Deep learning Assignment 1 (2023)

MPhil in Computational Biology

December 1, 2023

If there are errors found, I will update the assignment at: https://github.com/sje30/dl2023/.

Due date: 2024-01-23 23:45

Please submit your report to moodle, anonymised as before. Your report must be a maximum of twenty pages, excluding the appendix. Your appendix should contain only a copy of any code.

This assignment is worth 100% of your overall mark for this module.

1 Self-supervised learning [25 marks]

Build a multi-layer perceptron with the following specification: eight input units, three hidden units (sigmoidal activation) and eight output units (sigmoidal activation). There are eight 8-d input vectors $\mathbf{x}$, for each of the eight cases where one input unit is "1" and the other seven units are "0". Your task is to solve the self-supervision problem of reproducing the input vector on the output vector, so $\mathbf{t}^\mu = \mathbf{x}^\mu$ for $\mu \in [1, 8]$. Describe how your network attempts to solve this problem.

Demonstrate how an autoencoder can implement image compression (Cottrell et al., 1987). How good is your method, and how do your results vary with the number of hidden units in the bottleneck layer?

For this exercise, you should write your own code. you can adopt `https://github.com/sje30/dl2023/blob/main/code/xor/xor.R` if you wish.

2 Hopfield Networks [25 marks]

Construct a Hopfield network (Hopfield, 1982) with binary activations (+1 or -1), and test its ability to recall binary input patterns. Comment on the following features of your model:

1. Storage capacity: how many patterns can it store? How does the sparseness (fraction of units set to +1 rather than -1) of the pattern affect this result?
2. Robustness: how is pattern recall affected by the random loss of weights?
3. **(Advanced:)** Explore an alternative method for setting the weights and see how it affects network performance.

Consult Chapter 42 of David Mackay's book `https://www.inference.org.uk/itprnn/book.pdf` as well as the course notes to get started.

3 t-SNE, PCA and UMAP [25 marks]

1. Summarise the steps involved in Principal Components Analysis (PCA) for dimensionality reduction.
2. Summarise the steps involved in the t-SNE algorithm (van der Maaten and Hinton, 2008).
3. Summarise the steps involved in the UMAP algorithm (Becht et al., 2018).
4. Document and compare the ability of PCA, t-SNE and UMAP to summarise a high-dimensional biological dataset of your choosing. (Explain briefly your motivation for the dataset you choose to study.)
5. What are the pros and cons of each method? What would influence your decision as to which method to use?

PCA is available in R e.g. using `prcomp()`. Packages which you might wish to use: `https://github.com/lejon/TSne.jl`
`https://github.com/jkrijthe/Rtsne`
`https://github.com/tkonopka/umap`
`https://github.com/dillondaudert/UMAP.jl`

4 Open challenges in deep learning [25 marks]

Write a 3-page (max) essay discussing what you think are the open challenges are when applying deep learning to computational biology. To start with, please refer to (Sapoval et al., 2022), but please also find other relevant references to support your essay.

References

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*

Cottrell, G. W., Munro, P., and Zipser, D. (1987). Learning internal representations from gray-scale images : An example of extensional programing. *Ninth Annual Conference of the Cognitive Science Society, Seattle, 1987*.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.*, 79(8):2554–2558.

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., and Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.*, 13(1):1728.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605.

All references are available from https://paperpile.com/shared/mTCCFS