M5MS10 Machine Learning 2017
Assessed Coursework 2

Deadline for submission: 4pm Tuesday 7th March 2017.

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

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 document your approach and results.

Question 2

The Mauna Loa dataset contains atmospheric carbon dioxide (CO2) concentrations derived from the Scripps Institution of Oceanography's continuous monitoring program at Mauna Loa Observatory, Hawaii from 1958 to 1993. This record constitutes the longest continuous record of atmospheric CO2 concentrations available in the world. You can find this dataset on Blackboard. In this question you will investigate a probabilistic model that is often used within the atmospheric and environmental sciences.

(a) Write down and describe the equations involved in the Bayesian approach to linear regression, which may be used to describe an underlying hidden function of time given noisy data points. You may assume an independent Gaussian prior over each of the weights and independent noise on each data point.

(b) By considering the covariance of the function defined in part (a) at two time points,

show how we can derive an equivalent kernel function using an inner product of basis functions. Hence show that the Bayesian linear regression approach is equivalent to the Gaussian process approach with a suitable choice of kernel function.

(c) Using a Gaussian process, construct a predictive model of CO2 emissions over the subsequent 20 years. Give full details of any assumptions and equations you make use of when fitting your model to the data. In 2013 the concentration of CO2 topped 400 ppm for the first time in human history. Comment on how this observation compares with the prediction made by your model?

Question 3

"With respect to social consequences, I believe that every researcher has some responsibility to assess, and try to inform others of, the possible social consequences of the research products he is trying to create."

from Prof. Herbert A. Simon's autobiography, "Models of My Life"

Write up to 500 words discussing some of the possible consequences on society of a wide-spread application of AI and machine learning, and how these technologies might impact our lives, both positively and negatively. Feel free to refer to any mathematical approaches, methodologies and ideas presented in the lectures.

Question 4

Bitcoin is an experimental new digital currency that is in active development. It is created and held electronically. No one controls it. Bitcoins aren't printed, like dollars or euros - they're produced by people, and increasingly businesses, running computers all around the world, using software that solves mathematical problems. It's the first example of a growing category of money known as cryptocurrency. Bitcoin should be seen as a high risk asset, whose price can be very volatile, which nonetheless offers a potential opportunity for brave investors and traders.

On Blackboard you will find a dataset consisting of all Bitcoin transactions from a particular exchange on a particular day, together with the volume of each trade, and whether the transaction used the bid price or not. (Note that there are always two prices on offer - the *bid* is the lower price indicating how much someone would like to buy a bitcoin for, while the *ask* is the higher price indicating how much someone would like to sell their bitcoin for. Therefore when backtesting we may assume we can buy at the ask price and sell at the bid price.)

Choose **one** approach that we have covered in class to create a predictive model for future bitcoin prices. Think very carefully about how you should train and test your model. Describe in detail how your chosen model could be applied, discuss the problems

you face, and highlight its 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?