CS342 Machine Learning: Assignment #1
Supervised Learning: Classification and Regression

Due Date: February 10, 2016 @ 12:00pm

Assignment Date: January 27, 2016

Office Hours:
CS 3.07, Monday & Friday 10:00-11:00

Instructor: **Dr Theo Damoulas** (T.Damoulas@warwick.ac.uk)
Tutors: **Helen McKay** (H.McKay@warwick.ac.uk), **Shan Lin** (Shan.Lin@warwick.ac.uk)

The first assignment for CS342 is a programming assignment building on the concepts and models we have done in the first 3 weeks of lectures (OLS/Ridge regression/Lasso/k-NN/DTs). The due date is Wednesday 10 February at 12:00 noon and your submission should be in a zip file via the Tabula system containing the Python codes requested and a text file with what we should expect to get when running your codes. You can also include in the text file any additional observations and comments on your adopted strategies. We run Plagiarism detection software so please do not share your codes or solutions.

Marks are broken down by the different assignment components for a total of 15% of the total module marks. The next assignment is worth 25% and the final exam 60%. Further information and supporting material are available on the module website. We will discuss the assignment in the labs and the Friday class on 29th January for any questions and clarifications. Good luck!

1 Regression

In the first part of the assignment, we will focus on Regression using the models you should already be familiar with (OLS/Ridge/Lasso/k-NN/Decision Trees). This part of the coursework builds directly on the Lab material so you should be in a good position to complete this assignment task with ease.

Download the Data zip file from the module website next to *Assignment 1* box.

In this task you will work with the Abalone dataset. Information on the attributes and the history of the dataset is provided, together with the data, inside the Abalone folder. The goal is to *predict the age of abalone (number of rings) from physical measurements.*

→ [10 out of total 15 marks] write a function called *regressAba* that takes as inputs the Abalone data and outputs the mean and standard deviation μ_{CV}, σ_{CV} of the coefficient of determination R^2 from running 10-fold cross-validation (10-fold CV) with the *best settings* for the following models: OLS, Ridge regression, Lasso, Decision Trees, k-NN. You are predicting the last column (which is the number of rings) given the remaining 8 attributes.

[Marking: 4 marks for reading in the data, doing cross-validation with at least 1 model and returning the CV results (no matter how good/bad). 1 mark for every additional model CV results, and 2 marks for achieving top CV performances]

Hints: Categorical attributes can be encoded using 1-of-K (also known as one-hot) encoding. See the preprocessing function called *sklearn.preprocessing.OneHotEncoder*. You can also try different normalisations and feature expansions/engineering to improve your average 10-CV predictive performance for different models. **Only include in your submitted function the final choices/implementations you have made for each model. Include a text file with what we should get when running your codes as output and any additional information on the strategy you adopted.**

Clarifications: If you run 10-fold CV you have 10 validation sets and hence 10 values for any metric on these sets (e.g. R^2). The μ_{CV}, σ_{CV} is just the average and the standard deviation (see Lab 1-3) of these 10 metrics. If you need to choose any "tuning" parameter for any model (e.g. regularisation strength or the tree depth or k the number of neighbours) you should do that on a separate CV routine so that you *only report/output the best cross-validated performance from each model given your selected tuning parameters.*

2 Classification

In the second part of the assignment you will perform Classification using the classification models you were taught at class (k-NN, Decision Trees).

Download the Diabetes dataset that we used at Lab 2.

→ [5 out of total 15 marks] Write a function called *classifyDiabetes* that returns μ_{CV}, σ_{CV} from 10-fold CV using the measure of F1 score (see Lecture notes) for the following models: k-NN and Decision Trees.

[Marking: 3 marks for reading in the data, doing cross-validation with at least 1 model and returning the F1 CV results (no matter how good/bad). 2 marks for achieving top CV performances]

The same Clarifications and Hints as above apply.