Assignment 2. Nearest Neighbours and Cross-validation

Due Date: October 16, 11:59 pm

Description:

In this assignment you are required to implement cross-validation and k-nearest neighbours (k-NN) for regression. You will use data generated synthetically as for Assignment 1. The prediction is to be performed based on only one feature, denoted by x. The target t is a noisy measurement of the function $f_{true}(x) = sin(4\pi x)$. Thus, t satisfies the following relation

$$t = \underbrace{sin(4\pi x)}_{f_{true}(x)} + \epsilon \tag{1}$$

where ϵ is random noise with a Gaussian distribution with 0 mean and variance 0.09. Note that $f_{true}(x)$ is the optimal predictor for the data generated with equation (1) (i.e., the predictor that achieves the smallest expected squared error). However, in this experiment, you will only have a training data set and you will design the predictor only based on the knowledge of the training set.

Construct a training set consisting of only 201 examples $\{(x^{(1)}, t^{(1)}), \cdots, (x^{(200)}, t^{(200)})\}$, where $x^{(1)}, \cdots, x^{(201)}$ are uniformly spaced in the interval $[0, 1]$, with $x^{(1)} = 0$, $x^{(201)} = 1$, and $t^{(1)}, \cdots, t^{(201)}$ are generated using relation (1). Construct a test set consisting of 101 examples with the feature x uniformly spaced in the interval $[0, 1]$ and targets generated randomly according to relation (1). **When generating the random data use a four-digit number containing the last 4 digits of your student ID (in any order), as the seed for the pseudo number generator.**

You are required to implement k nearest neighbours (k-NN) for all k, $1 \leq k \leq 60$. Next perform 5-fold cross-validation to choose the best k-NN model.

You have to write a report to present your results and their discussion. The report should also contain a figure with the plots of the training error and the cross-validation error for all k-NN models. Identify in your report the set of values k for which underfitting, respectively overfitting occurs, and the set of values reaching a good trade-off between overfitting and underfitting (in other words, "the region of optimal capacity). Justify your choice.

You have to specify the model that you deem to be the best and indicate the test error. Justify your choice. Next train this model on the whole training set and plot (in the same figure) the prediction versus x for the training points and for the test points (in different colours) along with the "true" curve. Compare the performance (using the test set) of this predictor with the performance of the best predictor you obtained in Assignment 1.

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 (with extension ".py") containing your code, and a short demo video. The video should be 2 min or less. In the video, you should scroll down your code and explain what each part does. Submit the files in the Assignments Box on Avenue.