# HOMEWORK 3:
# SUPERVISED LEARNING

CPEN 355 @ UBC
TAs: Chun-Yin Huang(Primary), Wenlong Deng

## Instructions

- **Homework Submission:** Submit your code and report to Canvas. You will use Co-lab to implement the coding tasks. Please check Piazza for updates about the homework.
  - Upload a zip file containing two files: Your report in .PDF format and your notebook in .Ipynb format.
  - To ensure the reproducibility of your results: (1) set a seed for numpy and python random modules on top of your Colab notebook (2) restart and run all the cells of your notebook once before submission.
- **Collaboration policy**: The purpose of student collaboration is to facilitate learning, not to circumvent it. Studying the material in groups is strongly encouraged. It is also allowed to seek help from other students in understanding the material needed to solve a particular homework problem, provided no written notes (including code) are shared, or are taken at that time, and provided learning is facilitated, not circumvented. The actual solution must be done by each student alone. The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved.
- **Description:** In this assignment, you will practice how to build different classification models on a training set, choose best hyper-parameter using validation set, and evaluate the model on testing set. You will use Sklearn to implement the cross-validation and supervised learning models.

# 1 Coding practice [10 pts]

In this assignment, we will be working on the dataset of model price classification. The task is to classify the level of cost given the features of mobiles. In machine learning terms, it is a classification problem. The data are provided in the data folder. The training and testing data are in the different files. You can use `numpy` or `pandas` libraries to load the csv files.

The targets in the data have values:

- 0 (low cost)
- 1 (medium cost)
- 2 (high cost)
- 3 (very high cost)

which can be read from column `price_range`.

The features are the things like:

- battery_power: Total energy a battery can store in one time measured in mAh
- blue: Has bluetooth or not
- clock_speed: speed at which microprocessor executes instructions
- dual_sim: Has dual sim support or not
- fc: Front Camera mega pixels
- four_g: Has 4G or not
- ...

which can be read from the columns after dropping column `price_range`. Before feeding the features in the machine learning models, you need to do zero-mean unit variance normalization: https://en.wikipedia.org/wiki/Feature_scaling.

**(a)** Train a linear kernel SVM using `sklearn.svm.LinearSVC`. Please report the accuracy on testing data when choosing $C$ in the range of $(10^{-5}, 10^{-4}, 10^{-3}, \ldots, 10^{5})$ using `matplotlib.pyplot` where x-axis is $C$ and y-axis is accuracy (range from 0 to 1). The accuracy can be calculated using `sklearn.metrics.accuracy_score`.

**(b)** Train a RBF (Gaussian) kernel SVM using `sklearn.svm.SVC`. Please report the accuracy on testing data when choosing $\gamma$ in the range of $(10^{-1}, 10^{0}, \ldots, 10^{4})$ using `matplotlib.pyplot` where x-axis is $\gamma$ and y-axis is accuracy (range from 0 to 1). The accuracy can be calculated using `sklearn.metrics.accuracy_score`.

**(c)** Train a Random Forest Classifier using `RandomForestClassifier`. Please report the accuracy on testing data when choosing number of trees (`n_estimator`) in the range of $(10, 100, 500, 1000)$ using `matplotlib.pyplot` where x-axis is `n_estimator` and y-axis is accuracy (range from 0 to 1). The accuracy can be calculated using `sklearn.metrics.accuracy_score`.

**(d)** Please implement 5-fold cross-validation for hyper-parameter selection using `sklearn` on the training set and select the best parameters for problem (a)-(c) separately and report the corresponding testing accueracy.

# Machine Learning Question [2 pts]

*Some questions have Multiple Choices*
Question 1: What is the difference between Linear Regression and Support Vector Regression (SVR)?[1]

(A) Linear Regression can only model linear relationships, while SVR can model both linear and non-linear relationships.

[1] There could be more than one correct choice.

(B) Linear Regression is a classification technique, while SVR is a regression technique.

(C) Linear Regression does not use a margin for error, while SVR incorporates an error margin in the regression.

(D) Linear Regression is computationally more complex than SVR.

Question 2: Choose the disadvantage(s) of Decision Trees.

(A) Decision Trees are robust to outliers.

(B) Factor analysis.

(C) Decision Trees are prone to overfit.

(D) All of the above.

# Note

1. Remember to submit your assignment by 11:59pm of due date. Late submission will affect your scores.
2. If you submit multiple times, ONLY the content and time-stamp of the latest one would be considered.
3. We strictly follow the rules of UBC Academic Misconduct.