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

Comp3330/Comp6380 Machine Intelligence, Semester 1, 2019

Project 1b: Machine Learning Project

Deadline: 30 April 2019, (submit via blackboard)

Maximum possible marks: 12

Description

In this assignment we want to gain basic experience in testing out ANNs and SVMs for classification. The main part for marking this assignment is a report and the quality of the experimental results. The recommended length of the report is: about 4-12 pages for COMP3330 students, and about 6-14 pages for COMP6380 students (depending on teamsize). Include all files in your submission that are required for verifying your results. Aim at providing quality results and describe and discuss them clearly and concisely in your report following instruction of the individual questions below.

Be prepared that depending on your architecture training the ANNs might require some time. We recommend using Python and scikit-learn. However, any language/library combination is acceptable but 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 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.

This assignment can be done in teamwork with other students from this class (1-5 people per team) and we encourage you to do this. Best you include a statement agreed by all team members about who contributed what. Any additional help that you use also has to be explicitly acknowledged in your submission.

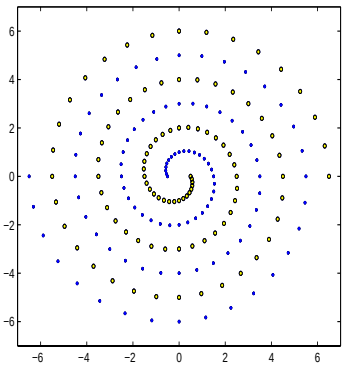
Warning: You will find that some of the questions can lead into open ended research and some of the experiments may take significant time on the computer. It is your responsibility to decide on a sensible balance of quality and depth of your investigation of each individual question so that the assignment can be completed within the given time.

Please submit your assignment electronically via the assignment section in blackboard. Include all relevant software, results, and data. Please consult your course outline for additional information and let us know if you would have any questions or if anything would require further clarification.

Q1 Variations of the Two-Spiral Task [total 4 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 series of variations of the 2-spiral task that starts with a lot of points and then reduces the number of points i.e. the data set becomes thinner step-by-step for each experiment. You could run, for example, 5 experiments, where the first uses 500 points and the last uses 20 points. Then solve the associated classification task using ANNs and discuss your approach and solution in comparison to a). We would like to see how many points are necessary so that it still works. [1.5 marks]

c) (ANN vs. SVM): Compare ANNs and SVMs on solving the two classification tasks above. [2 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 (Shuttle) Data Set [4 marks]

The shuttle dataset contains 9 attributes all of which are numerical. The first one being time. The last column is the class which has been coded as follows :

1: Rad Flow

2: Fpv Close

3: Fpv Open

4: High

5: Bypass

6: Bpv Close

7: Bpv Open

Approximately 80% of the data belongs to class 1. Therefore the default accuracy is about 80%. The aim here is to obtain an accuracy of 99 - 99.9%. Dataset contains 58000 instances and is divided into training (43500) and test (14500) subsets.

The data is available at the UCI Machine Learning Repository

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

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 or a Neural Network, or some combination. Discuss how well your classifier performs on training and test data by using some suitable form of cross-validation, considering false positives and false negatives or confusion matrices.

Acknowledgment: Thanks to Jason Catlett of Basser Department of Computer Science, University of Sydney, N.S.W., Australia for providing the shuttle dataset. Thanks also to NASA for allowing us to use the shuttle datasets.

Q3 Select Your Own Data [total 4 marks]

For this question please perform a comparison study of SVMs and ANNs on a data set of your choice. You can find data sets e.g. at:

- UCI repository `https://archive.ics.uci.edu/ml/datasets.html`
- 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. 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.