#MNIST 手写数字识别
##数据的读入
用padas来读取csv格式的文件

`train = pd.read_csv("..\input\train.csv")`

`test = pd.read_csv("..\input\test.csv")`

读取的数据类型为padas的DataFrame类型

##提取出label
训练集的第一列为label 其column为 label

`Y_train = train['label']`

删掉label这一列

```
    X_train = train.drop(label = 'label', axis = 1)
```

删除train 释放空间

```
    del train
```

##检查数据
```
    Y_train.value_count()
    X_train.isnull().any().describe()
    test.isnull().any.describe()
```

any()函数作用为若传入参数全为false 则输出为false 若有一个为true则输出为true 与full函数相反

describe()函数生成数据中元素的总数量（排除掉NAN）

value_count()函数统计相同的值的个数
以上均需要print函数才能输出

##正则化
```
X_train = X_train/255
test = test/255
```

##reshape

```
    arr.ravel()  # 此函数为将arr拉平为一维数组

    X_train.value.reshape(-1, 28, 28, 1)
    test.value.reshape(-1, 28, 28, 1)
```
-1代表自动推算出正确维度

```
    print(X_train.shape(), text.shape())#检查维度

    Y_train = Y_train.tocategorical(Y_train, num_classes=10)  #将标签向量化
```

##分离出训练集和验证集
```
    random_seed = 2
    X_train, X_val, Y_train, Y_val = train_test_split(X_train, Y_train, test_size =
0.1, random_state=random_seed)
```
从训练集中随机分成两部分 一部分作为训练 另一部分用来做验证集 验证集比例为0.1

四个参数分别为训练集，训练集标签　验证集所占比例 随机种子