University of Tehran

School of Electrical and Computer Engineering

Pattern Recognition

MNIST Dataset Description

Corresponding TAs:

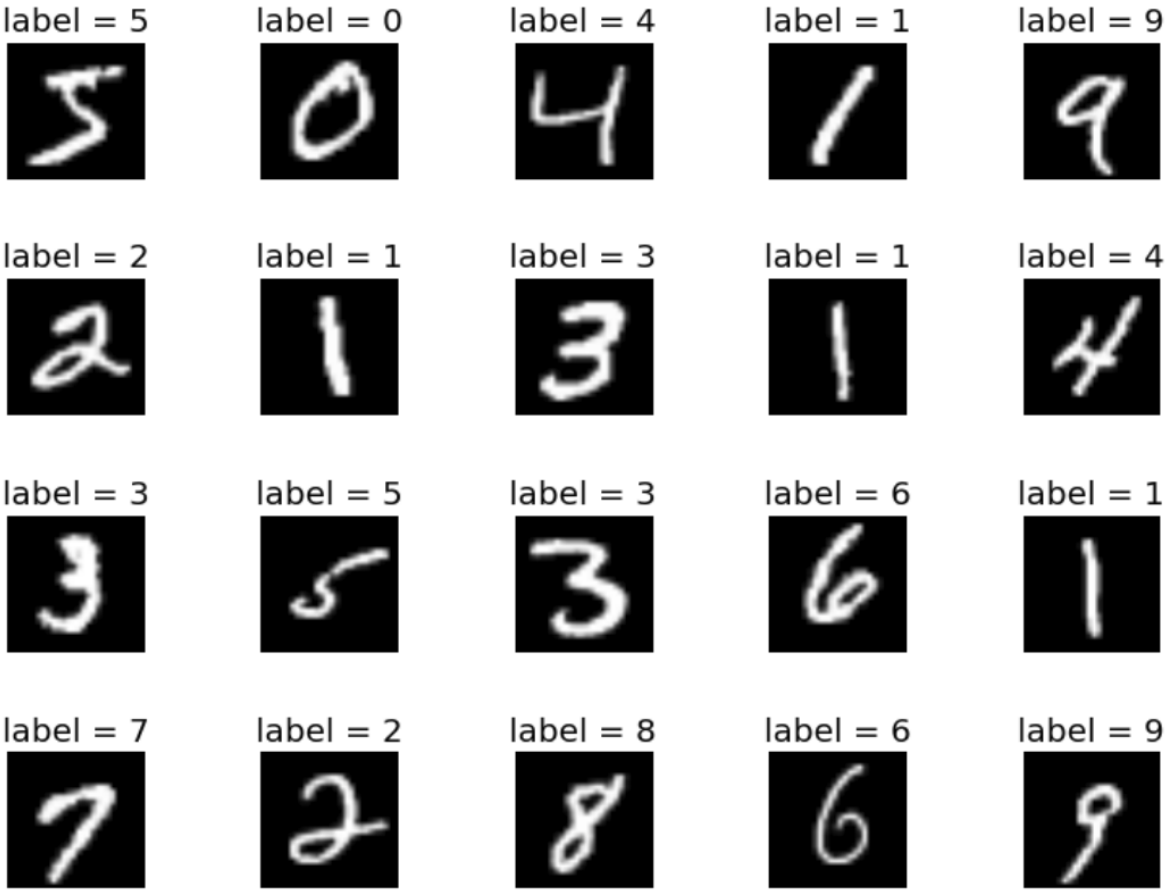
Marjan Shahi – marjan.shahi71@gmail.com

Niloufar Shahdoust – n.shahdoust@ut.ac.ir

Aban 97

What is MNIST?

The MNIST database of handwritten digits (including digits 0 to 9, so has 10 classes), has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image (28*28).



It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Before Using Dataset:

In this course, we used two versions of this dataset. The original dataset, named “MNIST”, and the small version, we name “TinyMNIST”. The TinyMNIST

dataset contains 5,000 train samples, and 2,500 test samples, and the images are scaled from 28*28 to 14*14.

Important Notes:

- From now on, in every implementation of your assignments, you need to apply the algorithms to whether “TinyMNIST” or “MNIST”. (It is clearly stated in your assignment which one to use).
- The dataset folder contains 4 “.csv” files, “testData.csv”, “testLabels.csv”, “trainData.csv”, and “trainLabels.csv”. The “Data” files are matrices of size (Number of Samples, Number of Features) containing flattened image samples. The “Labels” files are vectors of size (Number of Samples) containing label for each sample.
- The dataset is already shuffled and normalized to the range (0, 1), so there is no need to reshuffle or normalize it.