C S 487/519 Applied Machine Learning
Compare classifiers in scikit-learn library

1 Objective

In this *individual* homework, you are required to understand and compare several classification algorithms that are provided by the Python scikit-learn library.

2 Requirements

2.1 Tasks

(1) (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), and (iii) decision tree. In total these are three classifiers.

(2) (15 points) Each classifier needs to be tested using two datasets: (1) the `digits` dataset offered by scikit-learn library, and (2) another dataset on your own choice.

(3) (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.

(4) (15 points) (**CS 519 only**) Understand the source code of DecisionTreeClassifier (You can follow the source link).

   (a) (5 points) Please denote **two** strategies that this classifier implements to pre-prune or post-prune the tree.

   (b) (10 points) For each strategy, please clearly identify the repository file and the lines of code that implement such strategies.

   (c) Put your understanding in a report file (`report.pdf`). The file content should be succinct.

(5) (5 points) Write a readme file `readme.txt` with detailed instructions to run your program.

2.2 Other requirements

- Your Python code should be written for **Python version 3.5.2 or higher**.
- Please write proper **comments** in your code to help the instructor and teaching assistants to understand it.
- Please properly organize your Python code (e.g., create proper classes, modules).
- You can put your code to Jupyter Notebook or a `.py` file.

3 Submission instructions

Put all your files (Python code, readme file, report, etc.) to a zip file named `hw.zip` and upload it to Canvas.

4 Grading criteria

(1) CS 519 students need to answer all the questions. CS 487 students do not need to answer questions marked with **(CS 519 only)** although you have the freedom to work on them. Your scores will be scaled to 100. If CS 487 students answer the questions marked with **(CS 519 only)**, you will not have any points deducted if your answers are wrong; you will not get any extra points either if your answers are correct.

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

(3) FIVE points will be deducted if files are not submitted in the required format.

(4) If the total points are more than 100. Your grades will be scaled to the range of [0,100].

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