蒋豆芽

机器学习——聚类 已购

来自 【机器学习面试题汇总与解析（蒋豆芽面试题总结）】 | 37 浏览 | 0 回复 | 2021-05-23

蒋豆芽

+关注

机器学习面试题汇总与解析——聚类

1. **k-means介绍一下**☆☆☆☆☆
2. **k-means优缺点**☆☆☆☆☆
3. **k-means的簇怎么选**☆☆☆☆☆
4. **k-means如何调优**☆☆☆☆☆
5. **知道哪些聚类模型**☆☆☆☆☆
6. **K-means的过程**☆☆☆☆☆
7. **K-means如何选取K值**☆☆☆☆☆
8. **kmeans聚类如何选择初始点**☆☆☆☆☆
9. **kmeans聚类，聚的是特征还是样本？特征的距离如何计算？**☆☆☆☆☆
10. **聚类算法知道哪些**☆☆☆☆☆
11. **Kmeans算法和EM算法的关系**☆☆☆☆☆
12. **写Kmeans代码**☆☆☆☆☆

==========

- 本专栏适合于**Python已经入门**的学生或人士，有一定的编程基础。
- 本专栏适合于**算法工程师、机器学习、图像处理求职**的学生或人士。
- 本专栏针对面试题答案进行了**优化，尽量做到好记、言简意赅。这才是一份面试题总结的正确打开方式。这样才方便背诵**
- 如专栏内容有错漏，欢迎在评论区指出或私聊我更改，一起学习，共同进步。
- **相信大家都有着高尚的灵魂，请尊重我的知识产权，未经允许严禁各类机构和个人转载、传阅本专栏的内容。**

==========

关于**机器学习算法**书籍，我强烈推荐一本**《百面机器学习算法工程师带你面试》**，这个就很类似面经，还有讲解，写得比较好。私聊我进群。

蒋豆芽

关于**聚类**，先学习一下理论知识：推荐看看文章，**《百面机器学习算法工程师带你面试》**里面关于聚类的内容，写得最好

参考资料

五种聚类方法：https://www.sohu.com/a/225353030_99992181

聚类：https://www.jianshu.com/p/4f032dccdcef

读者可以把参考文章看看

个人理解

聚类方法其实有很多，但**k-means**是最简单、最常考的一种。

k-means的优缺点

优点：

1. 原理简单，实现容易

缺点：

1. 收敛较慢
2. 算法时间复杂度比较高 O(nkt)
3. 不能发现非凸形状的簇
4. 需要事先确定超参数K
5. 对噪声和离群点敏感
6. 结果不一定是全局最优，只能保证局部最优

1. **k-means介绍一下**☆☆☆☆☆

   参考回答

   基本**K-Means**算法的思想很简单，事先确定**常数K**，**常数K**意味着最终的聚类**类别数**，首先随机选定初始点为质心，并通过计算**每一个样本**与**质心**之间的相似度(这里为**欧式距离**)，将样本点归到最相似的类中，接着，重新计算每个类的**质心**(即为类中心)，重复这样的过程，直到**质心**不再改变，最终就确定了**每个样本所属的类别**以及**每个类的质心**。由于每次都要计算所有的样本与每一个质心之间的相似度，故在大规模的数据集上，**K-Means**算法的收敛速度比较慢。

   答案解析

蒋豆芽

类似的问题还有：

2. **k-means优缺点**☆☆☆☆☆

   参考回答

   k-means的优缺点

   优点：

   1. 原理简单，实现容易

   缺点：

   1. 收敛较慢
   2. 算法时间复杂度比较高 O(nkt)
   3. 不能发现非凸形状的簇
   4. 需要事先确定超参数K
   5. 对噪声和离群点敏感
   6. 结果不一定是全局最优，只能保证局部最优

   答案解析

   无。

3. **k-means的簇怎么选**☆☆☆☆☆

   参考回答

   手肘法

   答案解析

   无。

4. **k-means如何调优**☆☆☆☆☆

   参考回答

   1. **数据归一化和离群点的处理**

      上面也说了**k-means**是根据**欧式距离**来度量数据的划分，均值和方差大的数据会对结果有致命的影响。同时，少量的噪声也会对均值产生较大的影响，导致中心偏移。所以在聚类前一定要对数据做处理。

蒋豆芽

答案解析

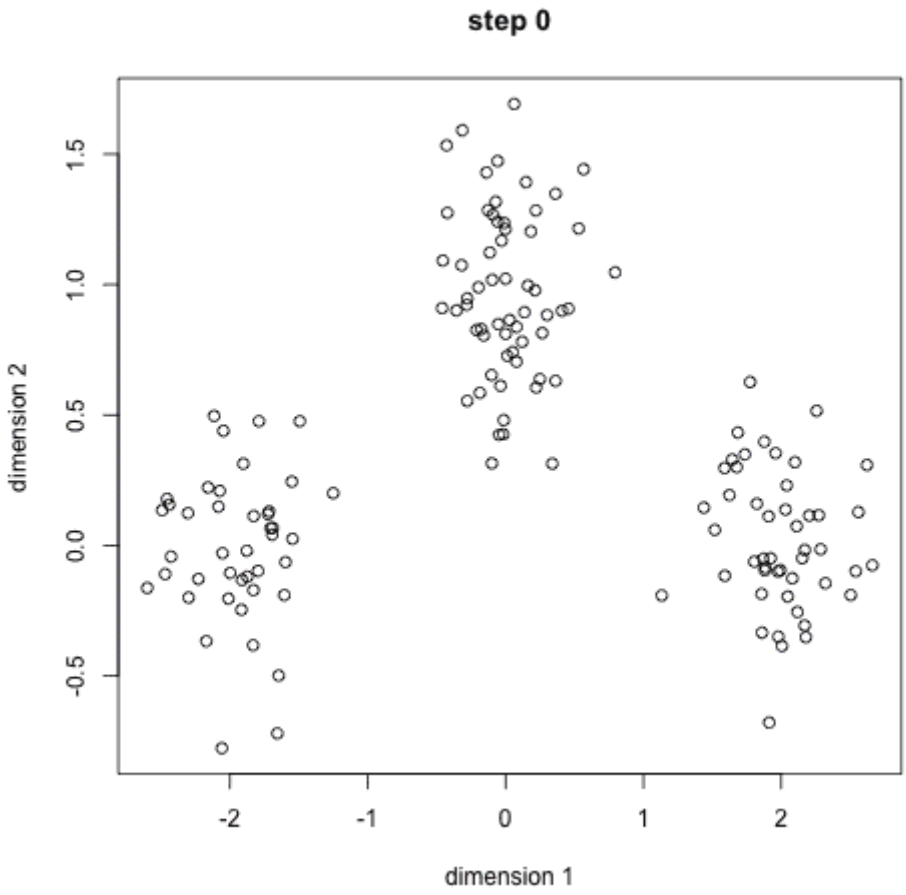
无。

5. **知道哪些聚类模型** ☆☆☆☆☆

参考回答

1. **K-Means聚类**

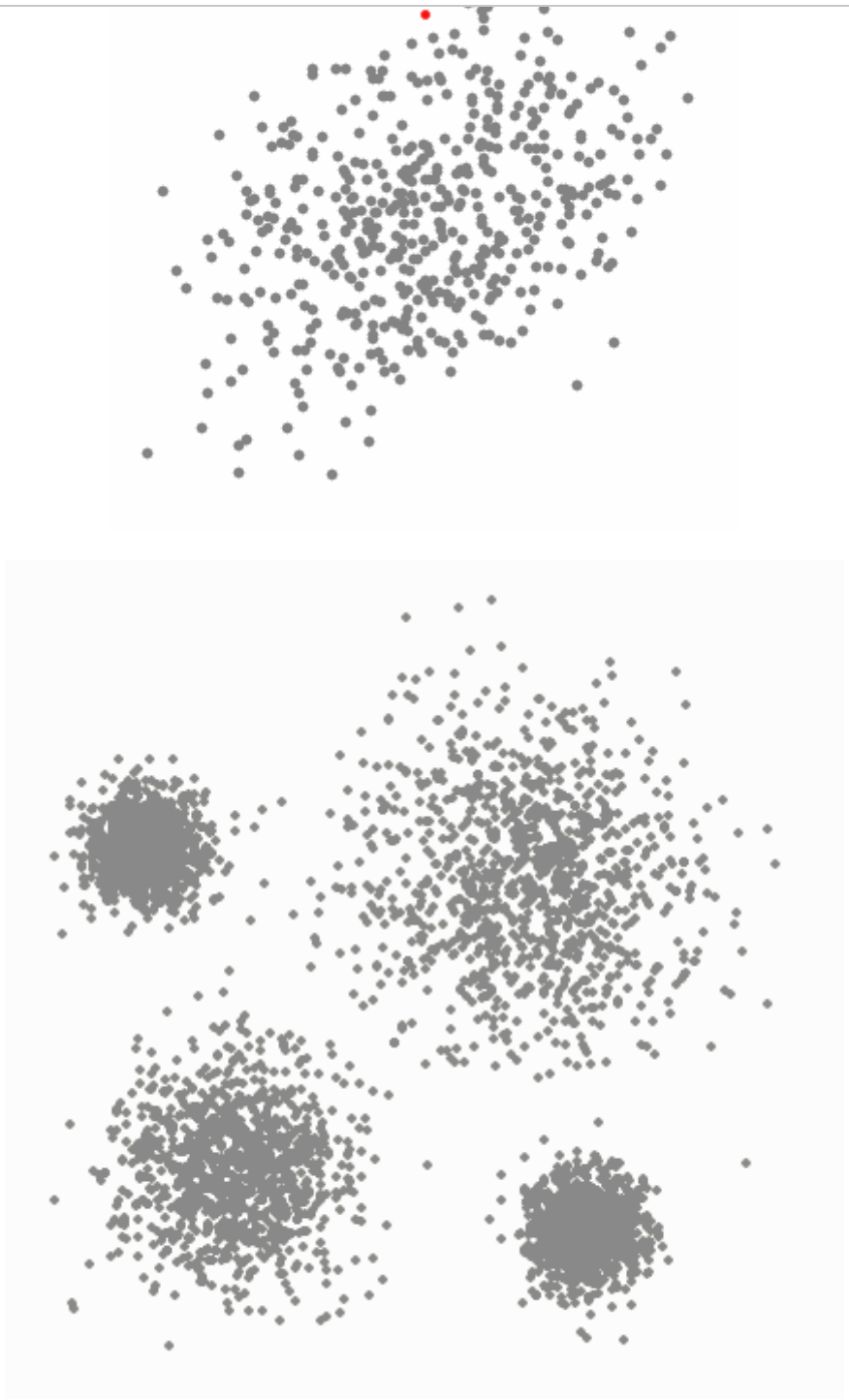
基本**K-Means**算法的思想很简单，事先确定**常数K**，**常数K**意味着最终的聚类**类别数**，首先随机选定初始点为质心，并通过计算**每一个样本**与**质心**之间的相似度(这里为**欧式距离**)，将样本点归到最相似的类中，接着，重新计算每个类的**质心**(即为类中心)，重复这样的过程，直到**质心**不再改变，最终就确定了**每个样本所属的类别**以及**每个类的质心**。由于每次都要计算所有的样本与每一个质心之间的相似度，故在大规模的数据集上，**K-Means**算法的收敛速度比较慢。



2. **均值漂移聚类**

Mean-Shift聚类是基于滑动窗口的算法，试图找到数据点的密集区域。这是一种基于质心的算法，意味着其目标是定位每个簇的中心点，通过将滑动窗口的均值点作为候选点来迭代更新中心点。在后处理阶段将消除近似重复的窗口，最终形成一组中心点及其相应的簇。

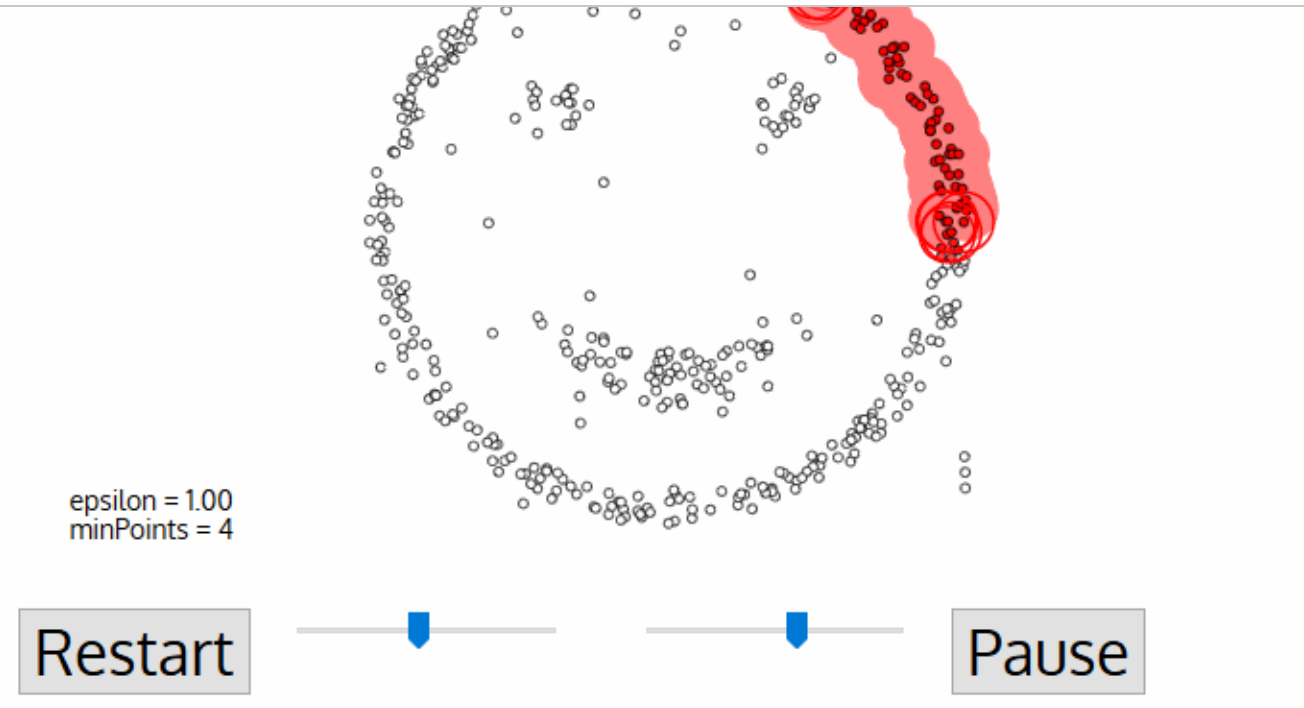
蒋豆芽



与**K-means聚类**相比，**Mean-Shift**的最大优势就是可以**自动**发现簇的数量而不需要人工选择。簇的中心向最大密度点聚合的事实也是非常令人满意的，因为它可被非常直观地理解并很自然地契合数据驱动。

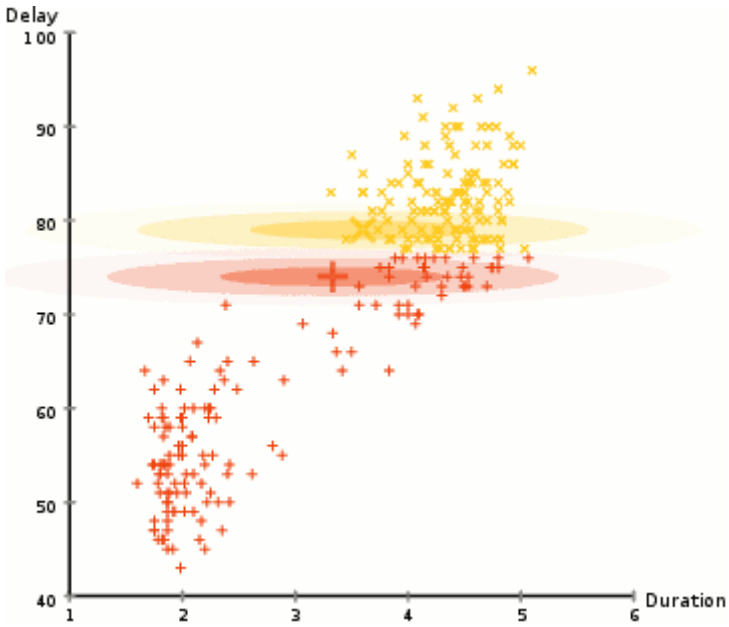
3. **具噪声基于密度的空间聚类算法**

   DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一种基于密度的聚类算法，类似于Mean-Shift，但具有一些显著的优点。



4. **高斯混合模型的期望最大化聚类**

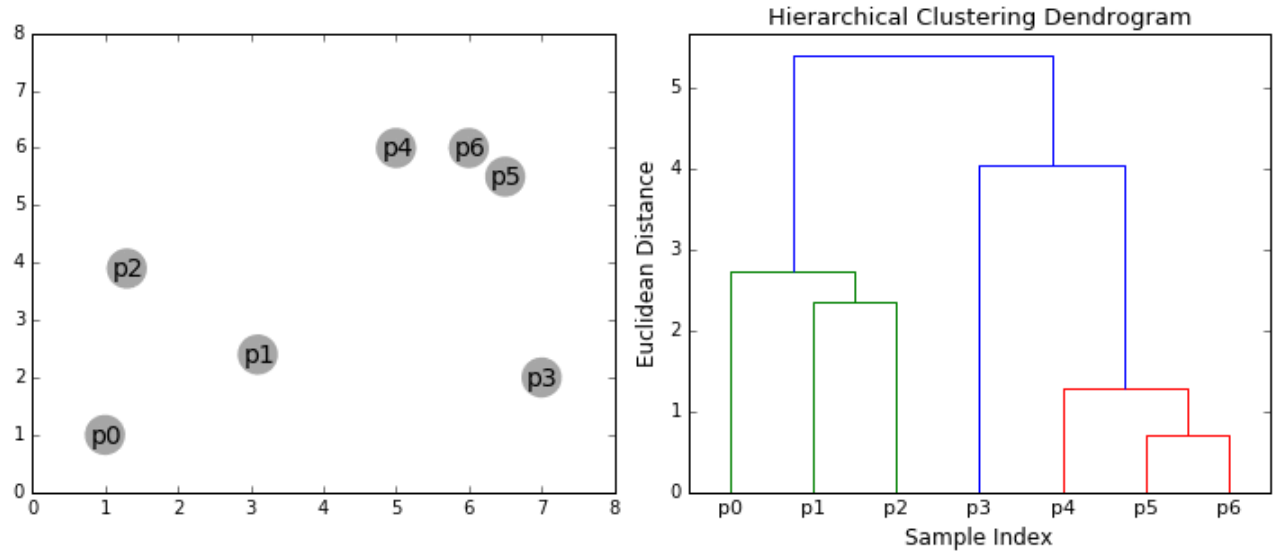
K-Means的主要缺点之一是其简单地使用了**平均值**作为簇的中心。 高斯混合模型（GMMs）相比于K-Means来说有更多的灵活性。 对于GMMs，我们假设数据点是服从高斯分布的（对于用均值进行聚类，这一假设是个相对较弱的限制）。 这样，我们有两个参数来描述簇的形状：均值和标准差！ 以二维为例，这意味着簇可以采用任何类型的椭圆形（因为我们在x和y方向都有标准偏差）。 **因此，每个簇都有一个高斯分布。**



5. **凝聚层次聚类**

分层聚类算法实际上分为两类：自上而下或自下而上。自下而上算法首先将每个数据点视为单个簇，然后不断合并（或聚合）成对的簇，直到所有簇合并成一个包含所有数据点的簇。因此自下而上的层次聚类被称为分层凝聚聚类或HAC。该簇的层次结构被表示为树（或树状图）。

蒋豆芽



分层聚类不要求我们指定聚类的数量，因为我们在构建一棵树，我们甚至可以选择哪个数量的簇看起来最好。另外，该算法对距离度量的选择不敏感，它们的效果都趋于相同，而对其他聚类算法而言，距离度量的选择则是至关重要的。

分层聚类方法的一个特别好的应用是源数据具有层次结构并且用户想要恢复其层次结构，其他聚类算法则无法做到这一点。这种层次聚类是以较低的效率为代价实现的，与**K-Means**和**GMM**的线性复杂性不同，它具有O（n3）的时间复杂度。

答案解析

无。

6. **K-means的过程** ☆☆☆☆☆

参考回答

基本**K-Means**算法的思想很简单，事先确定**常数K**，**常数K**意味着最终的聚类**类别数**，首先随机选定初始点为质心，并通过计算**每一个样本**与**质心**之间的相似度(这里为**欧式距离**)，将样本点归到最相似的类中，接着，重新计算每个类的**质心**(即为类中心)，重复这样的过程，直到**质心**不再改变，最终就确定了**每个样本所属的类别**以及**每个类的质心**。由于每次都要计算所有的样本与每一个质心之间的相似度，故在大规模的数据集上，**K-Means**算法的收敛速度比较慢。

答案解析

无。

7. **K-means如何选取K值** ☆☆☆☆☆

参考回答

蒋豆芽

答案解析

无。

8. **kmeans聚类如何选择初始点** ☆ ☆ ☆ ☆ ☆

   参考回答

   1. **初始点随机分布**
   2. **kmeans++方法**：在一开始确定**簇**时，让所有**簇**中心坐标**两两**距离最远。

   答案解析

   无。

9. **kmeans聚类，聚的是特征还是样本？特征的距离如何计算？** ☆ ☆ ☆ ☆ ☆

   参考回答

   聚的是**特征**。

   特征的距离计算方法一般是**方差**。

   答案解析

   聚类的**核心思想**是将具有**相似特征**的事物给聚在一起，也就是说聚类算法最终只能告诉我们哪些样本属于同一个类别，而不能告诉我们每个样本具体属于什么类别。

10. **聚类算法知道哪些** ☆ ☆ ☆ ☆ ☆

    参考回答

    答案参考上面。

    答案解析

    无。

11. **Kmeans算法和EM算法的关系** ☆ ☆ ☆ ☆ ☆

    参考回答

蒋豆芽

答案解析

EM算法是这样，假设我们想估计知道A和B两个参数，在开始状态下二者都是未知的，但如果知道了A的信息就可以得到B的信息，反过来知道了B也就得到了A。可以考虑首先赋予A某种初值，以此得到B的估计值，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到收敛为止。

12. **写Kmeans代码** ☆☆☆☆☆

参考回答

答案参考文章内容。https://www.cnblogs.com/ahu-lichang/p/7161613.html

答案解析

无。

资源分享 python 机器学习 算法工程师 春秋招 面试题 软件开发 面经

举报

收藏 赞

相关专栏

机器学习面试题汇总与解析（蒋豆芽面试题总结）

27篇文章 | 90订阅

已订阅

0条评论 默认排序

没有回复

请留下你的观点吧~

发布

蒋豆芽

牛客博客，记录你的成长

关于博客 | 意见反馈 | 免责声明 | 牛客网首页