读书报告

71117203-袁佳怡

1. （必填）自己提出的问题的理解：
2. 提出的问题1：如何更好地理解kd树的最近邻搜索（不是很理解里面的递归回退的过程）

讨论后的理解：在kd树中搜索最近邻，就是先从树根厨出发，根据大小比较值，选择出左偏右偏之后，到达一个叶子节点，然后不断回溯，同时根据超球体判断是否有可能在该节点的另一个子节点上，如果是则对子节点同样地遍历回溯，没有的话，不断更新最小距离，不断地递归回溯继续判断，知道到达根节点，即遍历完了整棵树中可能的最近邻居的点。这个算法主要是因为省略了一部分实例点的搜索，才在一定程度上降低了搜索的时间复杂度。

1. （必填）别人提出的问题的理解
2. 问题2：怎么理解超矩形区域的意思？

自己的理解：在凸优化里面提到，超平面就是二维里面的直线，三维里面的一个面，就是n维欧式空间中的n-1维度的线性子空间。可以简单理解为，垂直于每个坐标系做一个切面形成的高维图形。

1. 问题3：第三章开头说k近邻法是一种基本的分类与回归的方法，那如何将k近邻用于回归当中呢？是在找到k近邻之后求平均值这样的作法吗？

自己的理解：KNN算法不仅可以用于分类，还可以用于回归。通过找出一个样本的k个最近邻居，将这些邻居的某个（些）属性的平均值赋给该样本，就可以得到该样本对应属性的值。

1. 问题4: 在最近邻的搜索算法中，当目标点和“当前最近点”之间的距离为半径构成的圆和另一子节点的区域相交时，会移动到另一个子节点当中递归地进行最近邻搜索。请问是怎么在另一子结点中搜索的？

自己的理解：将搜索目标移动为另一子结点，同样的操作直接遍历另一个子节点。

1. 问题5: 我的问题是k值的选择与数据有没有什么联系，除了交叉验证法多次计算选取最优k值之外有无别的方法？

自己的理解：可能是有一个经验范围，在那个经验范围里面进行交叉验证，选取比较合适的k值。

1. 问题6：为什么说用较小的k值，即较小的邻域中的训练实例进行预测，学习的近似误差会减小，反之是估计误差会减小？

自己的理解：近似误差是训练误差，估计误差是测试集的测试误差。在k近邻算法中，设置的k值越小，得出的模型越复杂，因为k值越小会导致特征空间被划分成更多的子空间（可以理解为模型的项越多）。而k值越大得到的模型其实是越简单的。所以当k值越小，对于训练集的预测更加准确，近似误差越小（因为选择了更加复杂的模型去预测训练集）。当k值越大，对于训练集的预测则不会那么准确，所以近似误差会越大。

1. （必填）读书计划

1、本周完成的内容章节：《统计学习方法》第三章

2、下周计划：《统计学习方法》第四章

四、（选做）读书摘要

1、k近邻算法，简称KNN：给定一个训练数据集，对于新的输入实例，在训练数据集中找到与该实例最接近的k个实例，通过这k个实例投票决定该输入实例的类别。

2、当选取的值较小时，相当于用较小邻域的训练实例进行预测，更容易受噪声干扰（比如邻近的实例点恰好是噪声就会出错），即越小则模型过拟合的风险越大。

当选取的较大时，相当于用较小邻域的训练实例进行预测，这时候与输入实例较远（相似度较小）的训练实例也会对预测产生影响，从而降低模型准确率。

在应用中，k值在比较小的数值范围内取，并且结合交叉验证确定最优k值。

3、当训练集很大时，计算输入实例和每一个训练实例的距离相当耗时。为了提高近邻搜索的效率，我们使用特殊的结构存储训练数据来减少计算距离的次数，比如树方法。

kd树（k-dimension tree）是一种对维空间中的实例点进行存储以便对其进行快速检索的二叉树形数据结构。构造树相当于不断用垂直于坐标轴的超平面将维空间切分，构成一系列的超矩形区域，树上的每一个结点对应于一个超矩形区域。该超矩形区域垂直于当前划分维度的坐标轴，并在该维度上将空间划分为两部分。

根据书本知识，对kd树的构建，和如何运用kd树进行k近邻的搜索，其中主要的思想是利用不断地递归回溯，一定程度上减少了点的遍历，从而使得复杂度降低。复杂度为O(longN)，到这时考虑实例点随机分布的，kd树更适用于训练实例数远大于空间维数时的k近邻搜索，当空间维数接近训练实例数时，它的效率几乎为线性扫描。