Collective Report A

March 15, 2018

1 Unsupervised learning

This report summarises the work done in Labs 4 and 5. You will be working with the same data set of hand written digits as in the lab, a reduced version of the MNIST database (`http://yann.lecun.com/exdb/mnist/`). The data set is available in both cvs and matlab format from the course web page page. Please note that for educational purposes only, labels are provided. In general, this is not the case for problems where we use unsupervised learning approaches. Therefore your replies should **not** make use the labels as part of the arguments.

1.1 PCA

The task is to perform a Principle Component Analysis on the MNIST reduce data. You may want to keep the most important components and discard the others. Then plot 3D combinations of them and decide which is more informative for distinguishing the clusters that are formed. Please address the following points:

1. A brief description of the algorithm you implemented.
2. A 3D figure of the transformed data with the 1st, 2nd and 3rd PC as axes.
3. A 3D figure of the transformed data with the 2nd, 3rd and 4th PC as axes.
4. A 3D figure of the transformed data with the 1st, 3rd and 4th PC as axes.
5. A 3D figure of the transformed data with the 2nd, 3rd and 5th PC as axes.
6. A 3D figure of the transformed data with the 3rd, 4rd and 5th PC as axes.
7. How many clusters do the data form? Which of the figures is more meaningful? Please justify your response.

1.2 Competitive learning

Your task is to implement the standard competitive learning algorithm on a one layer network and use it to classify the set of hand-written digits. You may choose the number of output units such that you can capture all digits and tune the network such that it will learn quickly and result in as few dead units as possible. To do this, you will have to implement a method for detecting dead units among the prototypes. Please note: do not forget to normalise the weight vector and the data at the beginning of the process. Please provide the following:

1. A brief description of the algorithm you implemented. Is it batch or online? What is the difference between the two?

2. Why is it important to normalise data and weight vectors for competitive learning to work? What could happen otherwise?

3. A description of the technique(s) you used to optimise the network and what you achieved by using them. Provide a measurement of improvement (i.e. How many dead units you have on average. Perform statistics over at least 10 different initial conditions).

4. Your method for identifying dead units without the need of visualising the prototypes.

5. A figure showing the average weight change as a function of time. When has your network sufficiently learned from the data? For instance, if you implement an online version of the rule you may use a running mean to produce a smooth curve of the weight changes through time. Such a curve might be more informative on semi-log or log-log axes.

6. A figure of the prototypes and a comment on what they represent. How many prototypes did your network find?

7. The correlation matrix of the prototypes. How can you use this information to find similarities between the prototypes?

2 Report

This is an individually written report of a scientific standard, i.e. in a journal-paper like format and style. It is recommended that you use the web to find relevant journal papers and mimic their structure. Results should be clearly presented and justified. Figures should have captions and legends. Your report should **NOT exceed 6 pages** (excluding Appendix and Cover Page). Additional pages will not be assesed. Two-column format is recommended, the minimum font size is 11pt and margins should be reasonable. Kindly note that the readability and clarity of your document plays a major role in the assessment.

In the report you should include:

1. A response to all points requested by the assignment (including graphs and explanations). It is suggested to adopt a similar numbering scheme to make clear that you cave responded all questions.

2. An Appendix with snippets of your code referring to the algorithm implementations, with proper comments/explanations.

3. A proper description of how your results can be reproduced, see also “Important Note”.

4. A description of how in your view the two techniques may be used in combination to improve results.

Important Note: Please make sure that your results are reproducible. Together with the assignment, please upload your code well commented and with sufficient information (e.g. Readme file) so that we can easily test how you produced the results. If your results are not reproducible by your code (or you have not provided sufficient information on how to run your code in order to reproduce the figures), the assessment will not receive full points.

3 Suggested language

The recommended programming language is Python or Matlab. However, on your own responsibility, you may submit the project in any language you wish, but you should agree it beforehand with the Professor.

4 Marking

Assignments will be marked in a 0-100 scale. Results and explanations contribute for up to for 60 points, scientific presentation and code documentation for up to 20 points and originality in modelling the task for up to 20 points.

To maximise your mark, make sure you explain your model and that results are supported by appropriate evidence (e.g. plots with captions, scientific arguments). Figures should be combined wherever appropriate to show a clear message. Any interesting additions you may wish to employ should be highlighted and well explained.

5 Submission

The **deadline** for handing in **one copy** of the assignment to the Department of Computer Science Office (first floor, Regent Court, Portobello) is **Monday 23 April 2018, 3pm**, **please note the deadline extension**. A **second copy** of the assignment (in PDF format) and the corresponding code should be uploaded in MOLE prior to the deadline, with appropriate amount of detail for running your code and reproduce your results.

6 Plagiarism, and Collusion

You are permitted to use Matlab or Python code developed for the lab as a basis for the code you produce for this assignment with **appropriate acknowledgement**.You may discuss this assignment with other students, but the work you submit must be your own. Do not share any code or text you write with other students. Credit will not be given to material that is copied (either unchanged or minimally modified) from published sources, including web sites. Any sources of information that you use should be cited fully. Please refer to the guidance on “Collaborative work, plagiarism and collusion” in the Undergraduate or MSc Handbook.