Group coursework 3

- Please submit your coursework on Moodle **by Midday on 22nd of March**.
- Please upload your answers to Question 1 (2) and Question 2 in **one pdf file**.
- Please also upload **two R scripts** in `.R` files for Question 1 (1) and Question 2.
- 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 [**4 marks**]

(1) Consider the Gini index, classification error, and entropy in a binary classification setting. Create a single plot that displays each of these quantities as a function of $\hat{p}_{j1}$. The x-axis should display $\hat{p}_{j1}$, ranging from 0 to 1, and the y-axis should display the value of the Gini index, classification error, and entropy. [2 marks]

(2) Describe the patterns of the three curves you obtained in (1). [2 marks]

Question 2 [**16 marks**]

Use the `OJ` data from the `ISLR` package. This data contain 1070 purchases information to study which orange juice a customer would buy. The `Purchase` variable is a factor with levels `CH` and `MM` indicating whether the customer purchased Citrus Hill or Minute Maid Orange Juice. 17 features of the customers and products are recorded. The details of this dataset can be found in `https://rdrr.io/cran/ISLR/man/OJ.html`. Split the data to a training set (70%) and a test set (30%).

(1) Fit a support vector classifier to the training data by tuning the `cost` from $(0.01, 0.1, 1, 10)$. Compute the test error rate using the tuned value for `cost`. [1 mark]

(2) Fit a support vector machine with a radial kernel to the training data by tuning the `cost` from $(0.01, 0.1, 1, 10)$. Use the default value of `gamma`. Compute the test error rate using the tuned value for `cost`. [1 mark]

(3) Fit a support vector machine with a polynomial kernel to the training data by tuning the `cost` from $(0.01, 0.1, 1, 10)$. Use `degree=2`. Compute the test error rate using the tuned value for `cost`. [1 mark]

(4) Draw one plot with three ROC curves for the test predictions in (1), (2) and (3). Comment on the plot. [2 marks]

(5) Fit a decision tree to the training data with an optimal tree size determined by cross-validation. Create a plot of the pruned tree and interpret the results. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes. Compute the test error rate of the pruned tree. [3 marks]

(6) Fit a random forest to the training data with the following `mtry` values: 1, 2, 3, 4, 5 and 6. Compute the test error rates and comment on the results. Create a plot showing variable importance for the model with the best test error and comment on the plots. [4 marks]

(7) Draw one plot with two ROC curves for the test predictions in (5) and (6). Comment on the plot. [2 marks]

(8) Comparing the results of (4) and (7), what do you find? [2 marks]