C S 487/519 Applied Machine Learning I
Fall 2018
Project 3: Compare classifiers in scikit-learn library

1 Objective

In this *individual* project, you are required to understand and compare several classification algorithms that are provided by the Python scikit-learn library (`http://scikit-learn.org/stable/`).

2 Requirements

- (50 points) Write classification code by utilizing several scikit-learn classifiers: (i) perceptron, (ii) support vector machine (linear and non-linear using Radial Basis Function (RBF) kernel), (iii) decision tree, (iv) K-nearest neighbor, and (v) logistic regression.
- (15 points) Each classifier needs to be tested using two datasets: (1) the `digits` dataset offered by scikit-learn library, and (2) one dataset containing time-series instances. Example of the second dataset can be the *REALDISP Activity Recognition Dataset* (`https://archive.ics.uci.edu/ml/datasets/REALDISP+Activity+Recognition+Dataset`).
- (15 points) Properly analyze the classifiers' behavior by applying the knowledge that we discussed in class. Such analysis should include at least accuracy and running time.
- (15 points) Understand the source code of `DecisionTreeClassifier` (You can follow the `source` link in `http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html`).
  - (5 points) Please denote **two** strategies that this classifier implements to pre-prune or post-prune the tree.
  - (10 points) For each strategy, please clearly identify the repository file and the lines of code that implement such strategies.
- (5 points) Write a readme file `readme.txt` with the commands to run your code.
- Your Python code should be written for Python version 3.5.2 or higher.
- Please properly organize your Python code (e.g., create proper classes, modules).

3 Submission instructions

- In your github repository, create a project folder `proj3`.
- Put all your files (Python code, readme file, report, etc.) in your project folder.
- Submit the link to your github repository folder through Canvas.

4 Grading criteria

(1) The score allocation has already been put beside the questions.

(2) Please make sure that you test your code **thoroughly** by considering all possible test cases. Your code may be tested using more datasets.

(3) At least 5 points will be deducted if submitted files (including files types, file names, etc.) do not follow the instructions.