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

Comp3330/Comp6380 Machine Intelligence, Semester 1, 2015

Homework Assignment 1: Machine Learning for Data Analytics

Deadline: Week 7 (Monday 20 April 2014, 23:59)

Maximum possible marks: 10

Description

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 4-10 pages for Comp3330 students, and about 6-12 pages for Comp6380 students. 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 the pyBrain library to implement your neural network, rather than attempting yourself in Python or another language. Although you were briefly introduced to pyBrain in the lab sessions 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 if you choose to go in that direction. 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 basic 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 (1-3 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.

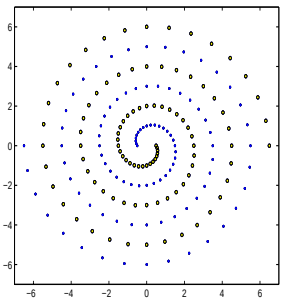
Warning: You will find that some of the questions can lead into open ended research and some of the experiments may take significant time on the computer. It is your responsibility to decide on a sensible balance of quality and depth of your investigation of each individual question so that the assignment 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 let us know any questions or if anything requires further clarification.

Q1 Variations of the Two-Spiral Task [2 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? (dataset will be supplied in blackboard) [10% of Q1 mark]



b) (ANN training): Generate your own variation of the 2-spiral task. Then solve the associated classification task using ANNs and discuss your approach and solution in comparison to a). [20% of Q1 mark]

c) (ANN training): Generate a 4-spiral data set. Show and discuss how well your task can be solved with a suitable ANN. [20% of Q1 mark]

d) (ANN vs. SVM): Compare ANNs and SVMs on solving the three classification tasks in a) - c). [50% of Q1 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).

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

Q2 Autoencoder [2 marks]

Generate a dataset that uses sparse coding to encode the numbers 0-15:

0 = (1,0,0,...,0)
1 = (0,1,0,...,0)
...
15 = (0,0,0,..., 1)

Train a 16-H-16 multilayer perceptron (i.e. 16 input units, a hidden layer of H units and an output layer of 16 units) on the identity function that maps 0 to 0, 1 to 1, ..., 15 to 15.

Part I: Determine experimentally what is the minimal number of hidden units, H, required for training the network successfully (Hint: Check chapter 4 of the book (Mitchell, 1997)). Try different training algorithms.

Part II: Conduct training experiments using 16-H1-H2-16 ANNs with two equally sized hidden layers. Determine experimentally what is the minimal number of hidden units in H1 and H2 required for training the network successfully. How does it compare to the above experiments with your 16-H-16 multilayer perceptron?

In your report describe what you did in the experiments and what was the outcome. Finally discuss what role the hidden layers plays in this experiment and what role this type of network could possibly play in real applications.

Postgraduate question [Counts 50% of Q2 for postgraduates]
Conduct the same experiment using input/output layer sizes of 4, 8, 32, 64, ... instead of 16 and report and discuss performance differences.

Q3 Multi-Class Classification for Human Activity Recognition Using Smartphones Data Set [6 marks]

A recent study (Anguita et al., 2013) provided a public domain dataset for human activity recognition. It contains recordings of 30 subjects performing activities of daily living while carrying a waist-mounted smartphone with embedded inertial sensors. The data set is large with 10299 records. Each record has 561 attributes that correspond to user behaviour data recorded during a fixed-width sliding time window.

There are six classes of different behaviours:

1. WALKING
2. WALKING_UPSTAIRS
3. WALKING_DOWNSTAIRS
4. SITTING
5. STANDING
6. LAYING

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

https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

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 561-attribute-long data record or a part of it and the output is the correct class number (1-6). 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:

1. How accurate is the classifier and how well does it generalise?
2. How fast is it and could it be used on a smartphone?

Postgraduate question [Counts 17% of Q3 for postgraduates]
Discuss if/how it may be possible to identify individuals within the group of 30 subjects.

Note

To save SVMs and Neural Networks for submission, use the following code:

Listing 1: Saving A Trained Neural Network

```
import pickle
pickle.dump(neural_net, open('myneuralnet','w'))
```

Listing 2: Saving A Support Vector Machine Model

```
from svm_util import svm_save_model
svm_save_model('mysvm',trained_svm)
```

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.

Literature

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013*.

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.