模式识别第五次作业

孙浩淼 201928013229100

Email: sunhm15@gmail.com

1 K-means clustering

1.1 原理

引入如下假设:

- 各类出现的先验概率相等
- 每个样本点以概率为 1 属于一个类别（后验概率 0-1 近似）

此时有性质,

$$P(\omega_i|x_k,\hat{\mu}) = \begin{cases} 1, & x_k \in \omega_i \\ 0, & x_k \notin \omega_i \end{cases} \tag{1}$$

$$\hat{P}(\omega_i) = \frac{n_i}{n}, \tag{2}$$

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} x_k^{(i)}, \tag{3}$$

$$\hat{\Sigma}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}(x_k^i - \hat{\mu}_i)(x_k^i - \hat{\mu}_i)^T \tag{4}$$

此时样本属于哪一类需要计算 $\|x_k - \hat{\mu}_i\|^2$ 来判断。因此需要通过迭代来得到 c 个高斯成分的均值。测试过程中, 以这些均值作为 c 个类 (簇) 的类中心, 计算每个样本点到类中心的欧氏距离, 将样本点归入到距离最近的类。从而完成 K-均值聚类的计算工作。

1.2 计算步骤

Algorithm 1 K-均值聚类

```
1: 初始化参数：n, c, μ_1, μ_2, ..., μ_n
2: repeat
3:     采用最近邻方法对 n 个样本进行分类。
4:     根据采样的样本更新 μ_i
5: until μ_i 保持不变
6: return μ_1, μ_2, ..., μ_n
```

1.3 影响因素

- 初始点 $\mu_1, \mu_2, ..., \mu_n$
- 聚类类别数目 c
- 采样样本数目 n

2 谱聚类算法

2.1 计算步骤

Algorithm 2 经典算法

Require: 相似矩阵 W, 聚类类别数目 k
计算拉普拉斯矩阵: $L = D - W$
计算 L 的前 k 个特征向量, $u_1, u_2, ..., u_k$
计算新的特征空间 $U = [u_1, u_2, ..., u_k] \in R^{n \times k}$
对于 $i = 1, 2, ..., n$，令 y_i 为新特征空间的第 i 行
利用 k-means 算法将 $\{y_i\}$ 聚为 k 个类别：$A_1, A_2, ..., A_k$
return $A_1, A_2, ..., A_k$

Algorithm 3 Shi 算法

Require: 相似矩阵 W, 聚类类别数目 k
计算拉普拉斯矩阵: $L = D - W$
计算归一化的拉普拉斯矩阵: $L_{norm} = D^{-1}L$
计算 L_{norm} 的前 k 个特征向量, $u_1, u_2, ..., u_k$
计算新的特征空间 $U = [u_1, u_2, ..., u_k] \in R^{n \times k}$
对于 $i = 1, 2, ..., n$，令 y_i 为新特征空间的第 i 行
利用 k-means 算法将 $\{y_i\}$ 聚为 k 个类别：$A_1, A_2, ..., A_k$
return $A_1, A_2, ..., A_k$

Algorithm 4 Ng 算法

Require: 相似矩阵 W, 聚类类别数目 k
计算拉普拉斯矩阵: $L = D - W$
计算归一化的拉普拉斯矩阵 $L_{sym} = D^{-1/2}LD^{-1/2}$
计算 L_sym 的前 k 个特征向量, $u_1, u_2, ..., u_k$
计算新的特征空间 $U = [u_1, u_2, ..., u_k] \in R^{n \times k}$
对于 U 矩阵的每一行进行归一化
对于 $i = 1, 2, ..., n$，令 y_i 为新特征空间的第 i 行
利用 k-means 算法将 $\{y_i\}$ 聚为 k 个类别：$A_1, A_2, ..., A_k$
return $A_1, A_2, ..., A_k$

2.2 影响因素

- 局部连接 k-近邻范围
- 点对权值计算方法
- 归一化方法
- 聚类数目
- 聚类方法

3 划分评估

3.1 平方误差准则

$$|S_1| = 18, \quad |S_2| = 18, \quad |S_3| = 17.3 \tag{5}$$

因此，若采用平方误差准则，第三种划分最好。

3.2 类内散度矩阵行列式最小准则

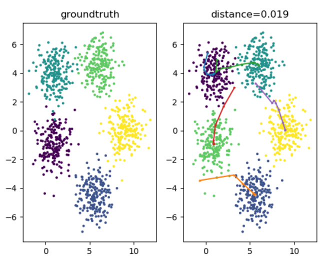
$$|S_{w1}| = 16, \quad |S_{w2}| = 16, \quad |S_{w3}| = 21.3 \tag{6}$$

因此，若采用类内散度矩阵行列式最小准则，前两种划分最好。

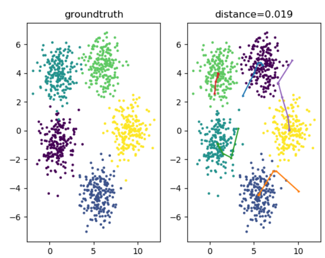
4 编程：K-means

4.1 实验结果

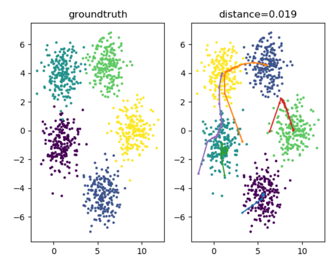
采用不同的初始值，进行了六次实验，实验结果显示如图 1 所示。



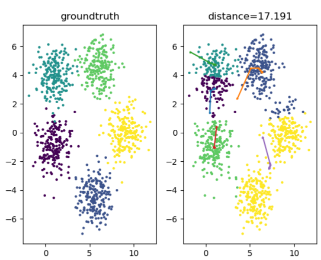
(a) 实验 1



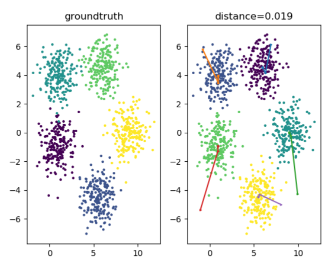
(b) 实验 2



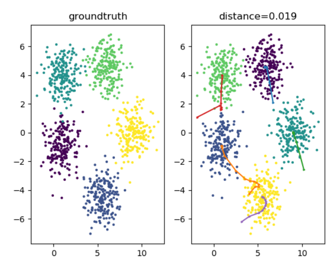
(c) 实验 3



(d) 实验 4



(e) 实验 5



(f) 实验 6

图 1 实验结果

根据实验结果可以看出，大部分初始值都可以收敛到最终的结果，且均方误差较小（0.019），只有一种情况下没有收敛到想要的结果。五个类别分别有（202，201,199,197,201）个数据。

5 编程：谱聚类

5.1 分类结果

结果如图 2 所示，左上为真实数据的特征，右上为谱聚类空间的特征分布，下方两个图时预测的结果。可以看到，谱聚类的效果很好。

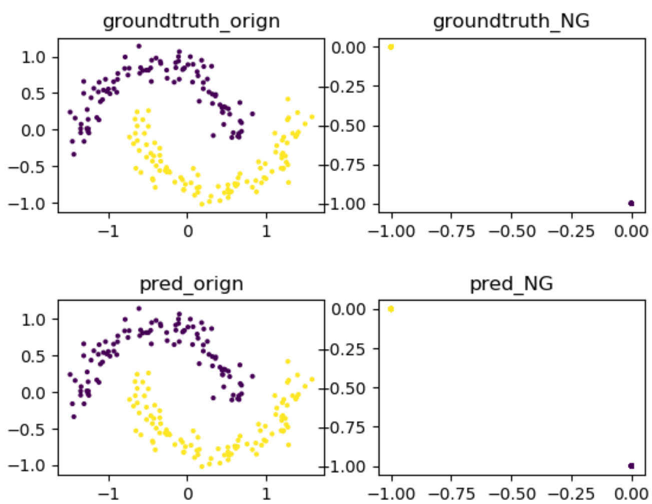
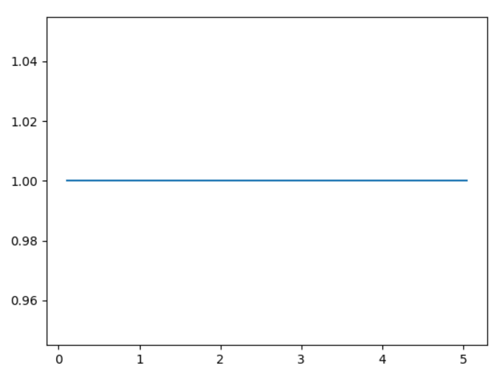
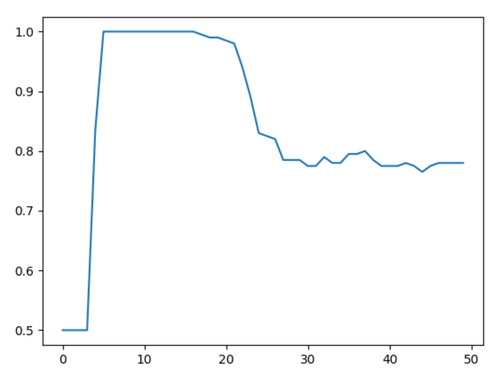


图 2 NG 聚类结果

5.2 不同因素对实验结果的影响



(a) sigma



(b) k

图 3 实验结果

根据实验结果可以看出，sigma 对于聚类结果影响不大，而 k 对聚类结果的影响比较大，k 过小的时候效果一般，而 k 较大又不能很好的区分多个类别（因为不同类别之间会产生混叠）