M5MS10 Machine Learning 2018
Assessed Coursework 2

Deadline for submission: 4pm Thursday 15th March 2018.

Email a typed report (in PDF format) including annotated computer code for carrying out the tasks detailed below to: *b.calderhead@imperial.ac.uk*

Undergraduate students please also submit a printed copy to the undergraduate office.

All questions carry an equal number of marks and the report should be no longer than 20 pages of text and images, excluding references and an appendix for additional code, which should be referred to as appropriate.

Please note the following:

- **Longer reports will not necessarily earn better marks - succinctness and clarity of expression will be rewarded more highly.**
- **All machine learning algorithms should be coded by yourself from scratch, although you may use built in functions for linear algebra operations, sorting and optimisation.**
- **When asked to comment on results, you should focus in particular on how the assumptions and underlying mathematical model may have affected any inferences made.**
- **Any queries about the coursework should be posted on the Machine Learning Blackboard forum, so that everyone can see my replies.**

Question 1

A major use of machine learning in the biological sciences is for automatic image processing, for example distinguishing cells (and their components, such as the nuclei) from background material. In this task you will import a digital image and use the Gaussian mixture model and K-means algorithm to (a) obtain an automatic segmentation of the image, (b) automatically count the number of cells in the image.

You can import the digital image `FluorescentCells.jpg`, which can be downloaded from the Blackboard in JPG format, using the following instructions.

If you are using *R*, first load the *jpeg* package by typing `library(jpeg)`. You can now load the image by using the command `img <- readJPEG("FluorescentCells.jpg")`. Make sure the working directory is pointing to the correct location of the image file. The variable `img` will represent the image as a multidimensional array, where the first 2 dimensions represent the coordinates of the pixel, and the 3rd dimension represents the colour using 3 numbers in the range $[0, 1]$. i.e. the colour vector of the 2nd pixel in row 4 can be displayed by typing `img[4,2,]`. You can plot the image by firstly creating a plot, `plot(c(0, 512), c(0, 512), type = "n", xlab = "", ylab = "")`, then creating a raster image, `rasterImage(img, 0, 0, 512, 512)`.

If you are using *Matlab*, first load the image using `img = imread('FluorescentCells.jpg')`, and display it using the command `imshow(img)`. Matlab also represents the image as a multidimensional array, however the pixel intensity vector is now a vector of type `uint8` so that each value is an integer in the range $[0, 255]$. This can cause problems when using algorithms on this type of integer data. A way around this is to convert the values into real numbers using the command `img = double(img)`, apply your algorithm, then convert the results back into the correct format, using `img = uint8(img)`.

Describe how a Gaussian mixture model and K-means algorithm can be used to produce a segmentation of the image such that each pixel intensity *vector* is replaced by the corresponding cluster mean or centroid vector. Produce pictures of your results showing the segmented image for a few selected values of K, using both algorithms, and comment on your results. Now using the appropriate output of your algorithm, apply another Gaussian mixture model to automate the counting of the number of cells in the image, and clearly explain your approach and the results you obtain.

Question 2

"Intelligent Machines: Forget Killer Robots - Bias is the Real Danger in AI."
- headline from a recent newspaper article.

Write up to 500 words discussing whether or not you agree with this headline regarding the imminent wide-spread application of AI and machine learning. You might, for example, focus on specific applications of machine learning and how biases might occur. Higher marks will be given for succinct, well thought through and well structured arguments. In particular, you should make reference to the concepts and methodologies presented in the lectures.

Question 3

Cryptocurrencies are experimental new digital currencies that are in active development. Examples include Bitcoin, Ethereum and Litecoin. They are created and held electronically, and are decentralised such that no one entity controls their production. They should be seen as high risk assets, whose prices can be very volatile, which nonetheless offer a potential opportunity for brave investors and traders.

On Blackboard you will find a dataset consisting of Ethereum prices collected every hour from a particular exchange for the past 2 months. The data includes open price (at the beginning of the last hour), the high and low price (during the last hour), the close price (at the end of the last hour), the volume (number of units traded during last hour) and the time stamp.

Choose **one** machine learning approach to create a predictive model for future Ethereum prices. You may for example use a model we have covered in class, or one of their extensions. Think very carefully about how you should define, train and test your model. For example, you could consider this problem within a regression framework, or within a classification framework. Describe in detail how your chosen model could be applied, what inference method you are using, and how it can be derived for this model. Discuss the problems you face and highlight the model's potential advantages and disadvantages. Discuss the accuracy of the results you obtain. Comment on whether you would use this to actually trade? If not, why not?