UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
MASTERS IN DATA SCIENCE

STA5076Z SUPERVISED LEARNING

Due date: Friday, 30 June 2017 at 4pm
Late submissions will not be accepted

INSTRUCTIONS:

- Prepare reports for both topics below. All computer code must be submitted as an appendix and should include comments that clearly explain the purpose of the code.
- You are expected to work on this on your own. Please attach a plagiarism declaration to your report. This can be collected from stats reception. Plagiarism of any form will be reported to the university court.

1. Handwritten Digit Classification

The MNIST (Modified National Institute of Standards and Technology) dataset comprises tens of thousands of grey-scale, handwritten digits that have been size-normalised and centred in a fixed-size image. Each image is 28 pixels in height and width, with 784 pixels in total. Each pixel has an integer value ranging from 0 to 255 indicating the darkness of that pixel, where darker pixels have larger numbers. Each image has also been labelled with the digit that it represents. Your goal is to develop an accurate classifier that can be used to identify the digit in a new handwritten image.

You must download the training and test datasets from Vula (`mnist_data.zip`). The training and test sets contain data for 60,000 and 10,000 handwritten digits, respectively. Each dataset contains 785 columns. The first column, called `label`, indicates the digit that is represented by each image. The remaining 784 columns contain the darkness values for each pixel. These are labelled as `pix<i>_<j>`, where `<i>` is the row number and `<j>` is the column number of the pixel, counting from the top left of the image (`pix1_1`) to the bottom right (`pix28_28`).

You have been randomly assigned 5 digits to classify (see `digits.pdf` on Vula). You should begin by restricting your datasets to only contain the images for the digits that you have been assigned. Your final training and test sets should contain roughly 30,000 and 5,000 images, respectively.

You must submit a neat report with your findings. In your report, you should:

- Compare the performances of different classifiers and determine which is optimal. You may use any supervised or unsupervised techniques to tackle this problem
- Clearly motivate your model choices and hyperparameter settings

- Clearly explain your findings. Why does method A outperform method B?
- Use appropriate graphs and tables to summarise your findings

You will be heavily penalised if you simply dump model outputs into a document without explaining your results. You need to demonstrate that you understand what you are doing!

Creativity and out-of-the-box thinking will be rewarded (if correct)!

2. Insert Your Own Problem Here

You are required to source an interesting supervised learning problem of your own. Your dataset could come from your organisation or from an online repository such as www.kaggle.com. It must contain a *numerical* outcome variable, and a sufficiently large number of predictors and observations. You should model these data using various supervised learning techniques and identify the best method for your problem.

You must submit a neat report with your findings. In your report, you should:

- Very clearly explain your problem. Remember that your marker will not be familiar with your specific problem
- Clearly describe your dataset, both in writing and with graphs and/or tables
- Compare the performances of different models and determine which is optimal. You may use any supervised or unsupervised techniques to tackle this problem
- Clearly motivate your model choices and hyperparameter settings
- Clearly explain your findings. Why does method A outperform method B?
- Use appropriate graphs and tables to summarise your findings

Each student should work on a different dataset. If you intend to use a dataset from a public repository, please let your lecturer know which dataset you are using so that it can be reserved for you only (first come, first served!).

If you intend on working with confidential data from your organisation, note that only the course lecturers will have access to your final report and that this will be returned to you once it is marked. The data and findings will be treated as confidential and will not be disclosed to anyone at any time.