MSc in Data Science
Machine Learning
Academic Year: 2019-2020

Exercise 3: Digit Recognizer task[1]

Delivery Date: **24/1/2020**

Objective

In this exercise, your goal is to correctly identify digits from a dataset of handwritten images.

Dataset

MNIST ("Modified National Institute of Standards and Technology") is the de facto "hello world" dataset of computer vision. Since its release in 1999, this classic dataset of handwritten images has served as the basis for benchmarking classification algorithms. As new machine learning techniques emerge, MNIST remains a reliable resource for researchers and learners alike.

For this exercise, a subset of this dataset will be used. The data file can be found in the following link:

https://github.com/MSc-in-Data-Science/class_material/blob/master/semester_1/Machine_Learning/datasets/digit_recognizer_dataset.csv

The dataset contains gray-scale images of hand-drawn digits, from zero through nine. It contains 42,000 images.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.

The data set, has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image.

Each pixel column in the training set has a name like pixelx, where x is an integer between 0 and 783, inclusive. To locate this pixel on the image, suppose that we have decomposed x as $x = i * 28 + j$, where i and j are integers between 0 and 27, inclusive. Then pixelx is located on row i and column j of a 28 x 28 matrix, (indexing by zero).

For example, pixel31 indicates the pixel that is in the fourth column from the left, and the second row from the top, as in the ascii-diagram below.

Visually, if we omit the "pixel" prefix, the pixels make up the image like this:

---

[1] https://www.kaggle.com/c/digit-recognizer

```
000 001 002 003 ... 026 027
028 029 030 031 ... 054 055
056 057 058 059 ... 082 083
 |   |   |   |  ...  |   |
728 729 730 731 ... 754 755
756 757 758 759 ... 782 783
```

You are expected to:

- Experiment with different models and settings and decide on the best model for this dataset
- Write a brief report (2-4 pages maximum) describing the choices you made and the evaluation you performed