Lab 5

(Deadline 17:00 10/01/2018)

Task

You will be provided with a machine learning benchmark dataset (details below). The task focuses on the implementation and critical analysis of multiple regression methods: We expect that each student will implement, evaluate, and analyse each of the following algorithms:

- Nearest neighbour
- Linear regression
- Regression forest
- Gaussian process

The results should be presented as a report with a 3000 word limit in a pdf document. Please also include your code in Jupyter workbook format as well.

Suggested document structure

Please structure your report sensibly; here is a suggested structure but feel free to deviate if it makes sense.

Section 1: Introduction

Explain the problem

Section 2: Methods

Explain regression forests and Gaussian process regression; Why (or why not) you would use these algorithms.

Section 3: Validation on a toy problem

Design a toy problem (e.g., a two dimensional regression problem) and provide a discussion on validating your implementations on this problem. Please explain the rationale of your toy problem design.

Section 4: Experiments and analysis

Please report your experimental results. This could include

1) Your hyperparameter selection strategies.
2) Performance comparison of different algorithms in terms of accuracy, computational complexity, etc.
3) Analysis: How the results are influenced by different hyper-parameter choices? Why does one algorithm perform better than the others on the given dataset?

Mark Scheme

This project is worth 40% of your marks for the project:

- 9 marks for the method,
- 14 marks for designing and validating on a toy problem,
- 14 marks for the experiments,
- 3 marks for the analysis,

for a total of 40 marks.

If a piece of work is submitted after the submission date (and no extension has been explicitly granted by the Director of Studies), the maximum possible mark will be 40% of the full mark. If work is submitted more than five working days after the submission date, the student will receive zero marks.

Data Set

Included on Moodle is the SARCOS data set, a regression problem where the task is to predict the torque of one motor of a robotic arm given physical joint details. Specifically, position, velocity and acceleration for 7 degrees of freedom.

The data set is provided as one csv file -you are responsible for splitting it for train/test and hyperparameter learning. Each row is an exemplar, and each column a feature. Your task is to predict the last column (#22) given the first 21 columns. Be aware that you will not want to use more than 10000 exemplars when training a Gaussian process.

The data set was originally obtained from http://gaussianprocess.org/gpml/data/