Project: Digit Classification

Machine Learning

Instructions

- You may use any typesetting software you wish, but I would encourage you to use R-Markdown or LaTeX.
- Provide complete code for each of the relevant questions under separate headings as an appendix to your write-up. For example, for Question 2, provide all code pertaining to this question as a single script in the appendix under the subsection heading 'R-Code: Question 2'. Start each question on a new page.
- You may NOT provide R output interspersed between your answers!. Please typeset relevant elements in the output either in-line, or tabulate results formally. Plots are very useful, but use them sparingly — make sure that a given plot is relevant to the question and pertains to text in your answer. Figures are meant to enrich your analysis, not leave it to the reader to analyse. Provide captions for all figures and tables. Square figures only.
- When you typeset R code use `courier` or an equivalent 'typewriter'-like font.
- All assignments must be accompanied by a signed plagiarism declaration — a template is provided on Vula.
- Hand-in dates will be announced on Vula.

Problem Set

Under the STA5003W Vula tab, go to 'Resources/Machine Learning/Data' and download the training and test datasets within the folder. Each row represents a number. For the training dataset, the first column shows the true value and columns 2 - 785 recreate a 28 by 28 grid that reveals a picture of the handwritten number. For the test dataset, the digit value has been masked.

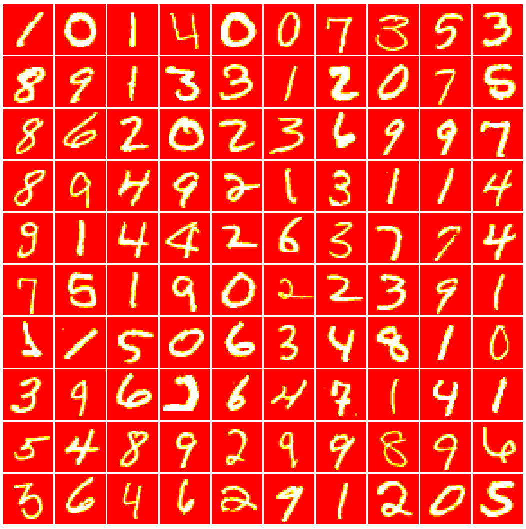

Figure 1: A sample of 100 digits within the training dataset.

Using the training data, develop a classification scheme that will decide whether a number is even or odd. Implement your decision rule on the test dataset and write your classifications for the 2500 test images to a *.csv* file.

Your assignment should implement the various classification methods that you have learned in this course. Describe the implementation of the various methods in detail. In addition, an in-depth comparison of the performances of the various methods should be given.

Email your R code, your report and a csv file giving your test data classifications. If you don't hand-in by the deadline, you get zero for the assignment!
Notes: Attach all R-code used as an appendix to your assignment. For the support vector machine (SVM), you may use the `solve.QP` function within the `quadprog` package to maximise the margin. You may also use the `randomForest` and `gbm` packages.