测试数据科学家聚类技术的40个问题（能力测验和答案）（下）

【AI100 导读】本次测试的重点主要集中在概念、聚类基本原理以及各种技术的实践知识等方面。本文为下部，包括21-40题。上部请查看：测试数据科学家聚类技术的40个问题（能力测验和答案）（上）

Questions & Answers

Q21. 给定具有以下属性的六个点：

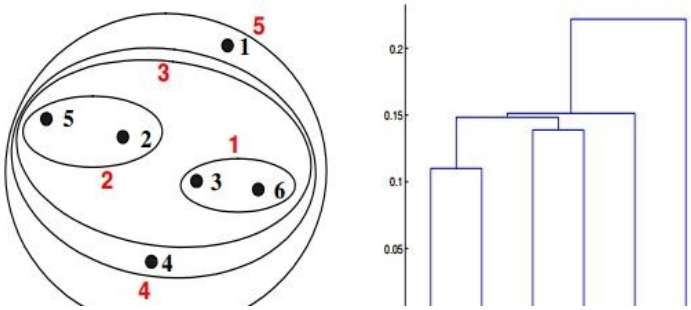
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table： X-Y coordinates of six points.

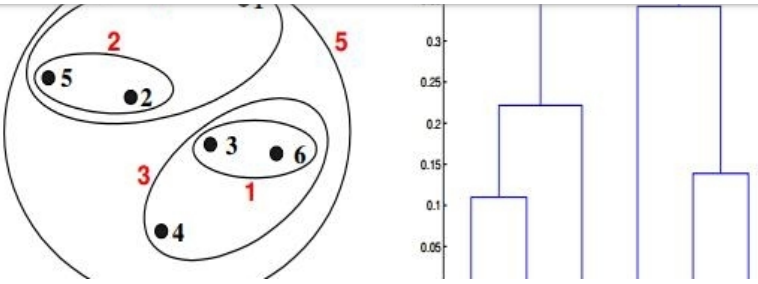
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

如果在层次聚类中使用组平均值接近函数，可以通过下面哪些聚类表示和树形图来描述？

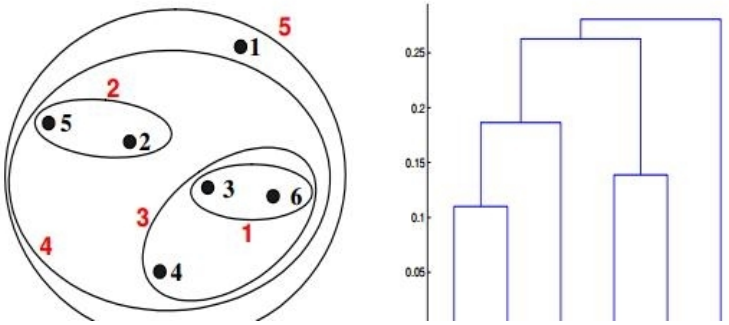
A.



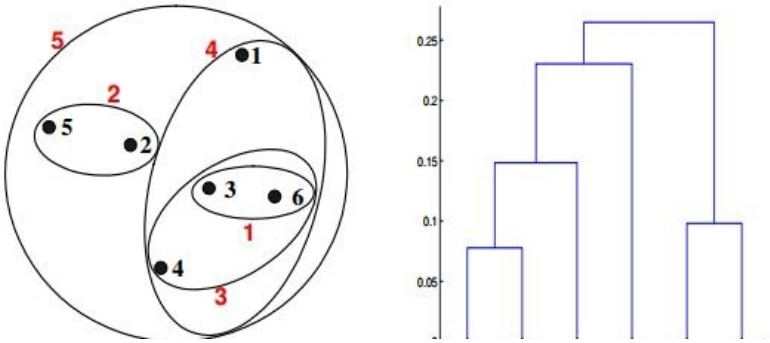
B.



C.



D.



对于层次聚类的的群平均值，两个簇的接近度指的是不同集群中的每一对点对的近似值的平均值。这是最大值和最小值方法之间的中间方法，下面的等式可以表示：

我们来计算一下某些簇之间的距离。dist({3, 6, 4}, {1}) = (0.2218 + 0.3688 + 0.2347)/(3 1) = 0.2751，dist({2, 5}, {1}) = (0.2357 + 0.3421)/(2 1) = 0.2889. dist({3, 6, 4}, {2, 5}) = (0.1483 + 0.2843 + 0.2540 + 0.3921 +0.2042 + 0.2932)/(61) = 0.2637。因为dist({3, 6, 4}, {2, 5})小于dist({3, 6, 4},{1}) 和 dist({2, 5}, {1})，所以这两个簇在第四阶段被合并到了一起。

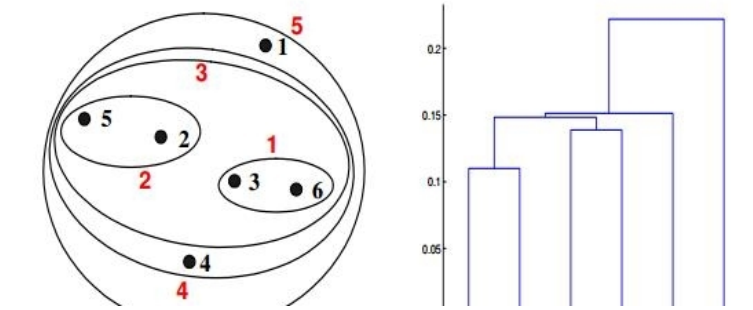
Q22. 给定具有以下属性的六个点：

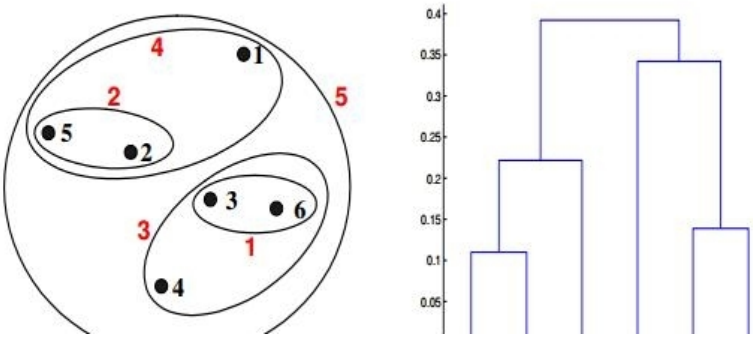
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

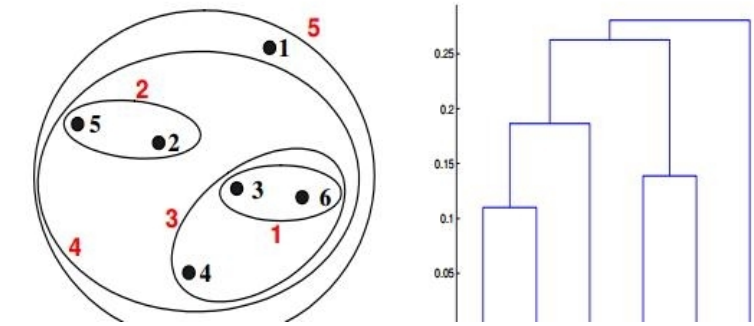
Table ： X-Y coordinates of six points.

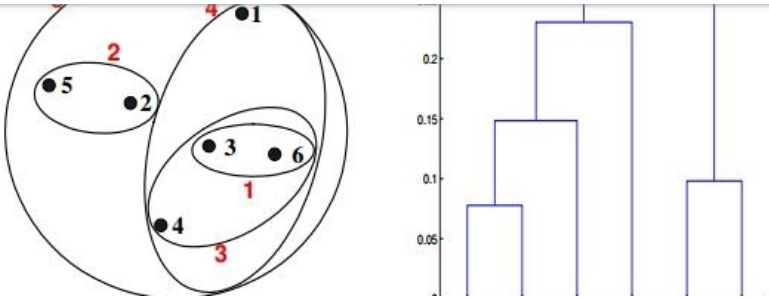
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

如果在层次聚类中使用 Ward 方法的接近函数，可以通过下面哪些聚类表示和树形图来描述？



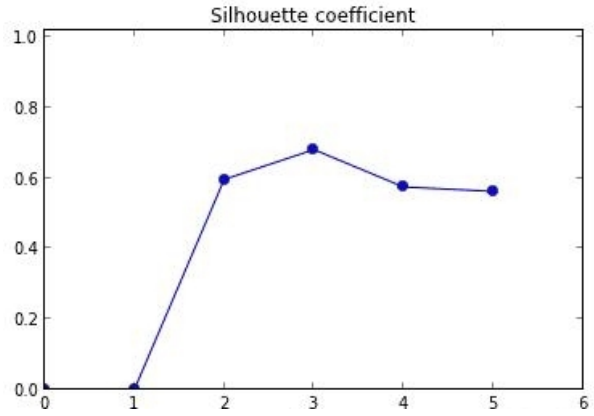






Ward 方法是一种质心算法。质心方法通过计算集群的质心之间的距离来计算两个簇的接近度。对于 Ward 方法来说，两个簇的接近度指的是当两个簇合并时产生的平方误差的增量。在6%的样本数据集中，使用 Ward 方法产生的结果和使用最大值、最小值、组平均值的聚类结果会有所不同。

Q23. 根据下图，簇的数量的最佳选择是？



A. 1

B. 2

C. 3

D. 4

答案：C

轮廓系数旨在将某个对象与自己的簇的相似程度和与其他簇的相似程度进行比较。轮廓系数最高的簇的数量表示簇的数量的最佳选择。

Q24. 在聚类分析之前处理缺失值的有效迭代策略有哪些？

A. 平均值插补法

B. 由最近的值进行分配

C. 用期望最大化算法进行插补

D. 以上都是

上面提到的所有方法都可以有效的在聚类分析之前处理缺失值，但是只有期望最大化算法是可以迭代的。

注意：软性分配可以被认为是被分配给每个聚类的概率：例如 K= 3 时对于一些点 Xn，p1 = 0.7，p2 = 0.2，p3 = 0.1

下面哪些算法允许软性分配？

高斯模糊模型模糊K均值

选项：

C. 1 2

D. 以上都不是

高斯模糊模型和模糊K均值都允许进行软性分配。

Q26. 假设你想使用K均值聚类算法将7个观测值聚类到3个簇中。在第一次迭代簇之后，C1、C2和C3具有以下观测值：

C1: {(2,2), (4,4), (6,6)}

C2: {(0,4), (4,0)}

C3: {(5,5), (9,9)}

如果继续进行第二次迭代，哪一个将成为集群的质心？

A. C1: (4,4), C2: (2,2), C3: (7,7)

B. C1: (6,6), C2: (4,4), C3: (9,9)

C. C1: (2,2), C2: (0,0), C3: (5,5)

答案：A

找到集群中数据点的质心 C1 = ((2+4+6)/3,(2+4+6)/3) = (4, 4)

找到集群中数据点的质心 C2 = ((0+4)/2, (4+0)/2) =(2, 2)

找到集群中数据点的质心 C3 = ((5+9)/2, (5+9)/2) =(7, 7)

因此, C1: (4,4), C2: (2,2), C3: (7,7)

Q27. 假设你想用K均值聚类方法将7个观测值聚类到3个簇中，在第一次迭代簇之后，C1、C2、C3具有以下观测值：

在第二次迭代中，观测点（9，9）到集群质心C1的 Manhattan 距离是？

A. 10

B. 5 * sqrt（2）

C. 13 * sqrt（2）

Q28. 如果聚类分析现在有两个变量V1和V2，对于K均值分析（k=3）的描述，下面哪些是正确的？

如果V1和V2完全相关，簇的质心会在一条直线上如果V1和V2完全不相关，簇的质心会在一条直线上

如果变量V1和V2完全相关，那么所有的数据点都会在同一条直线上，三个簇的质心也会在同一条直线上。

Q29. 应用K均值算法之前，特征缩放是一个很重要的步骤。这是为什么呢？

A. 在距离计算中，它为所有特征赋予相同的权重

B. 不管你用不用特征缩放，你总是会得到相同的簇

C. 在Manhattan距离中，这是重要的步骤，但是Euclidian中则不是

特征缩放保证了在聚类分析中每一个特征都有同样的权重。想象这样一个例子，对体重范围在55-100（kg）和身高在5.6到6.4（英寸）的人进行聚类分析。因为体重的范围远远高于身高的范围，如果不进行缩放，产生的簇会对结果产生误导。因此，使它们具有相同的级别就显得很有必要了，只有这样才能保证聚类结果权重相同。

Q30. 为了在K均值算法中找到簇的最优值，可以使用下面哪些方法？

A. Elbow 法

B. Manhattan 法

C. Ecludian 法

E. 以上都不是

在上面给出的选项中，只有 Elbow 方法是用来寻找簇数的最优值的。方差百分比是一个与簇数有关的函数，Elbow 方法关注的就是方差百分比：分析时应该选择多个簇，以便在添加另一个簇时，不会给出更好的数据建模。

Q31. 关于K均值聚类的描述正确的是？

K均值对簇中心初始化非常敏感初始化不良会导致收敛速度差初始化不良可能导致整体聚集不良

A. 1 3

B. 1 2

C. 2 3

D. 1 2 3

答案：D

上面给出的三个描述都是正确的。K均值对簇中心初始化非常敏感。而且，初始化不良会降低收敛的速度差并会使得整体聚集效果不佳。

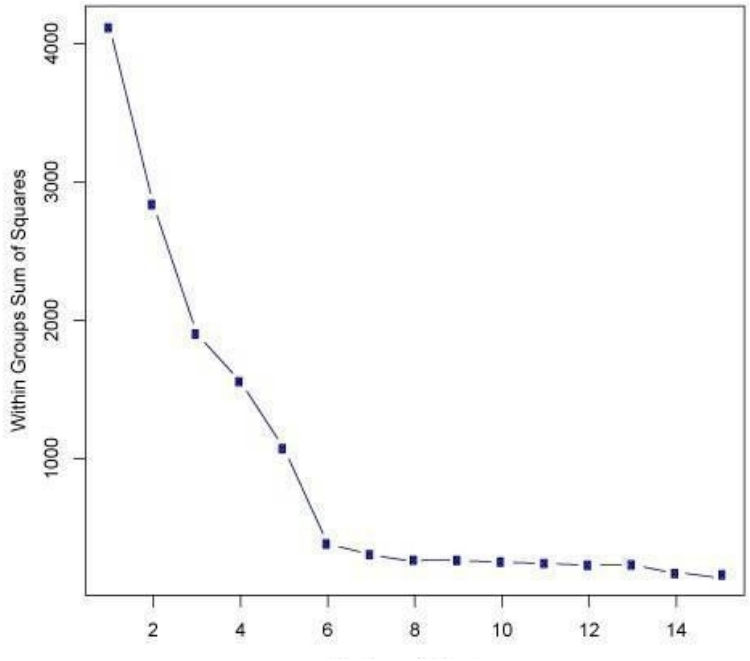
试着运行不同的质心初始化算法调整迭代的次数找出最佳的簇数

A. 2 3

B. 1 3

上面列举的所有选项都是为了获得良好的聚类结果而采用的标准实践。

Q33. 根据下图的结果，簇的数量的最好选择是？



A. 5

B. 6

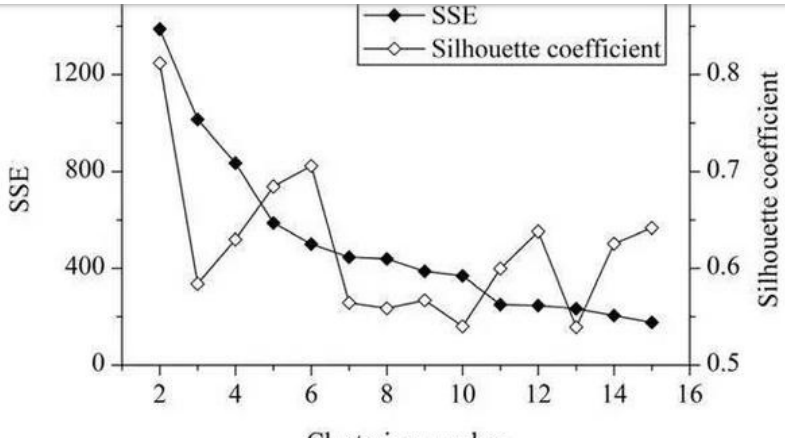
C. 14

D. 大于14

答案：B

根据上面的结果，使用 elbow 方法的簇数的最优选择是6。

Q34. 根据下图的结果，簇的数量的最好选择是？



A. 2

B. 4

C. 6

D. 8

一般来说，平均轮廓系数越高，聚类的质量也相对较好。在这道题中，对于研究区域的网格单元，最优聚类数应该是2，这时平均轮廓系数的值最高。但是，聚类结果（k=2）的 SSE 值太大了。当 k=6 时，SEE 的值会低很多，但此时平均轮廓系数的值非常高，仅仅比 k=2 时的值低一点。因此，k=6 是最佳的选择。

Q35. 对于在初始化中使用了 Forgy 方法的K均值算法，下面哪个顺序是正确的？

指定簇的数量随机分配簇的质心将每个数据点分配给最近的簇质心将每个点重新分配给最近的簇质心重新计算簇的质心

A. 1 2 3 5 4

B. 1 3 2 4 5

C. 2 1 3 4 5

用于K均值初始化的方法是 Forgy 和随机分区。Forgy 方法从数据集中随机选择k个观测值，并将其作为初始值。随机分区方法是先随机为每个观测值分配一个簇，随后进行更新，簇的随机分配点的质心就是计算后得到的初始平均值。

Q36. 如果你要用具有期望最大化算法的多项混合模型将一组数据点聚类到两个集群中，下面有哪些重要的假设？

A. 所有数据点遵循两个高斯分布

B. 所有数据点遵循n个高斯分布(n>2)

C. 所有数据点遵循两个多项分布

D. 所有数据点遵循n个多项分布(n>2)

同的类型。

Q37. 下面对基于质心的K均值聚类分析算法和基于分布的期望最大化聚类分析算法的描述，哪些是不正确的？

都从随机初始化开始都是可迭代算法两者对数据点的假设很强都对异常值敏感期望最大化算法是K均值的特殊情况都需要对所需要的簇数有先验知识结果是不可再现的。

B. 5

C. 1 3

D. 6 7

E. 4 6 7

F. 以上都不是

上面的描述中只有第五个是错的，K均值是期望最大化算法的特殊情况，K均值是在每次迭代中只计算聚类分布的质心。

Q38. 下面关于 DBSCAN 聚类算法的描述不正确的是？

集群中的数据点必须处于到核心点的距离阈限内它对数据空间中数据点的分布有很强的假设它具有相当高的时间复杂度O(n3)它不需要预先知道期望出现的簇的数量它对于异常值具有强大的作用

C. 4

D. 2 3

E. 1 5

F. 1 3 5

DBSCAN 可以形成任意形状的聚类，数据点在数据空间的分布很难预测。

DBSCAN 有比较低的时间复杂度 O(n log n)。

Q39. 以下哪项的F分数存在上限和下限？

A. [0,1]

B. (0,1)

C. [-1,1]

F分数的最小可能值是0，最大可能值是1。1表示每个数据点都被分配给了正确的聚类，0表示聚类分析的旋进和（或）回调为0。在聚类分析中，我们期望出现的是F分数的高值。

Q40. 下面是对6000个数据点进行聚类分析后聚集成的3个簇：A、B和C：

		A	B	C	SUM
Predicted	A	600	400	200	1200
	B	1000	1200	200	2400
	C	400	400	1600	2400
	SUM	2000	2000	2000	

集群B的F1分数是多少？

A. 3

C. 5

D. 6

True Positive, TP = 1200

True Negative, TN = 600 + 1600 = 2200

False Positive, FP = 1000 + 200 = 1200

False Negative, FN = 400 + 400 = 800

因此，

Precision = TP / (TP + FP) = 0.5

Recall = TP / (TP + FN) = 0.6

最后，

F1 = 2 * (Precision * Recall)/(Precision + recall) = 0.54 ~ 0.5

结语

真心希望这次测试的答案对你有用。本次测试的重点主要集中在概念、聚类基本原理以及各种技术的实践知识等方面。

测试数据科学家聚类技术的40个问题（能力测验和答案）（上）

本文作者 Saurav Kaushik 是数据科学爱好者，还有一年他就从新德里 MAIT 毕业了，喜欢使用机器学习和分析来解决复杂的数据问题。

本文由 AI100 编译，转载需得到本公众号同意。

编译：AI100

原文链接：https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/