# Group coursework 2

- Please submit your coursework on Moodle **by Midday on 26/02**.
- Please upload your text answers to Questions 1 and 2 in **one pdf file**.
- Please also upload one R script for Question 1, one R script for Question 2 i) and one R script for Question 2 ii) (**3 R scripts in total**).
- Make sure that you have included **sufficient** comments in the codes to make them **readable** by other people. There should be **no error messages** shown when I run your R scripts. You can assume that I have installed all required packages.

**Question 1** [**10 marks**]

Download the `newthyroid.txt` data from moodle. This data contain measurements for normal patients and those with hyperthyroidism. The first variable `class=n` if a patient is normal and `class=h` if a patients suffers from hyperthyroidism. The rest variables `feature1` to `feature5` are some medical test measurements.

i) Apply $k$NN and LDA to classify the `newthyroid.txt` data: randomly split the data to a training set (70%) and a test set (30%) and repeat the random split 20 times.

   For $k$NN, repeat 5-fold cross-validation five times to choose $k$ from $(3, 5, 7, 9, 11, 13, 15, 17, 19, 21)$. Use AUC as the metric to choose $k$, i.e. choose $k$ with the largest AUC.

   Record the 20 AUC values of kNN and LDA in two vectors. [4 marks]

ii) For the first random split, draw the ROC curves of $k$NN and LDA on one plot. [2 marks]

iii) Draw two boxplots on one plot based on the 20 AUC values of $k$NN and LDA. [2 marks]

v) What conclusions can you make from the classification results of $k$NN and LDA on the `newthyroid.txt` data? [2 marks]

**Question 2** [**5 marks**]

i) Complete the following `myFDA` function **without using any additional packages**. With the feature matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ $(N > p)$ and the label vector $\mathbf{y} \in \mathbb{R}^{N\times 1}$ of the training data, the `myFDA` function outputs the linear discriminant $\mathbf{w} \in \mathbb{R}^{p\times 1}$ for binary classification.

Hint: You can use the solution of $\mathbf{w}$ in slides directly.

[3 marks]

```
myFDA <- function(X,y){
#########################################################
# This function calculates the linear discriminant for binary
# classification.
# Input: Feature matrix, X (N by p) and label vector, y (N by 1)
# Output: Linear discriminant, w (p by 1)
#########################################################



return(w)
}
```

ii) Calculate the cosine similarity between the linear discriminant calculated from `myFDA(X=iris[51:150,-5],y=iris[51:150,5])` and that calculated from `lda(Species~.,data=iris[51:150,])`. [You can ignore the warning message from `lda` that the setosa class is empty.]

The cosine similarity between two vectors, $\mathbf{u} \in \mathbb{R}^{p\times 1}$ and $\mathbf{v} \in \mathbb{R}^{p\times 1}$, is defined as

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^T\mathbf{v}}{||\mathbf{u}||_2||\mathbf{v}||_2},$$

where $||\mathbf{u}||_2 = \sqrt{\mathbf{u}^T\mathbf{u}}$ and $||\mathbf{v}||_2 = \sqrt{\mathbf{v}^T\mathbf{v}}$. The value of the cosine similarity can be used to examine the relationship between two vectors. A value that is close to 1 or $-1$ indicates that the two vectors have the same directions or exactly opposite directions. Projections onto these two directions can be considered as the same. A value that is close to 0 indicates that the two vectors are orthogonal to each other. Projections onto such two directions are very different.

What conclusion can you make from your results? [2 marks]