4 | At least one variation of the Two-Spiral Task should be investigated. Experimentation with different ANN parameters (e.g., number of layers, batch size, etc.) based on suitable performance metrics (e.g., learning curves, confusion matrices, visualisations, etc.). This process should be documented, including plots and tables. Overall findings should be discussed and interpreted. |
| **Question 1 c)** | 3 | Experimentation with different SVM hyperparameters (e.g., C, Gamma) based on suitable performance metrics (e.g., learning curves, confusion matrices, visualisations, etc.) and comparison with the ANN approaches. This process should be documented, including plots and tables. Overall findings should be discussed and interpreted. |
| **Question 2 a)** | 1 | The dataset should be prepared appropriately, including dichotomising the target variable and identifying as well as normalising or otherwise preprocessing the explanatory variables. This process should be summarised. |
| **Question 2 b)** | 6 | Experimentation with different classifiers (e.g., ANN, SVM), corresponding hyperparameter choices (e.g., number of layers, batch size, C, etc.), and data preprocessing methods based on suitable performance metrics (e.g., learning curves, confusion matrices, visualisations, etc.). This process should be documented, including plots and tables. Overall findings should be discussed and interpreted. |
| **Report format** | 2 | The report should conform to the Springer LNCS format. Moreover, it should be clearly structured and well organised to be easily readable. The report length in pages (excluding Appendix) should be 4-8 pages for COMP3330 and 6-12 pages for COMP6380. Be aware that any work after the maximum page limit may not be included in marking. |
| **Python code** | 2 | The Python code used to produce the results discussed in the report is to be submitted as supplemental material. The code needs a clear structure and has to be free of errors. Use comments to document what is done and to structure the code. Marks will be deducted for poor documentation, structure, and reproducibility of the results discussed in the report. |
| **Writing and referencing** | 3 | Results and findings should be communicated clearly, accurately, and concisely. This includes correctness of grammar and spelling, and appropriate use of references, direct quotes, paraphrasing, etc. For COMP6380 students the citations and report presentation are expected to be of high professional standard. Students should note that there is a fine line between poor referencing and plagiarism. Reports that appear to be plagiarised will be referred to the Student Academic Conduct Officer, with possible outcomes such as a mark of zero for the entire assignment. Students are strongly advised to repeat the University's Academic Integrity Module, and to be sure never to take text or ideas from anywhere without clearly noting the source. |

## Literature

S. K. Chalup, and L. Wiklendt. Variations of the Two-Spiral Task. *Connection Science 19*(2), pp. 183-199, June 2007. Available at http://hdl.handle.net/1959.13/808886

K. J. Lang and M. J. Witbrock. Learning to tell two spirals apart. In: Touretzky, D., Hinton, G., Sejnowski, T. (Eds.), *Proceedings 1988 Connectionist Models Summer School*. Morgan Kaufmann, Los Altos, CA, pp. 52–59, 1988.

McDermott J, Forsyth RS. Diagnosing a disorder in a classification benchmark. Pattern Recognition Letters, 73 pp. 41-43, 2016.