OXFORD BROOKES UNIVERSITY
School of Engineering, Computing, and Maths
Module P08820: Introduction to Machine Learning, Semester 2 2019

Coursework: MNIST Written Character Classification

For this coursework, you are required to apply machine learning methods to a classification problem specified by a data set. This coursework is worth 100% of the module mark and is due in week 12. An interim assessment of part of the work will happen in week 6.

Data set

The dataset MNIST contains images of hand-written characters 0-9. The task is to take an image as input and determine which of the numbers 0-9 is written in it. The dataset is an industry standard and one of the most common benchmarks for new classification algorithms. The dataset and information about it can be found here:

http://yann.lecun.com/exdb/mnist/

The dataset is already pre-structured into training and test sets, which can be downloaded on the homepage. Performances of common algorithms on this database are well known and also documented as error-rates on the test set on the homepage.

Machine Learning and Evaluation

For this coursework you will apply existing algorithms in R to the MNIST machine learning problem and dataset. Image data from the dataset needs to be loaded and pre-processed.
Apply appropriate dimensionality reduction techniques for visualization of the data set.
Use pre-processed and dimension-reduced data to train a k-nearest-neighbour (k-NN) classifier on the dataset.
Use appropriate statistical tools to estimate its classification error and determine which value of k is best suited for the task.
Chose one additional, suitable classification algorithm and evaluate its effectiveness on the dataset.
The experiments need to allow for a clear comparison of these two algorithms and allow for a recommendation.
The entire experiment must be submitted as a set of R scripts from which it can be reproduced.

OXFORD BROOKES UNIVERSITY
School of Engineering, Computing, and Maths
Module P08820: Introduction to Machine Learning, Semester 2 2019

Report structure and assessment (100% of module mark)

1) Write a brief introduction that introduces [15%]

a) The notion of a classification problem
b) The notion of and rationale for separate test and training data sets
c) Explains what MNIST is about and how it is an example of the previous two points

2) Apply dimensionality reduction to the data set in R and visualise the different classes using two dimensions, based on it

[20%]

3) Realise and describe an experiment in R that evaluates the classification error rate for the k-nearest-neighbour (k-NN) classifier on the MNIST dataset. Use appropriate pre-processing of the data. Determine the most suitable value for k experimentally using a suitable error measurement. Use appropriate illustrations and diagrams as well as statistics

[30%]

4) Realise and describe an experiment in R that evaluates a second, suitable classification algorithm of your choice on the MNIST dataset. Use appropriate illustrations and diagrams as well as statistics in order to compare to the previous results

[20%]

5) Write a brief conclusion on the results and compare to results published for other algorithms on the dataset's homepage. Which approach and parameter value is best suited? What properties other than just the classification error could be important when deciding what method is most suited? Explain possible current limitations of your solutions and possible further strategies to improve on the results

[15%]

Submission

Detail the five components in a single-file report of no more than 6 pages, including illustrations. Append all source code to reproduce your experiments at the end (not included in the 6 page limit).
Submit the final report by the deadline specified below, digitally as single PDF file939 on Moodle.

Report submission deadline: Friday April 5 2019, 17:00:00

Feedback: Monday April 29th (two weeks, accounting for Easter break)