第四次作业题目

手写体字符识别MNIST数据集是计算机视觉最为经典的一个数据库。里面包含0~9的10个手写字符。每张图由28×28的灰度图片组成。本题为了简化问题，只考虑训练集60000张图（见附件的数据train-images.idx3-ubyte），每个字符6000张图，共10个字符。请用kmeans进行对手写体字符进行聚类，其中固定k=10，即聚成10簇。要求：

1. kmeans的代码必须自己写，不能利用相应的kmeans库函数。
2. 聚类完，请分析聚类的精度。

3）提示：每张图看出28×28=784维度的高维空间的数据点，60000张图就相当于60000个数据点。

精度计算：

1. 首先聚成10簇，即为C1,C2,…,C10。每个簇Ci有Ni个点。比如，C2的簇有N2个点，N2个点中含有数字7（7指的是附件train-labels.idx1-ubyte的真实标签，和数据train-images.idx3-ubyte中60000个点依次对应）的点数最多，则认为这个C2簇的所有点的预测标签为7.
2. 依次分配10个标签给这10个簇（可能有相同标签的簇，无所谓）。
3. 10个簇的分好预测标签了，也就相当于所有的数据点（图片）分好了标签。对比预测的标签和数据集真实的标签(见附件的标签train-labels.idx1-ubyte，和数据train-images.idx3-ubyte中60000个点依次对应)，计算预测正确的概率。

python提示：如何读取数据附件的一张图

import numpy as np

import struct

import matplotlib.pyplot as plt

path = 'I:\\博二\\统计分析方法助教\\第四次作业\\data\\train-images.idx3-ubyte'

with open(path,'rb') as f1:

buf1 = f1.read()

image\_index = 0

image\_index += struct.calcsize('>IIII')

temp = struct.unpack\_from('>784B', buf1, image\_index)

# '>784B'的意思就是用大端法读取784( 28\*28 )个unsigned byte

im = np.reshape(temp,(28,28))

plt.imshow(im , cmap='gray')