方骏-2020年3月22日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：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？

讨论后的理解：这里的近似误差指的是训练误差，估计误差指的是测试误差。对于近似误差来说，这里我们认为选择k个最近的点作为训练行为，如果k较小，那么训练产生的误差自然很小，但是由于可选的训练点比较少，那么预测的分类可能会出现错误，所以估计误差会比较大。如果k值较大，那么训练误差自然变大，预测的分类自然准确性会高很多，所以估计误差较小。

1. 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：对于二维来说有线，但是上升到三维就是一个三维空间的平面，超就是多维的空间，即存在于高维空间中的一个平面，类似于二维的平面，三维的平面。

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：通过找出一个样本的k个最近邻居，将这些邻居的某个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。这就是一种k近邻的回归方法。

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：首先，移动到另一个子节点当中，说明该子节点负责的区域存在一个更近点，就必须从该结点往下进行遍历，根据当前最近的距离来筛选移动的分支，按照交集的分支往下遍历，即可找到更近的一个点。然后再以该点为最近点，继续向上递归。

1. 读书计划

1、本周完成的内容章节：《统计机器学习》第三章

2、下周计划：《统计机器学习》第四章

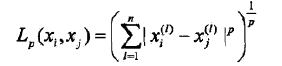
四、读书摘要及理解

1、k近邻法是一种基本分类与回归方法。对于分类问题，k近邻法的输入为实例的特征向量，对应于特征空间的点，输出为实例的类别，可以取多类。分类时，对新的实例，根据其k个最近邻的训练实例的类别，通过多数表决等方式进行预测。可见，k近邻法不具有显式的学习过程。k近邻法实际上利用训练数据集对特征向量空间进行划分。

2、k近邻法使用的模型实际上对应于对特征空间的划分，模型由三个基本要素——距离度量、k值的选择和分类决策规则决定。

2.1、k近邻法中，当训练集、距离度量、k值及分类决策规则确定后，对于任何一个新的输入实例，它所属的类唯一地确定，这就是相当于根据上述要素将特征空间划分为一些子空间，确定子空间里的每个点所属的类，这就是k近邻法形成的一个模型。

2.2、k近邻法的距离度量一般采用的是欧氏距离。特征空间中两个实例点的距离是两个实例点相似程度的反映。同时可以采取更一般的距离：



这里p=1时，就是曼哈顿距离；p=2时，就是欧氏距离。由不同的度量选择的近邻点是不同的，可根据需要去选择。

2.3、k值得选择会对k近邻法的结果产生重大影响。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，近似误差增大。原因在问题中已经说过。所以k值得选择尤为重要，通常通过交叉验证来确定k得取值。

2.4、k近邻法中的分类决策规则往往是多数表决，由输入实例的k个临近的训练实例中的多数类决定输入实例的类。要使误分类率最小即经验风险最小，就要使最大，多数表决规则就等价于经验风险最小化。

3、kd树是一种提高k近邻搜索效率的数据结构。

3.1、kd树的构造方法如下：构造根节点，使根结点对应于k维空间中包含所有实例点的超矩形区域，在超矩形区域上选择一个坐标轴和在此坐标轴上的切分点，确定一个超平面，这个超平面通过选定的切分点并垂直于选定的坐标轴，将当前超矩形区域切分为左右两个子区域；这是，实例被分到两个子区域，这个过程直到子区域内没有实例时终止。选择训练实例点在选定坐标轴上的中位数为切分点，这样的kd树是平衡的。

3.2、给定一个目标点，搜索其最近邻。首先找到包含目标点的叶结点，然后从该叶结点出发，一次回退到父结点，不断查找与目标点最近邻的结点，当确定不可能存在更近的节点时终止。这样搜索就被限制在空间的局部区域上，效率大为提高。

先回退到父结点，如果父结点的另一个子结点的超矩形区域与超球体相交，那么在相交的区域内寻找与目标点更近的实例点，如果存在这样的点，将此点作为新的当前的最近点。算法转到更上一级的父结点，继续该递归过程。