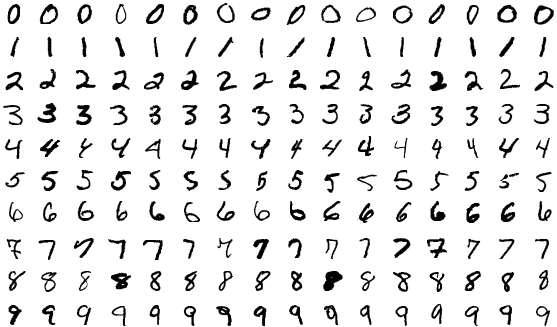
**ASSIGNMENT 1--MNIST BY SVM**

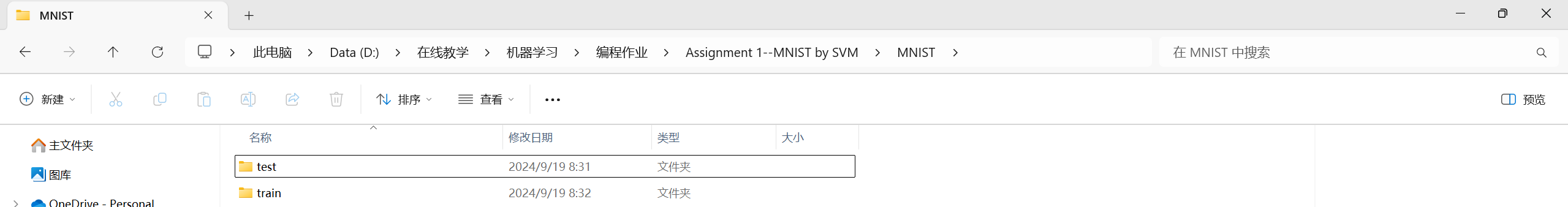
**Data:** The MNIST dataset consists of a vast collection of handwritten digits from 0 to 9. It’s like a giant library filled with pictures, where each picture is a grayscale image measuring 28x28 pixels. The MNIST dataset contains 60,000 images for training and another 10,000 for testing. The sample images are shown as the following figure.



**Fig. 1 Sample Images of MNIST**

**Task:** Using multi-class SVM to build a system which can classify the MNIST dataset. You need to use the 60000 images in the train folder to train your model, and use the 10000 images in the test fold to test your model.

The train folder and test folder are prepared in the MNIST dataset as shown below.



The following python code to preprocess the data are given for your reference.

from sklearn.model\_selection import train\_test\_split

#Training

dir\_path = 'D:\\datasets\\MNIST\\train' #Please specify your data directory.   
file\_ls = os.listdir(dir\_path)  
data = np.zeros((60000, 784), dtype=float)  
label = np.zeros((60000, 10), dtype=float)  
flag = 0  
for dir in file\_ls:  
 files = os.listdir(dir\_path+'\\'+dir)  
 for file in files:  
 filename = dir\_path+'\\'+dir+'\\'+file  
 img = mpimg.imread(filename)  
 data[flag,:] = np.reshape(img, -1)/255  
 label[flag,int(dir)] = 1.0  
 flag+=1

ratioTraining = 0.95  
xTraining, xValidation, yTraining, yValidation = train\_test\_split(data, label, test\_size=1 - ratioTraining, random\_state=0) # split the training data into 95% for training and 5% for validation.

#Testing  
dir\_path = 'D:\\datasets\\MNIST\\test'  
file\_ls = os.listdir(dir\_path)  
xTesting = np.zeros((10000, 784), dtype=float)  
yTesting = np.zeros((10000, 10), dtype=float)  
flag = 0  
for dir in file\_ls:  
 files = os.listdir(dir\_path+'\\'+dir)  
 for file in files:  
 filename = dir\_path+'\\'+dir+'\\'+file  
 img = mpimg.imread(filename)  
 xTesting[flag,:] = np.reshape(img, -1)/255  
 yTesting[flag,int(dir)] = 1.0  
 flag+=1

By using the above code, for each image we could obtain we could obtain a vector of 784 dimensions (because each image has 28x28=784 pixels). Then please use multi-class SVM to build a model to classify each image in the test folder into one of ten categories (represented as 0-9).

**Requirements：**

1. Using LIBSVM to build a system which is trained by the 60000 images in the train folder, and could classify each image into one of ten categories.
2. Write a report which consists of around 2 pages to specify the method you are using, and the results which include but not limited to **the accuracy, confusion matrix and ROC curve of your system on the 10000 test images**.

**Scoring Criteria:**

**Your score = 50% creativity of your method + 10% clarity of your report + 40% accuracy**