SCHOOL *of* ELECTRICAL ENGINEERING & COMPUTING
FACULTY *of* ENGINEERING & BUILT ENVIRONMENT
The UNIVERSITY *of* NEWCASTLE

Comp3330/Comp6380 Machine Intelligence, Semester 1, 2020

Project A and Written Assignment A: Introductory Machine Learning Project

Deadline: Week 7 (check date and time and submit via blackboard)

Maximum possible marks: 20 = 10 (group project part) + 10 (individual paper)

Description

In this assignment we want to gain basic experience in testing out ANNs and SVMs for classification. There are two components for marking this assignment. Both components have to be submitted separately in Blackboard:

Project 1A [10 marks]. The group's experimental work should be summarised in a brief group summary report that provides an overview what your team has done in the experimental component of your project. It should explain how the software and data is structured and detail your group's individual member contributions to the project outcome. Marked will be the quality, volume and depth of the experimental results achieved by the group. If the member contributions are very different the marks will be weighted accordingly. The recommended length of the group summary report is about 4 pages. Include in your submission the summary report and all essential files that are required for reproducing and verifying your results. It is sufficient if one team member submits. Make sure that all other group member names are listed and each member has signed (!) the agreed group report with individual member contribution statements.

Report 2A [10 marks]. Each student writes an individual paper that describes, analyses and critically discusses the project results achieved by the student and his/her team in detail. We expect about 5-8 pages from COMP3330 students, and about 8-12 pages from COMP6380 students. Aim at providing a high quality report that describes and discusses your results and approaches from part 1A clearly and concisely following instructions of the individual questions Q1, Q2, Q3 below. Your individual report should be formatted in Springer LNCS format. Any literature citations, e.g., in the background, method or discussion sections should follow a consistent citation style. The LNCS conference paper template is available e.g. at the following link: `https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines` Please submit your individual paper in pdf format in Blackboard.

Notes: Be prepared that training ANNs can require some time. We recommend using Python and scikit-learn. However, any language/library combination is acceptable while it is expected that you are able to acquire the necessary details how to use the software or programming language of your choice from relevant on-line help or literature. Plot error curves that indicate convergence times (how many iterations did it take?). For demonstrating how well your trained ANN model generalises you can visualise the results of your tests (you can submit several plots from different networks or different training schemes) or you may consider suitable statistical

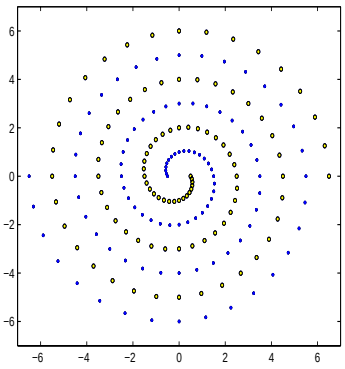
measures. Always discuss your results and highlight the most important outcomes.

Warning: You will find that some of the questions can lead into open-ended research and some of the experiments may take significant time on the computer. It is your responsibility to decide on a sensible balance of quality and depth of your investigation of each individual question so that the assignment can be completed within the given time. If computing resources become a severe issue provide pilot results as proof of concept, explain the issue and solution attempts and put more emphasis on the analysis and discussion.

Q1 Variations of the Two-Spiral Task [total 3 marks]

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

a) (ANN training): Start with the "original dataset" of Lang and Witbrock (1988) with 194 training points (see Figure below). 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 blackboard.) [0.5 marks]



b) (ANN training): Generate a variation(s) 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 to generate the spirals using code such as available at `https://gist.github.com/45deg/e731d9e7f478de134def5668324c44c5` [1 mark]

c) (ANN vs. SVM): Critically compare ANNs and SVMs on solving the two classification tasks above and discuss the outcome. [1.5 marks]

For each subquestion try out different architectures, parameters, and methods. Compare and discuss their performance (speed, generalisation). It is recommended that you focus for each part of your experiments on *about two* different aspects that you investigate in more detail (this could be e.g. variation of the step size, number of hidden layers/units, use of momentum, different kernels or kernel parameters in SVMs, ...). 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.

A background paper with literature links, description of the data and some hints about successful network architectures is, for example, the following survey (Chalup and Wiklendt, 2007).

Q2 Statlog (Landsat Satellite) Data Set [3 marks]

The data consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification associated with the central pixel in each neighbourhood. The aim

is to predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number.

CLASS

There are 6 decision classes: 1, 2, 3, 4, 5 and 7. There are no examples associated with with class 6 in this dataset - they have all been removed because of doubts about the validity of this class.

1: red soil

2: cotton crop

3: grey soil

4: damp grey soil

5: soil with vegetation stubble

6: mixture class (all types present)

7: very damp grey soil

NUMBER OF EXAMPLES

- training set = 4435
- test set = 2000

NUMBER OF ATTRIBUTES

36 (= 4 spectral bands x 9 pixels in neighbourhood)
The attributes are numerical, in the range 0 to 255.

DOWNLOAD

The data is available at the UCI Machine Learning Repository

`https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29`

TASK

Your task is to use the training data to train your classifier and report the maximum accuracy that you can achieve on the test data. Document the process of researching and creating this classifier. For solving this you can train a SVM and/or a Neural Network, or some combination. Discuss how well your classifier performs on training and test data by using some suitable form of metrics e.g. considering false positives and false negatives, confusion matrices, learning curves etc.

Acknowledgment:
Ashwin Srinivasan
Department of Statistics and Data Modeling
University of Strathclyde
Glasgow

Scotland
UK
ross '@' uk.ac.turing

The original Landsat data for this database was generated from data purchased from NASA by the Australian Centre for Remote Sensing, and used for research at:
The Centre for Remote Sensing
University of New South Wales
Kensington, PO Box 1
NSW 2033
Australia.

Q3 Select Your Own Data [total 4 marks]

For this question please perform a comparison study of SVMs and ANNs on one data set of your choice (for COMP3330 students) or two data sets (for COMP6380 students). You can find data sets e.g. at:

- UCI repository `https://archive.ics.uci.edu/ml/datasets.php`
- Kaggle `https://www.kaggle.com/datasets`

a) Submit your full study with all specifications so that the marker is able to verify it.

b) Describe and discuss your approach in a concise report that is detailed enough to allow your solution to be replicated. Include a detailed analysis of your classifier.

Note

Marks will be awarded for the performance of the classifier, evidence of researching better solutions for the classifier, and evidence of understanding the training process and the effects of the various training parameters. For details please consult the marking guides that will be provided separately. Depending on the configuration of your solution you may be asked to give a demo to the tutors for evaluation. If you have any questions about the specific submission format of your solution please consult with the tutor. Make sure you submit before the deadline.

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.

T. Mitchell. Machine Learning, McGraw Hill, 1997.