Assignment 3. Nearest Neighbours and Cross-validation

Due Date: November 3, 11:30 pm

Late submission

If you submit the assignment after the deadline, the following penalty is applied:

◇ 10% penalty if the submission is before November 10, 11:30 pm (if the mark before applying the penalty is 78 out of 100, after applying the penalty it is 78 - 7.8 = 70.2 out of 100);

◇ 50% penalty if the submission is after November 10, 11:30 pm and before Dec. 68, 11:30 pm.

Description:

In this assignment you are required to implement cross-validation and k-nearest neighbours (k-NN) for regression. You will use the Boston housing data set, which is available in scikit-learn (use "load_boston"). You will use cross-validation to select the best k-NN model among all $k, 1 \leq k \leq N$, where N is the size of the training set.

At the beginning you have to split the data into the training and test sets. Use an $80\% - 20\%$ split **Whenever you use randomization in your code, use a number formed with the last 4 digits of your student ID, in any order, as the seed for the pseudo number generator.**

You have to write a report to present your results and their discussion. You have to specify the model that you deem to be the best and indicate the test error. Justify your choice. The report should also contain a plot of the cross-validation and training errors for all k-NN models, versus k. Identify in your report the set of values k for which underfitting, respectively overfitting occurs, and the set of values reaching a good bias variance trade-off. Justify your choice.

Besides the report, you have to submit your numpy code. The code has to be modular and use vectorization whenever is possible. Write a function for each of the main tasks. Also, write a function for each task that is executed multiple times. The code should include instructive comments.

Submission Instructions:

- Submit the report in pdf format, the Python file containing your code (extension .py), and a short demo video. The video should be 1 min or less. In the video, you should scroll down your code briefly explaining it, show that it runs and that it outputs the results for each part of the assignment. The main Python file in the project should be clearly distinguishable. Some feedback might be written on your report, therefore, please DO NOT ZIP YOUR FILES.
  Naming convention:
  "studentMacId_studentNumber_A3_report", "studentMacId_studentNumber_A3_code".