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

Comp3330/Comp6380 Machine Intelligence, Semester 1, 2017

Homework Assignment 1: Machine Learning for Data Classification

Deadline: 11 April 2017, 1pm (submit via blackboard)

Maximum possible marks: 10

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.

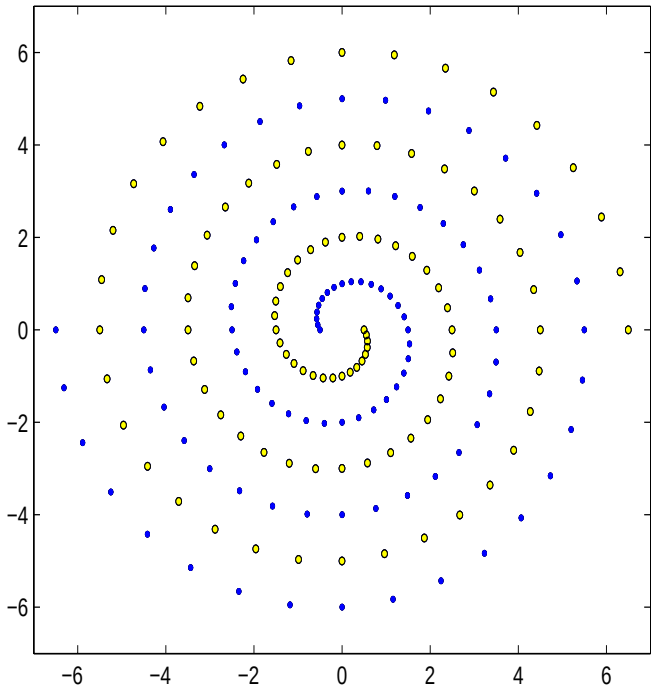
Warning: You will find that some of the questions can lead into open ended research and some of the experiments may take significant time on the computer. It is your responsibility to decide on a sensible balance of quality and depth of your investigation of each individual question so that the assignment 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 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 3-spiral data set. Show and discuss how well your task can be solved with a suitable ANN. [0.5 marks]

c) (ANN training): Generate your own variation of the 2-spiral task (e.g. use a different type of spiral or more or less points). Then solve the associated classification task using ANNs and discuss your approach and solution in comparison to a). [1 mark]

d) (ANN vs. SVM): Compare ANNs and SVMs on solving the three classification tasks in a) - c). [1 mark]

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 Wearable Computing Classification task [3 marks]

A recent study (Ugulino, W. et al., 2012) provided a public domain dataset for human activity recognition. It contains recordings of 4 subjects performing activities of daily living while wearing accelerometers mounted on their waist, left thigh, right arm, and right ankle. The data set is relatively large with 165633 records. Each record has 8 attributes that correspond to user behaviour data.

There are six classes of different behaviours:

1. SITTING-DOWN
2. STANDING-UP
3. STANDING
4. WALKING
5. SITTING

You will be provided with a version of the data on Blackboard. The original data is available at the UCI Machine Learning Repository

http://archive.ics.uci.edu/ml/datasets/Wearable+Computing

Your task is to submit the most successful classifier you can create, and document the process of researching and creating this classifier. I.e. the aim is to have a function where the input is a 18-attribute-long data record or a part of it and the output is the correct class number (1-5). For solving this you can train a SVM or a Neural Network, or some combination (in this case please provide code to load and run it on another dataset). Discuss how well your classifier performs e.g. by using some suitable form of cross-validation.

Q3 Captcha Data Classification task [total 4 marks]

Develop a solution of the captcha character classification task (data provided on Blackboard). These characters are generated from a real Captcha algorithm which distorts the letters, moves them, converts them to outlines, blurs them, and draws a line through them. We have generated 20 samples for each character, and randomly split the dataset in half. You will use half of this data (available on blackboard), and the quality of your solutions will be tested against the other half (this secret set is locked away and only the tutors will have access when they evaluate your solution). Not all letter samples have been split evenly in this distribution, but you will have at least 3 of each letter. This data is indicative only and may not contain the full range of distortions for each character.

Only alphabet characters are used, and these may be offset by 20 or more pixels in any direction. An approximate horizontal outline is used for the characters to further confuse the letter shapes and lines. Lines are drawn with random curvature, and up to 10 pixels of horizontal and/or vertical stretching is performed. Finally the letters are distorted by a non-linear stretch and blurred slightly. Some real world Captcha systems have been cracked using similar approaches once the letters are separated. Only a small number of examples are given, and you must decide how best to use the information for training. This data is indicative only and may not contain the full range of distortions for each character.

a) Submit your full solution with all specifications so that the marker is able to verify it (you are allowed to submit two (!) solutions e.g. one ANN and one SVM or e.g. two ANNs).

   For marking we will evaluate all submitted solutions on the secret test set which is the other half that was randomly selected from the full original dataset and put aside for testing and marking. Please remember that ANNs will initialize randomly every time, so the same training process may produce different performance on separate experimental runs. [A ranking of all submitted solutions will determine the fraction of 2.5 marks awarded for this subquestion. Only your best solution will count.]

b) Describe and discuss your approach in a concise report that is detailed enough to allow your solution to be replicated. You should also include a description of what choices you made and how you came up with your solution. [1.5 marks.]

Note

To save SVMs and Neural Networks for submission, use the following code (if using Python and sklearn):

Listing 1: Saving A Trained Classifier

```
import pickle
pickle.dump(clf, open('file_out', 'wb'))
```

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

W. Ugulino, D. Cardador, K. Vega, E. Velloso, R. Milidiu, H. Fuks. Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements. Proceedings of 21st Brazilian Symposium on Artificial Intelligence. Advances in Artificial Intelligence - SBIA 2012. In: Lecture Notes in Computer Science. , pp. 52-61. Curitiba, PR: Springer Berlin / Heidelberg, 2012.

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.