模式识别实验报告

实验一 K-均值聚类

学院：计算机科学与技术

姓名：张文强

学号：18S003044

一、实验内容

1、 使用 C 或 Matlab 编程实现 K-均值聚类算法：要求独立完成算法编程，禁止调用已有函数库或工具箱中的函数；

2、 使用仿真数据测试算法的正确性：将下列 19 个样本聚成 2 个聚类：

$$\mathbf{x}_1=(0,0)^t,\mathbf{x}_2=(1,0)^t,\mathbf{x}_3=(0,1)^t,\mathbf{x}_4=(1,1)^t,$$
$$\mathbf{x}_5=(2,1)^t,\mathbf{x}_6=(1,2)^t,\mathbf{x}_7=(2,2)^t,\mathbf{x}_8=(3,2)^t,$$
$$\mathbf{x}_9=(6,6)^t,\mathbf{x}_{10}=(7,6)^t,\mathbf{x}_1=(8,6)^t,\mathbf{x}_{12}=(7,7)^t,$$
$$\mathbf{x}_{13}=(8,7)^t,\mathbf{x}_{14}=(9,7)^t,\mathbf{x}_{15}=(7,8)^t,\mathbf{x}_{16}=(8,8)^t,$$
$$\mathbf{x}_{17}=(9,8)^t,\mathbf{x}_{18}=(8,9)^t,\mathbf{x}_{19}=(9,9)^t$$

3、 MNIST 数据集测试：ClusterSamples 中的 10000 个 784 维特征手写数字样本聚类为 10 个类别，根据 SampleLabels 中的标签统计每个聚类中不同样本的数量。测试不同初始值对聚类结果的影响。

二、程序代码

（K-均值算法部分代码）

```
centers = [0] * n_clusters
clusters = [list() for i in range(n_clusters)]

def init_clusters_centers():
    picked = random.sample(range(n_points), n_clusters)
    for idx, picked_id in enumerate(picked):
        centers[idx] = data[picked_id]

    if DEBUG_INFO:
        print ('Picked Ids:')
        print (picked)
        print ('Init Centers:')
        for i,c in enumerate(centers):
            print ("%d:%s" % (i, c))

def distance_metric(x, y):
    dist = np.linalg.norm(x - y)
    return dist

def assign_cluster():
    for cluster in clusters:
        cluster.clear()

    for idx, p in enumerate(data):
        min_v = distance_metric(p, centers[0])
        min_i = 0
        for c in range(1,n_clusters):
            dist = distance_metric(p, centers[c])
            if dist < min_v:
                min_v = dist
                min_i = c
        clusters[min_i].append(idx)

def recalc_centers():
    for c in range(n_clusters):
        center = 0
        for p_id in clusters[c]:
            center += data[p_id]
        center /= len(clusters[c])
        centers[c] = center

def judge_converge(last_clusters, clusters):
    for c in range(n_clusters):
        if set(last_clusters[c]) != set(clusters[c]):
            return False
    return True


def main():
    last_clusters = [list() for i in range(n_clusters)]
    iteration = 0
    init_clusters_centers()
    while True:
        assign_cluster()
        recalc_centers()
        if judge_converge(last_clusters, clusters):
            break
        last_clusters = copy.deepcopy(clusters)
        iteration += 1
        print ("Iter : %d" % iteration)
```

三、实验结果

1、 仿真数据实验结果：（可以列出每个聚类中包含的样本，也可以画图显示不同聚类）

Cluster1:

中心为 (1.25 , 1.125),

包含点: { 0, 1, 2, 3, 4, 5, 6, 7 }

Cluster2:

中心为 (7.818182 , 7.3636365)

包含点: { 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 }

2、 MNIST 数据集实验结果：

每个聚类中包含不同类别样本数量统计表

	0	1	2	3	4	5	6	7	8	9
聚类 0	2	11	729	32	1	3	13	6	12	1
聚类 1	32	0	25	13	20	33	696	0	12	1
聚类 2	420	0	17	18	1	35	22	1	4	0
聚类 3	20	3	20	175	4	270	12	4	569	11
聚类 4	8	0	35	5	354	25	157	86	27	225
聚类 5	3	4	6	8	280	65	0	485	35	326
聚类 6	1	5	9	39	267	43	0	408	23	409
聚类 7	22	3	34	630	0	314	7	0	184	13
聚类 8	2	1100	118	69	32	148	60	68	110	25
聚类 9	450	0	1	1	0	5	12	0	3	3