李航统计学习：

第三章讨论部分：

读书报告内容：

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：

问题1:构造kd树的时候，选择超平面x轴的顺序对构造结果是否会有影响，即图3.4先以x2进行构造结果会不会不一样？

讨论后的理解：

对于分割超平面来说是不影响的，因为就像我们去切割蛋糕，必然不可能因为先横着切割，再竖着切割和先竖着切割再横着切割而有所不同的。但是从构造数据结构的角度讲，本质上是构造了一个平衡二叉树，从树的角度讲，不同选择x轴的顺序必然根节点就不一样，二叉树也必然不同。

由此引发的，在通过kd树寻找最近邻的时候，他们的实际时间复杂度一定是不同的。

1. 提出的问题2：

kd树有什么不足，最差情况举例？

讨论后的理解：

不足1：

Kd树只是比较平衡的树，并不一定是最优的时间复杂度的数据结构

不足2：

Kd树在k等于n的时候，效果非常的差，基本上等于对所有的样本进行线性扫描

不足3：

Kd树在最近邻预测的时候，超矩形和超球面易于相交，导致会有许多棱角相交而有一些，无关多余的搜索。

二、（必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：

问题3：P54平衡时的kd树未必是最优的，为何？

自己的理解：

因为平衡时候的kd树，本质上是一个平衡的二叉树，它只能保证在期望平均值的情况下达到一个比较优秀的解决，但是可能有这么一个情况，一个极度不平衡的二叉树，在极端情况下刚好可以通过一两次回溯找到最近邻，那么显然平衡时的kd树未必是最优的。

问题4：为什么选择数据的中位数作为切分点，有没有更针对性的策略

自己的理解：

选择数据的中位数作为切分点是因为，我们要获得的kd树本质上就是一个平衡二叉树，选择数据的中位数作为切分点只是因为这样可以让构成的kd树的左右两边子树更加的均匀。

问题5：Kd树的构造结果唯一吗？

自己的理解：

Kd树的构造结果必然不是唯一的，就想之前问题1说的，选择不同x轴切割的顺序会对kd树的构造产生不同的结果。

对于分割超平面来说是不影响的，因为就像我们去切割蛋糕，必然不可能因为先横着切割，再竖着切割和先竖着切割再横着切割而有所不同的。但是从构造数据结构的角度讲，本质上是构造了一个平衡二叉树，从树的角度讲，不同选择x轴的顺序必然根节点就不一样，二叉树也必然不同。

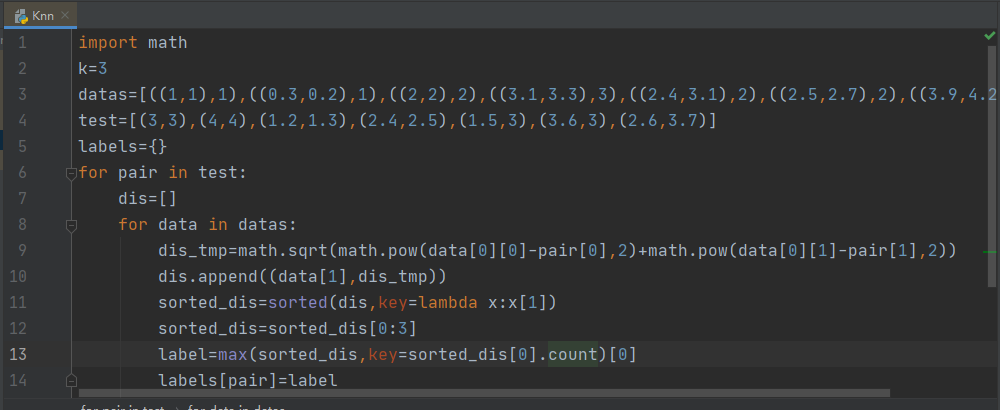
由此引发的，在通过kd树寻找最近邻的时候，他们的实际时间复杂度一定是不同的。

1. （必填）读书计划
2. 本周完成的内容章节：

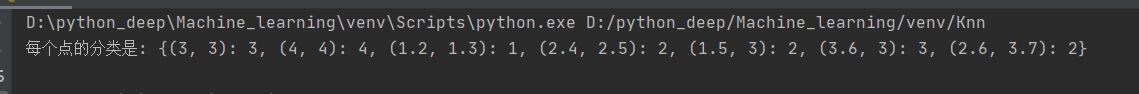
（1）第三章完成并且梳理知识点，寻找问题，自己思考并且于小组会之前完成了自己的思考。

2、下周计划：第四章的阅读.

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）



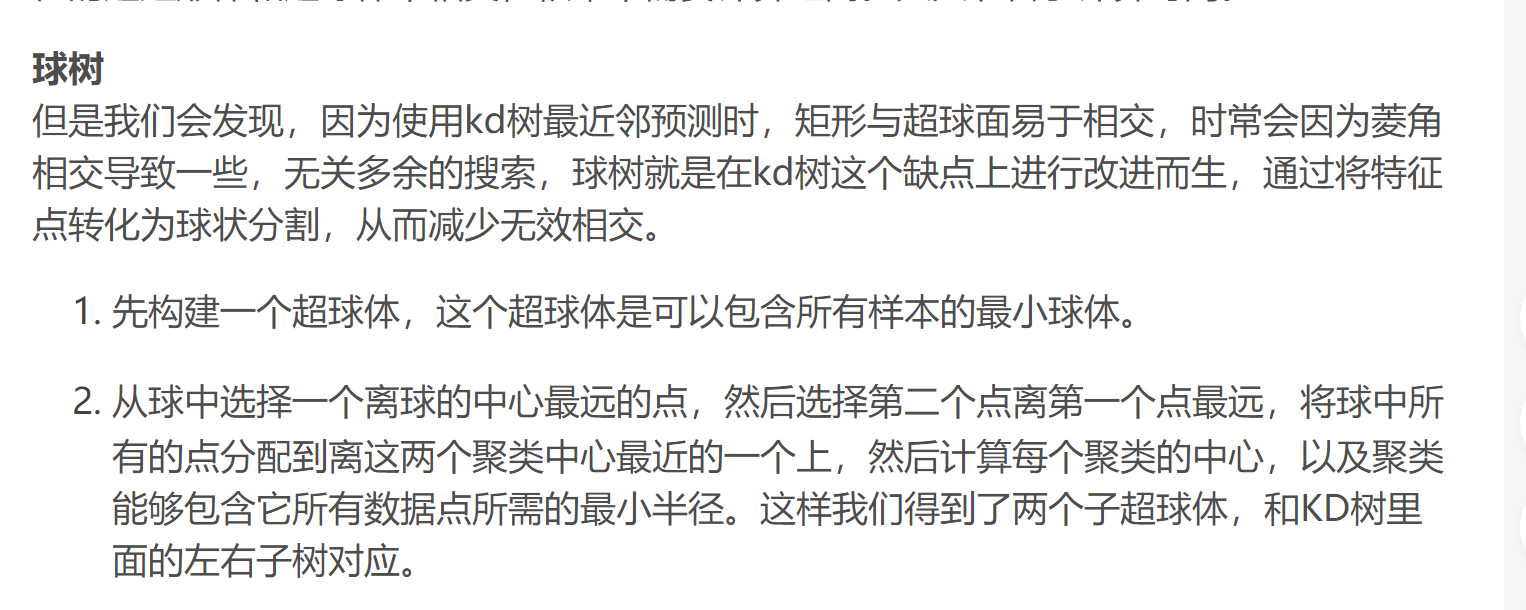
数据结果：



读书摘要：

Kd树是一种便于对k维空间中的数据进行快速检索的数据结构。Kd树是二叉树，表示对k维空间的一个划分，其每个节点对于k维空间划分中的一个超矩形区域。利用kd树可以省去对大部分数据点的搜索，从而减少搜索的计算量。

对于如何解决kd树由于棱角容易重合的问题，提出了一种球树的方法，来减少一些不必要的点带来的计算。



球树是指：



参考博客：<https://blog.csdn.net/u014486460/article/details/86260756>

Knn算法的主要缺点：

计算量大，尤其是特征数非常多的时候；

样本不平衡时，对稀有类别的预测准确率低；

KD树，球树之类的模型建立需要大量的内存；

使用懒惰的学习方法，基本不学习，导致预测时的速度比逻辑回归之类的算法慢；

比起决策树模型，KNN模型的可解释性不强。

知识图谱部分：

工作报告内容：

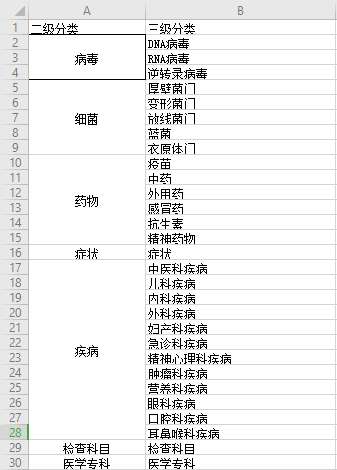
上周在细细粒度fine\_type的选取不同阈值的表现如下：



发现sdtype能够在有噪音的情况下，准确率会有一个非常好的表现，但是不足是召回率较低，即使在0.4较低阈值的情况下，仍然没有一个很好的表现。

经过思考，在sdtype中加入了概念构建时候的type的上下位。

具体如下：



更改后的算法思想：先通过细细粒度的type进行sdtype推断，能够推断出来的数据可以作为正确的结果，对于那些推断不出来的entity，再进行粗细粒度的sdtype推断，如果粗细粒度能够推断出来，再查看细细粒度推断出来的type是否是粗细粒度type的子分类，如果不是则无法正确推断，如果正确那么通过粗细粒度情况下的type和细细粒度的情况加权计算，算出置信度conf。大于阈值则正确推断。

通过这个思想，有效的提高了召回率，但是准确率有所降低，在不同阈值的表现情况如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 阈值 | 准确率P | 召回率R | F1值 |
| 0.3 | 0.583941605839416 | 0.4662004662004662 | 0.5184705119896306 |
| 0.4 | 0.6092307692307692 | 0.46153846153846156 | 0.5251989389920425 |
| 0.5 | 0.6239737274220033 | 0.4428904428904429 | 0.5180640763462849 |
| 0.6 | 0.6455696202531646 | 0.4160839160839161 | 0.5060240963855421 |
| 0.7 | 0.8793650793650793 | 0.32284382284382285 | 0.4722932651321398 |
| 0.8 | 0.9396984924623115 | 0.21794871794871795 | 0.3538315988647115 |

存在问题：

1. 准确率有所降低。可能原因是先验数据本身比较少，不足以进行更准确的推断。同时，一部分是因为，爬取的数据有的就不是医学类的，怎么推断肯定都是错的，需要推断的entity本身不完全是医学域的，这在较低阈值的情况下会进行错误判断。

下周工作内容：

1. 自己调研同时也找吴老师问清楚，寻找别的方法能够进行新冠域的type inference推断。
2. 找到方法试试能够跑起来和修改后的sdtype进行比较
3. 准备看看安卓开发，进行学习，准备做吴老师下周基于知识问答系统的安卓开发。