机器学习第六次作业

131220032 马浩杰

1 1.针对西瓜数据 3.0α计算样本 1-8 上两两之间的闵可夫斯基距离(书 9.18 式)

这道题目，由于计算量比较大，我使用 python 编写程序完成，程序放在 code/MinkowskiDistance 文件夹下.计算结果如下

1.1 P = 1

	1	2	3	4	5	6	7	8
1	0.0	0.161	0.259	0.231	0.386	0.517	0.527	0.509
2	0.161	0.0	0.252	0.224	0.379	0.51	0.52	0.502
3	0.259	0.252	0.0	0.08	0.127	0.258	0.268	0.25
4	0.231	0.224	0.08	0.0	0.155	0.286	0.296	0.278
5	0.386	0.379	0.127	0.155	0.0	0.175	0.141	0.123
6	0.517	0.51	0.258	0.286	0.175	0.0	0.166	0.06
7	0.527	0.52	0.268	0.296	0.141	0.166	0.0	0.106
8	0.509	0.502	0.25	0.278	0.123	0.06	0.106	0.0

1.2 P=2

	1	2	3	4	5	6	7	8
1	0.0	0.114	0.2059	0.1676	0.2827	0.369	0.3787	0.36
2	0.114	0.0	0.1793	0.1758	0.271	0.3962	0.3706	0.3752
3	0.2059	0.1793	0.0	0.0599	0.0921	0.2326	0.1914	0.204
4	0.1676	0.1758	0.0599	0.0	0.1154	0.2204	0.2114	0.2017
5	0.2827	0.271	0.0921	0.1154	0.0	0.1546	0.0999	0.1191
6	0.369	0.3962	0.2326	0.2204	0.1546	0.0	0.1176	0.0428
7	0.3787	0.3706	0.1914	0.2114	0.0999	0.1176	0.0	0.076
8	0.36	0.3752	0.204	0.2017	0.1191	0.0428	0.076	0.0

1.3 P=3

	1	2	3	4	5	6	7	8
1	0.0	0.1016	0.1981	0.1528	0.2597	0.3317	0.3424	0.3208
2	0.1016	0.0	0.1607	0.1683	0.244	0.3774	0.3328	0.3497
3	0.1981	0.1607	0.0	0.0559	0.084	0.2311	0.1722	0.1983
4	0.1528	0.1683	0.0559	0.0	0.1072	0.2091	0.1901	0.184

5	0.2597	0.244	0.084	0.1072	0.0	0.1532	0.0892	0.119
6	0.3317	0.3774	0.2311	0.2091	0.1532	0.0	0.105	0.0385
7	0.3424	0.3328	0.1722	0.1901	0.0892	0.105	0.0	0.0686
8	0.3208	0.3497	0.1983	0.184	0.119	0.0385	0.0686	0.0

2 2.针对西瓜数据 3.0 计算属性“色泽”上两个离散值“青绿”和“乌黑”之间的 VDM 距离

假设簇采用西瓜数据 3.0 的标签，因此只有两个簇，分别对应于好瓜坏瓜

m(色泽，青绿) = 6
m(色泽，乌黑) = 6

m(色泽，青绿，是) = 3
m(色泽，乌黑，是) = 4
m(色泽，青绿，否) = 3
m(色泽，乌黑，否) = 2

0.5 - 2/3 = -0.16666
0.5 - 1/3 = 0.17777
VDM(青绿,乌黑,1) = -0.166 + 0.177 = 0.01111
VDM(青绿,乌黑,2) = (-0.166)^2 + (0.177)^2 = 0.05938
VDM(青绿,乌黑,3) = (-0.166)^3 + (0.177)^3 = 0.00099

3 3.编程实现 k 均值算法在给定的数据集(glass, fourclass 和西瓜数据 4.0)上进行测试。设置三组不同的 k 值(数据类别数目的 1,2,3 倍)使用不同的初始化方式(初始化尽可能不同,如果是随机初始化,请给出随机数种子),分析结果。讨论 k 的取值和初始化方式对聚类结果的影响。本题中任选 1-2 种聚类指标进行结果分析即可.

A.说明

程序使用python2完成k均值算法的实现，在ubuntu16下进行测试.分别对三个文件进行三种k值的测试(注:运行大概需要几分钟)西瓜数据4.0

聚类结果使用聚类性能度量内部指标 DB,Dunn 指数来表示.

B.初始化方式

采用随机初始化，初始化种子设为 0

C.实验测试结果

Fourclass.csv 文件

	DBI	DI
K = 2	1.7591	0.0123
K = 4	2.6043	0.0196
K = 6	50.2268	0.0106

Glass.csv 文件

	DBI	DI
K = 6	323.2095	0.0315
K = 12	3613.1407	0.0464
K = 18	877.5205	0.0252

西瓜数据 4.0.csv

	DBI	DI
K = 2	1.5218	0.1931
K = 4	3.2976	0.3524
K = 6	2.5649	0.2304

D.结果讨论

从表中可以看出对于 fourclass 文件，总体上看 DBI 随 k 的增长越来越大，当 k 变为 6 的时候 DBI 增大了几十倍，而 DI 变化不大. 对于 Glass 文件，总体来看 DBI 比 fourclass 要大上不少，我估计因为 Glass 数据维度比较高，有的维度起到干扰作用，还有离群点的影响导致了这个结果，可能需要降维技术来修正，Glass 文件的数据可以看出 DBI 不一定随着 k 增大而增大，而是不断变化，DI 同样变化不是很大，对于西瓜数据，DBI 也是随着 k 变化变化比较大，而 DI 变化也比较大.

因此可以得出 k 对于 DBI 的影响比较大，而对与 DI 影响较小. 总体看来 k 对总体聚类效果影响非常大，需要谨慎选择.