In this assignment, we will use the mathematical notations from the following reference books:

- **Reference books:**

  1. *An Introduction to Statistical Learning with Applications in R* **(ISLR)**. Authors: James, G., Witten, D., Hastie, T., Tibshirani, R.

  2. *The Elements of Statistical Learning Data Mining, Inference, and Prediction* **(ESL)**. Authors: Hastie, T., Tibshirani, R., Friedman, J.

1 Programming

1.1 Classification

(PhD : 50 points, MS: 50 points)

In this assignment, you will experiment with three classification algorithms on a real-world dataset.

Dataset: We will use the *Tic-Tac-Toe Endgame Dataset* (for all classification algorithms) from UCI's machine learning data repository. The training/validation/test sets are provided along with the assignment in Blackboard as `hw1ttt_train.data`, `hw1ttt_valid.data`, and `hw1ttt_test.data` for the *Tic-Tac-Toe Endgame Dataset*. For description of the datasets, please refer to https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame.

Please follow the steps below:

(a) Data Pre-processing (No points) In assignment#1 you implemented data pre-processing techniques to preprocess the Tic-Tac-Toe Endgame Dataset. Please reuse the same codes for obtaining pre-processed matrices and vectors will be used as inputs to the classification algorithms.

(b) Performance Comparison Compare three algorithms - Decision Trees, Support Vector Machines, Linear Discriminant Analysis- on the provided dataset.

Decision Trees (20 points): Train decision trees and report the training, validation and test accuracy and AUROC scores for different split criterions (Gini index and cross-entropy), by setting the minimum number of leaf node observations to 1; 2; ... ; 10. You need to report the results for 2 x 10 = 20 different cases. When training decision trees, please do not use pruning.

Support Vector Machines (SVM) (20 points): Train SVM models and report the training, validation and test accuracy and AUROC scores by varying hyper-parameters such as penalty C and kernels.

Linear Discriminant Analysis (LDA) (10 points): Train a LDA model and report the training, validation, and test accuracy and AUROC scores.

1.2 Linear Model Selection and Regularization

(Extra credit: PhD : 25 points, MS: 25 points)

In the problem, you will predict the per capita crime rate in the Boston dataset.

Try out some of the regression methods such as Best Subset Selection, the LASSO, Ridge Regression, and Principal Components Regression on the Boston dataset. Present and discuss results (MSE of the different approaches) for the approaches that you consider.

Submission Instruction: You need to provide the followings:

- Provide your answers/plots/results to the programming problems in a PDF file, named as `IS777_hw#_fa19_YourLastName.pdf`. You need to submit the homework in electronic version as pdf file on Blackboard.
- Submit ALL the code and report via Blackboard. Recommend to use Python programming language. For your program, you MUST include the main function called `IS777_hw#_fa19.py` in the root of your folder. After running this main file, your program should be able to generate all of the results needed for this programming assignment, either as plots or console outputs. You can have multiple files (i.e your sub-functions), however, the only requirement is that once we unzip your folder and execute your main file, your program should execute correctly. Please double-check your program before submitting. You should only submit one `.zip` file. No other formats are allowed except `.zip` file. Also, please name it as `[lastname]_[firstname]_hw1_fa19.zip`.

Collaboration: You may collaborate. However, collaboration has to be limited to discussion only and you need to write your own solution and submit separately. You also need to list with whom you have discussed.