sklearn机器学习之kmeans

原创 NGSHotpot NGSHotpot机器深度学习生信 2017-06-17

sklearn机器学习之kmeans

帮助

sklearn 是NGS Hotpot知道的最好的python machine learning package，这个最好的评价标准是：1）性能稳定；2）持续更新；3）文档完整；4）代码优美且一致性很高。NGS Hotpot通过学习sklearn的文档收获匪浅，在自己的课题中也没少用sklearn去完成一些有意思的分析。虽然近期deep learning 大火，所有方法都有往那个方向靠的趋势，但在NGS Hotpot朴素的观点中，可能并不存在完美的普适的方法，只有针对问题比较合适的方法，经典方法（经过许多人填坑的方法）依旧有其魅力。

从本篇开始，我们将一篇介绍一个sklearn集成的算法以及应用，希望大家继续共同进步。 本篇从较为经典的k-means开始。

算法思想

1.数学本质

k-means把含有N个样本的数据集X 切分成k个类别 C，对C中的每个类其中心点为u, 最终使组内到中心点的平方和最小：

$$\sum_{i=0}^{N}\sum_{j=0}^{k} min(||x_{i,j}-\mu_j||^2)$$

2. 算法流程与伪代码

这是最简单的思想

```
随机选取k个初始点作为每个类的中心点

当任意一点的分类结果发生改变时

        对每个点

                对每个中心点

                        计算点到中心点的距离

                将点分类到距离最近的中心点所在的类

        对每个类计算新的中心点
```

这是最简单的思想

1. 对于初学者，怎么选择最终聚类的数量k是个问题；
2. 随机选取的初始点可能会导致得到的结果是个局部最优而非全局最优；

针对问题1，我们下一篇会介绍几个参考方法，例如Silhouette coefficient；对于问题2， 可行的方法包括多跑几次，用不同的初始点来看结果是否稳定，例如计算生物学常用的cluster 3.0选项里面的-r，另一种比较好的想法就是选择距离比较远的的点当作初始点，而这就是kmeans++的思想。

然后，其实针对上面两个问题，老版本的sklearn的文档里面其实提供过一个非常好的解决方法，即首先对数据进行PCA，找到解释variance 权重比较大的前几个components, 选这个几个components里面关键的点得到初始点，用这些初始点去跑kmeans。NGS Hotpot切身实践过，对于比较大的数据，这个方法很靠谱，由于初始点是确定的只需要跑一次所以效率很高结果也是确定性的。

使用案例

使用案例太多了，翻翻paper到处有，这里就不着墨了。

Github

使用电脑的同学可以通过Github链接直接查看高清版本！
NGSHotpot Github:
https://github.com/NGSHotpot
本文链接：
https://github.com/NGSHotpot/sklearn/blob/master/sklearn:%20k-means.md