UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
STA5073Z – Data Science for Industry 2024

Assignment 2
Neural Networks

Due Date: Monday 7 October 2024 12pm (noon)

In the real world, as a data scientist, consultant, machine learning engineer and many other roles, one will often be given some problem and be expected to solve it using some appropriate method. A large number of problems can be solved using simple methods, like tree based methods. In other cases more sophisticated methods perform best. Being equipped with knowledge in a variety of methods is very useful - think of it has a toolkit. For this assignment, you will demonstrate your ability to use one tool from your toolkit in particular, neural networks. You are provided with two problems, one regression and one classification problem for which little is known about the features, but your task is to address this problem using a neural network.

Note: You are expected to work on this **on your own**. Please sign the plagiarism declaration provided on Vula and append it to your report.

1 Classification task

I have provided a dataset in Vula (”Data-classification.csv”) which has a number of features and a single target (last column in dataset). Your task is to use neural networks to address this task.

2 Regression task

I have provided a dataset in Vula (”Data-regression.csv”) which has a number of features and a single target. Your task is to use neural networks to address this task.

3 Submission

Your submission should consists of the following two items:

1. Your **report**, written in whatever word processing software you like (e.g. word, latex). The report should have two sections for the two problems, and contain a description of the type of problem, the approach you took, your results, hyperparameter tuning, pre-processing and details of anything else you explored. A good report is one which describes as much details of the investigation as possible, for

example, discusses multiple models which were explored and how that affected performance. Please provide details for any decision you have made, for example, if you decided to use a 1,000 layer network, then explain why that was the case.

2. Your **code**, written in whatever software you like (e.g. kaggle, R, Quarto). Doing your project in R is recommended but Python submissions will be accepted. Provide as much detail as possible in your code. You will be allocated 1 point for each sensible step you take. You are free to do whatever you like. In fact, the more details and the more you try, the higher your chances of being awarded marks. Thus, someone who builds 1 model and simply states the result obtained will not get a lot of marks. On the other hand, someone who put in a fair amount of effort and tries multiple things and discusses in detail will get a lot of marks. Steps that don't make sense for the given problem will result in no mark for that step (for example, discussing the use of softmax activation in the case of regression). Add sensible comments and details to the code where you feel it is appropriate. Feel free to explore anything and provide a discussion for what you tried. Please attach your code to your submission.

You can assume that your report and code is what you would send out to a client for which you have solved both problems, and your work will be marked accordingly. You are expected to present a coherent report, i.e. a continuous story that anyone should be able to at least follow, and that will not leave someone knowledgeable in the field with questions as to why/how you did something, or what the results mean. Both tasks 1 and 2 will count 50% towards your grade for this assignment.