统计学习第三次图书报告

报告人：71117415-喻泽弘

读书时间：3.17-3.22

读书进度：统计学习第三章

问题与解答

- 个人提出的问题：
  - 为什么使用较小的k值，学习的近似误差反而会减小，反之是估计误差会减小？

    近似误差是训练误差，估计误差是测试集的测试误差。设定的k值越小，得出的模型是越复杂的，而k值越大，得到的模型其实越简单的，从而估计误差减小，近似误差增大。可以类比于泛化能力这方面，进行理解。
- 别人提出的问题：
  - 怎么理解超矩形区域的意思？

    超方形(Hypercube)（又叫立方形、正测形(MeasurePolytope)）是指正方形和立方体的n维类比（对于正方形，n=2，对于立方体，n=3）。它是一类闭合的、紧致的、凸的图形，它们的1维骨架是由一群在其所在空间对准每个维度整齐排列的等长的线段组成的，其中相对的线段互相平行，而相交于一点的线段则互相正交。在n维空间中单位超方形（棱长为1）的对角线长等于\sqrt{n}.
  - 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）

    这个是从底部向下回退才算搜索完了整个数据集，也就是说，这是一个搜索回溯的过程，回到每个节点都会更新当前值。

下周计划

完成统计学习第四章的阅读

学习总结

- k近邻法是基本且简单的分类与回归方法。K近邻法的基本做法是：对给定的巡礼实例点和输入实例点，首先确定输入实例点的k个最近邻训练实例点，然后利用这k个训练实例点的类的多数，来预测输入实例点的类
- k近邻模型的三要素：距离度量、k值得选择和分类决策规则。常用得距离度量是欧式距离及更一般得L_P距离。k值小时，k近邻模型更加复杂，k值大时，k近邻模型更简单。k值得选择反映了对近似误差与估计误差之间得权衡，通常由交叉验证选择最优得k。常用得分类决策规则时多数表决，对应于经验风险最小化