Department of Computer Science
University of Pretoria

Artificial Intelligence II
COS 711

Assignment 1

August 14, 2017

1 Objectives

This assignment aims to achieve the following general learning objectives:

- To gain practical experience in the use of the self-organising feature map (SOM);
- To gain experience in simple data preparation for machine learning and SOMs;
- To gain practical experience in the empirical comparison of machine learning approaches;
- To gain experience in formal, scientific report writing.

2 Plagiarism Policy

The Department of Computer Science considers plagiarism as a serious offence. Disciplinary action will be taken against students who commit plagiarism. Plagiarism includes copying someone else's work without consent, copying a friend's work (even with consent) and copying textual material from the Internet. Copying will not be tolerated in this course. For a formal definition of plagiarism, the student is referred to `http://www.ais.up.ac.za/plagiarism/index.htm` (from the main page of the University of Pretoria site, follow the *Library* quick link, and then click the *Plagiarism* link). You may use any third party software, tools or packages, as long as their copyright does not prohibit their use for such purposes. If you have any form of question regarding this, please ask one of the lecturers, to avoid any misunderstanding.

3 Submission Instructions

You will have to write a report for this assignment. The report should be in standard PDF format, preferably compiled using LaTeX. You will have to submit only the report. No additional files of any sort should be submitted. Do not submit files in any other format other than PDF. Failure to follow any of these instructions will result in a zero mark for the assignment.

Upload just your PDF file (named `s99999999.pdf`, where `99999999` is your student number), to the appropriate assignment upload on the course website. Multiple uploads are allowed, but only the last one will be marked. The deadline is **15 September 2017, at 23:00**.

4 Analysis

You must analyse the performance differences between different implementations of the self-organising map (SOM) on real-world data sets. Your analysis must meet the following requirements:

- You must compare the performance of the standard stochastic SOM [3], the one of the batch SOM implementations (either the original batch SOM [3] or the fast batch map [2] are commonly used approaches) when used for a simple classification task on a set of at least three benchmark data sets.
- You may select any three data sets from the UCI Machine Learning Repository [1] on which to perform your analysis.
- The simple classification task works as follows (see [4] for further details):
  - First train the chosen SOM algorithm on a benchmark data set.
  - Then label the nodes of the SOM. Use the example-centric neuron labelling approach [5] to attach a label to each neuron in the trained map.
  - Once labels have been applied to neurons, each pattern (tuple) from the data set can be mapped to its best matching unit (BMU). The pattern in question is given the classification of the BMU's label. Note that it is possible for a pattern to map to a BMU with no label, in which case the pattern is not classified.
  - Classification performance can be measured in various ways. Most measures are variations on the percentage of misclassified data patterns.
- Note that the preferred means of experimentally comparing classification algorithms is by means of a k-fold cross-validated set of experiments. Many sources discuss the application of k-fold cross-validation. Select statistical tests that are appropriate for your experimental configuration.

5 Report

You must write a report describing the analysis you performed in section 4. The report should be of an academic nature. This means that:

- The tone of language should be formal and scientific, and must use correct spelling and grammar;
- Any figures or graphs included should be clear and of a professional standard;
- Label all appropriate parts of graphs and other data visualisations, so that they can be easily interpreted (it is usually not sufficient to simply paste a screen capture of a visualisation from a data analysis program or package);
- The report's structure should include all the aspects typically required of an academic paper (these include a title, abstract, introduction, methodology discussion, results, and conclusions);
- An adequate background discussion should be provided (this means that you must broadly discuss every technique you use, and provide a reference to published sources that describe the techniques — it is **not** acceptable to cite the course slides or Wikipedia);
- You must describe all the data preparation steps you performed on the data set, and justify why you performed these steps;
- Provide full details on your experimental procedure;
- Provide a conclusion in which you summarise your findings, and critically discuss the outcomes of your analysis;
- Adequate references must be provided (this means that you must properly cite all techniques that you discuss, and all the reference details must be correct). Your references should be to published academic work (references to slides or websites, for example, are not appropriate).

It is recommended that you consult several existing conference and journal papers, as a guide to the type of style you should adopt. There are many such sources freely available online.

Your report should be concise and to the point. Avoid discussing irrelevant details, but make sure that you describe all the details of your analysis. Justify each choice you make, even if it seems obvious to you.

It is very strongly recommended that you use the LaTeX template for IEEE conference papers (available as `IEEEtran.zip` in the folder for this assignment, on the course website), to typeset your paper. If you use this template, you should aim for a report (including references) of no more than 10 pages. Do not pad your report to reach this limit — shorter reports are preferred over longer reports that do not focus on relevant details.

6 Marking

The following general breakdown will be used during the assessment of this assignment:

Category	**Mark Allocation**
Report structure background content, references, style, spelling, and grammar	30 marks
Results and conclusions	70 marks
TOTAL	**100 marks**

References

[1] D. W. Aha, C. L. Blake, S. J. Hettich, E. J. Keogh, C. J. Merz, and P. M. Murphy. UCI repository of machine learning databases, 1998. University of California, Irvine, Department of Information and Computer Sciences, Irvine, California, United States of America. `http://archive.ics.uci.edu/ml/index.php`.

[2] S. Kaski, J. Venna, and T. Kohonen. Coloring that reveals cluster structures in multivariate data. *Australian Journal of Intelligent Information Processing Systems*, 6:82–88, 2000.

[3] T. Kohonen. *Self-Organizing Maps*. Springer, second edition, 1997.

[4] W. S. van Heerden. Self-organizing feature maps for exploratory data analysis and data mining: A practical perspective. Master's thesis, University of Pretoria, 2017.

[5] W. S. van Heerden and A. P. Engelbrecht. An investigation into the effect of unlabeled neurons on Self-Organizing Maps. In *Proceedgins of IEEE Symposium Series on Computational Intelligence*, pages 823–830, 2016.