Lab 5 / Project

Deadline Friday 8th January 2021 (week 15), 8pm.
Submission via Moodle

Task

You have been provided with a dataset, which you are to use to predict crop yield given climate data, i.e. a regression problem. The task is to perform a "typical" machine learning project:

- Prepare the data
- Implement multiple algorithms, specifically:
    - Linear regression (As the *baseline*; you may reuse your earlier code)
    - Regression forest (Use technique in lecture; do not use a histogram)
    - Gaussian process
    - One of your own choosing (One we taught or something else)
- Evaluate the algorithms on the dataset and tune them to maximise performance
- Critically analyse the results

Submission

Your project should be presented as a report with a 3000 word limit in a **pdf document**. As it is only fair if everyone has the same information, be aware that the universities rules require that we ignore upto 10% over the specified word limit, i.e. the effective word limit is actually 3300. Please also submit your code as well (Jupyter workbook and/or normal .py files). The report is to be structured with the below section headings, where each heading is a question to be answered. Notes on how to answer them satisfactorily have been included below many of the questions.

1. **What exploration of the data set was conducted?**
   Statistics calculated, visualisations and any conclusions reached.
2. **How was the dataset prepared?**
   Any feature engineering; train/validation/test split.
3. **How does a regression forest work?**
   You should start by describing a decision tree and then how it becomes a regression forest. Pay attention to the regression part.
4. **How does a Gaussian process work?**
   Focus on intuition, not reciting the maths.
5. **Which additional algorithm did you choose and why?**
   This should reference the properties of the selected algorithm and why they suggest it would have good performance on the specific problem.
6. **What are the pros and cons of the algorithms?**
   You should be comparing the algorithms to each other.

7. **Describe the toy problem used to validate the algorithms, and explain its design?**
   A toy problem is used to validate a machine learning algorithm is working as expected, and debug it if not; it will usually be low dimensional, so it's easy to visualise and synthetic, with a simple y = f(x) function, so the correct answer is known. It does not have to be a function that the algorithm can learn exactly, as observing expected failure can be as informative as observing expected success. You should comment on if it achieves these properties.
8. **What evidence of correct, or incorrect, implementation did the toy problem provide?**
9. **How were the hyperparameters optimised?**
   The discussion of the technique should be complemented by graphs/tables generated during the process. Over and under fitting should be considered.
10. **What results are obtained by the algorithms?**
11. **How fast do the algorithms run and how fast could they run?**
    You should consider both your runtime and the big-O complexity of each algorithm, during both training and evaluating.
12. **Which algorithm would you deploy and why?**
13. **How could the best algorithm be improved further?**
    Discuss modifications that may improve accuracy, and why they may have that effect.
14. **If you were to try another algorithm then which one and why?**
15. **Are the results good enough for real world use?**
    This requires consideration of the context in which the algorithm would be used. You should state any assumptions you've made about that context.
16. **How could this solution fail?**
17. **What improvements could be made to the data set?**
    For instance, what features could be added, and why?

The word limit applies to the report as a whole, but it is suggested that a roughly even split between questions is sensible, i.e. 176 words per question. You are encouraged to include tables and graphs; remember to justify your answers.

Mark scheme

This project is worth 80% of your marks for the unit. The answer to each question will be marked as satisfactory or unsatisfactory; partial credit will be rare. Each satisfactory answer is worth four marks, for a total of 68 marks. An additional 12 marks are then reserved for correct implementation of the algorithms (4 marks per algorithm, excluding linear regression).

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 project will receive

zero marks. You must comply with the universities plagiarism guidelines: http://www.bath.ac.uk/library/help/infoguides/plagiarism.html

Data set

The task is to predict crop yield (tonnes per hectare) for the major maize crop of farms distributed across the entire planet (the major crop is the main crop of the year; farms may also grow a second crop with a lower yield). For each farm climate data is provided, specifically three features for each month:

- Total rainfall, in mm.
- Mean minimum temperature, in degrees celsius (minimum temperature for each day, averaged over the entire month).
- Mean maximum temperature. Likewise to above.

Year has also been included to account for improvements in farming techniques; the data covers three decades

The data is provided as one csv file (it includes a header) - 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 (#38) given the first 37 columns. There are 31744 exemplars.

This data set was created by merging and subsampling

- “The global dataset of historical yields for major crops 1981–2016” by Iizumi & Sakai.
- USA NOAA World Weather Records (WWR).