Statistical Modelling & Machine Learning Final Exam
(9:30 - 11:30AM 12/17/2020, Thursday)

- **Instruction:**
  - Solve all 4 problems.
  - Upload a file with your R code and answers for the problems on I-campus.

1.[20pts] Consider '`Q1.csv`' data file. To predict the variable `Y` based on `(X1, X2, X3)`, use a **data modelling technique**.

(1) Investigate whether there is an irrelevant input variable for the prediction of `Y`. If it exists, find it and justify why it is the irrelevant variable.

(2) Construct the best parametric regression model and estimate the model parameters.

(3) Show the residual plot for the best model obtained from part (2).

(4) Based on part (2), describe the functional relationships between `Y` and individual input variables in the model.

2.[10pts] Consider '`Q2.csv`' data file. In the dataset, there are a categorical `Y` variable with two groups and 10 continuous input variables (`X1 - X10`). Visualize effectively data points for the two groups on a 2-dimensional plot (i.e., the plot should show two distinctive groups; You should use different colors or symbols for the two groups).

3.[10pts] Consider '`Q3.csv`' data file. In the dataset, there are 500 observations with a binary output variable (`Y`) and 2000 input variables (`X1 - X2000`). Consider the logistic regression model as a predictive model.

(1) Identify 5 input variables with the highest variable importance (i.e., top 5 ranked input variables).

(2) Suppose that we are interested in interpretation and our goal is to find a small subset of input variables for the prediction of `Y`. Find a small subset of input variables for the prediction of `Y`.

(3) If the results from parts (1) and (2) are different, explain briefly why they are different.

4.[20pts] Consider '`Q4train.csv`' and '`Q4test.csv`' data files. The datasets have information of companies and it consists of a binary `Y` variable and 12 input variables. Our goal is to identify companies that will go bankrupt after 1 year based on 12 input variables describing their current financial status. The description of the variables are as follows:

- `Y`: Bankruptcy after 1 year (0: No bankruptcy; 1: Bankruptcy).

- `X1`: Gross profit (in 3 years) / total assets.
- `X2`: Operating expenses / total liabilities.
- `X3`: Profit on sales / sales.
- `X4`: Current year sales / Last year sales.
- `X5`: Total liabilities / ((profit on operating activities + depreciation) * (12/365)).
- `X6`: (sales - cost of products sold) / sales.
- `X7`: (current assets - inventory) / short-term liabilities.
- `X8`: Profit on operating activities / sales.
- `X9`: Sales / total assets.
- `X10`: Total costs /total sales.
- `X11`: Logarithm of total assets.
- `X12`: (equity - share capital) / total assets.

Build your best model using the '`Q4train.csv`' (training set) data file, and then apply it to '`Q4test.csv`' (test set) data file. Report $\mathbf{F-measure}$ for the test set (i.e., try to attain the highest test F-measure value).

Instruction for Q4:

- When you build your model, use `set.seed(1)` to obtain constant results.
- Use the test set only for calculating the F-measure (Do not use test set for model selection or decision of tuning parameter values).
- Both your model building process and F-measure will be evaluated.