Assignment 1– Introduction to Python & sklearn

Assignment overview. This assignment is designed to practice Python and to introduce you the sklearn library for machine learning in Python. This Assignment requires you to install sklearn, tensoflow, and download the **Iris**, and **Wine** data set to learn and predict some items by using algorithms implemented in sklearn. We also practice to display NMIST data that we need later.

Submission. Create a Jupyter Notebook file called ML_Assignment1.ipynb which should be submitted through Brightspace. Also, you need to submit a separate csv file with your wine label predictions.

Submission deadline. Wednesday, September 19, 11:00 am (1/2h before class!).

Late submission policy: Late submissions **cannot** be permitted as solutions will be discussed in class.

Academic Integrity: Dalhousie academic integrity policy applies to all submissions in this course. You are expected to submit your own work. Please refer to and understand the academic integrity policy, available at https://www.dal.ca/academicintegrity

If you have a question: You can ask our TA Yoshimasa Kubo for help during the labs or email him at ys842667@dal.ca. If you still have questions, feel free to email the instructor at tt@cs.dal.ca.

Questions:

1. **[10 marks]** MNIST (http://yann.lecun.com/exdb/mnist/) is a famous dataset and benchmark for pattern recognition. It contains images of hand written numbers that are normalized and centered in a 28x28 pixel array. The data set was derived from a NIST (National Institute of Standards and Technology) data set, hence the acronym for Modified NIST. There are many ways to load this data set. Find this data set and **write** a program that displays some of these examples.

2. **[20 marks]** In the example code for the Iris data classification, we used 10-fold cross-validation to evaluate the accuracy of predictions with a linear SVM. In the example, we used the sklearn method `model_selection.cross_val_score`.
   a. Explain briefly what k-fold cross validation is and what it is used for.
   b. Write a script that does a k-fold cross-validation without using the `cross_val_score` function and **compares** the results with the sklearn function.
   c. Compare the cross-validated results of the SVM and RF and comment on which method is better.

3. **[20 marks]** Please download the wine.zip file and extract it to the directory for this assignment. Read through the wine_names.txt file and come to understand the problem and the *wine* data contained in the wine.train dataset. Train one of the models **SVM**, **MLP**, or **RF** to develop the best possible model for classifying the *wine* data in the hold-out test data set of 58 records in the wine.test file given the training data. In other words, you must submit a list of 58 classifications (as a separate *.csv file) for the hold-out test set in the same order as received. We will use your answers to score how well your model performs.
   Describe briefly your methodology for determining the best model and submit your final prediction program as well as the .csv file with the labels. Everyone will be ranked based on how well they do on their classification of the hold-out test set and **2 additional marks** will be given to the best 10% of submissions.