The EMNIST dataset, is a set of handwritten character digits derived from the NIST Special Database 19  and converted to a 28x28 pixel image format and dataset structure that directly matches the MNIST dataset. All the image files and labels of the MNIST dataset is encoded into these 4 files. We need to be able to extract the images from the files to work with them.

### **File descriptions**

### Four files are provided:

* Test Images : t10k-images-idx3-ubyte
* Test Labels : t10k-labels-idx1-ubyte
* Train Images : train-images-idx3-ubyte
* Train Labels : train-labels-idx1-ubyte

The IDX file format is a simple format for vectors and multidimensional matrices of various numerical types.

**The basic format for labels:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Offset** | **Type** | **Value** | **Description** |
| 0000 | 4 byte integer | 0x00000801(2049) | magic number (MSB first) |
| 0004 | 4 byte integer | 10000 or 60000 | number of items (test or train) |
| 0008 | unsigned byte | ?? | label |
| 0009 | unsigned byte | ?? | label |
| ... | ... | ... | ... |
| xxxx | unsigned byte | ?? | label |

**The basic format for images:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Offset** | **Type** | **Value** | **Description** |
| 0000 | 4 byte integer | 0x00000801(2049) | magic number (MSB first) |
| 0004 | 4 byte integer | 10000 or 60000 | number of items (test or train) |
| 0008 | unsigned byte | ?? | label |
| 0009 | unsigned byte | ?? | label |
| ... | ... | ... | ... |
| xxxx | unsigned byte | ?? | label |