K近邻算法

K近邻算法是一种常用的监督学习算法。其工作机制非常简单: 给定测试样本，基于某种距离度量找到与测试样本最近的k个训练样本，然后可以根据k个样本决定分类。譬如可以选择k个样本中最多的测试样本进行分类。

下面我们使用k近邻算法识别手写数字。手写数字为一个32\*32维的向量，我们可以把它看成一个1024维的向量。然后任意两个样本的距离完全可以通过计算1024维上的两个点之间的距离获得，然后我们可以找到与测试样本距离最近的的10个样本，其中最多的分类我们认为他就是测试样本的分类。具体代码如下：

**import** numpy **as** np  
**import** os  
**from** numpy **import** linalg  
  
**global** g\_train  
**global** g\_label  
  
**def** makeTrain(dir):  
 dirs=os.listdir(dir)  
 files = filter(**lambda** item:**not** os.path.isdir(item), dirs)  
 mat = []  
 label = []  
  
 **for** file **in** files:  
 arr = []  
 f = open(dir+**"/"**+file)  
 **while True**:  
 line = f.readline()  
 **if not** line:  
 **break  
 for** i **in** range(len(line)-1): *# line[len(line)]='\n'* arr.append(int(line[i]))  
 mat.append(arr)  
 label.append(int(file.split(**"\_"**)[0]))  
 **return** (np.array(mat),np.array(label))  
  
**def** testClassify(dir):  
 dirs=os.listdir(dir)  
 files = filter(**lambda** item:**not** os.path.isdir(item), dirs)  
 mat = []  
 right=0  
 wrong=0  
 **for** file **in** files:  
 arr = []  
 f = open(dir+**"/"**+file)  
 **while True**:  
 line = f.readline()  
 **if not** line:  
 **break  
 for** i **in** range(len(line)-1): *# line[len(line)]='\n'* arr.append(int(line[i]))  
 mat.append(arr)  
 testLabel = file.split(**"\_"**)[0]  
 label=classify(np.array(arr),10)  
 **if** testLabel == label:  
 right=right+1  
 **else**:  
 wrong=wrong+1  
 print(**"right="**,right,**", wrong="**,wrong)  
  
**def** classify(vec,k):  
 *# 计算各个训练数据与测试数据的距离* m = len(g\_label)  
 dis = []  
 **for** i **in** range(m):  
 dis.append([linalg.norm(vec-g\_train[i]),g\_label[i]])  
 dis = sorted(dis, key=**lambda** v:v[0])  
  
 *# 计算相似度最高的k个值，这里写入map做累积* dic = {}  
 **for** j **in** range(k):  
 **if not** str(dis[j][1]) **in** dic:  
 dic[str(dis[j][1])]=1  
 **else**:  
 dic[str(dis[j][1])]=dic[str(dis[j][1])]+1  
 **return** max(dic.items(), key=**lambda** x: x[1])[0]  
  
**if** \_\_name\_\_ == **"\_\_main\_\_"**:  
 *# 1 formate trainning date* (g\_train,g\_label) =makeTrain(**"/Users/zcy/Desktop/study/git/mlearning/res/trainingDigits1"**)  
  
 *# 2 test* testClassify(**"/Users/zcy/Desktop/study/git/mlearning/res/testDigits1"**)

计算946个测试样本集，有926个为正确计算，20个为错误计算，识别为97.89%