读书报告

71117203袁佳怡

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：提升树与Adaboost之间是什么关系？

讨论后的理解：两个都是提升方法，但他们实现的机制是不同的，两者都是基于加法模型和前向分布算法，但是不同的损失函数解法是不一样的。

前向分布算法+决策树（分类树or回归树）=提升树

提升树是以分类树或者回归树为基本分类器的提升算法，分类问题用二叉分类树，回归问题用二叉回归树，最简单的回归树可看做一个根节点连接的左右子树的二叉树，即所谓的决策树桩。Adaboost是提升思想的算法模型，经典的AdaBoost一般用于分类问题，并没有指定基函数，或者说是分类器，它可以从改变样本的权值的角度和前向分布算法的角度来解释。当确定基函数是回归或分类树时，结合前向分布算法就得到提升树算法。

1. （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）：
2. 问题2：Adaboost算法中的α是负值的话，那么对应的Gm不是也会很大的影响最后的f(x)吗？而且这样的话计算下一轮m+1的时候，误分类的点的权值小于正确分类的点的权值。

自己的理解：在Adaboost中，用若干个弱分类器组成一个强分类器，对于那些学习效果很差的学习器，也就是说它们总是将数据分类错误，实际上在这个时候去α为负数的话就是我们想要的结果。所以权值是可以为负的。

1. 问题3：定理8.1如何说明AdaBoost算法可以在每一轮选取选取适当的Gm使得Zm最小，从而使训练误差下降最快？

自己的理解：在每一轮中都选取让分类误差上界下降最快的那个，每一步都让误差分类的上界最小化，类似于一个夹逼定理的过程，右边是对误差最大分类，形成的误差分类的上界。、

1. 问题4：如何理解强学习是弱学习的充要条件？

自己的理解：每个输入空间都有强学习器，比如可以用一个二叉树，每一个叶节点对于输入控件一个变量的输出结果，弱学习器的话，相当于对强学习器进行剪枝。

1. 问题5：AdaBoost算法的迭代次数如何判断？

自己的理解：对于不断迭代的过程，有精度要求，只要达到了一个要求的值，就可以停止迭代。

1. （必填）读书计划

1、本周完成的内容章节：《统计学习方法》第八章

2、下周计划：《统计学习方法》第九章

四、（选做）读书摘要Adaboost提升方法

1、提升学习：将一系列弱学习算法进行综合与提升。

提升学习要考虑的两个问题：每一轮如何改变训练数据的权值或概率分布；如何将弱分类器组合成一个强分类器。

2、Adaboost算法：是AdaptiveBoost的缩写，表明该算法是具有适应性的提升算法，通过学习一系列弱分类器，线性组合成一个强分类器，在每一轮中更新训练数据的权值分布，提升误分类样本比重，在线性组合中，所有系数和并不为1。Adaboost的训练误差以指数速率下降。Adaboost的算法解释是：模型为加法模型、损失函数为指数函数、学习算法为前向分布算法（每步只学习一个基函数及其系数）。

算法的步骤如下：

1. 给每个训练样本（x1,x2,,,,xN）分配权重，初始权重w1均为1/N。
2. 针对带有权值的样本进行训练，得到模型Gm(初始模型为G1)。
3. 计算模型Gm的误分率
4. 计算模型Gm的系数αm=0.5log[(1-em)/em]
5. 根据误分类e和当前权重向量更新权重向量。
6. 计算组合模型的误分率
7. 当组合模型的误分率或迭代次数低于一定阈值，停止迭代，否则，回到步骤2

3、提升树算法：提升树是指基本模型是分类树或者回归树的提升组合模型，提升树被认为是统计学习中性能最好的模型之一。属于加法模型，采用前向分布算法，以决策树为基函数；

梯度提升树（GBDT）：利用最速下降法的近似方法来实现每一步的优化，关键在于用损失函数的负梯度在当前模型的值作为回归问题中提升树算法中的残差的近似值，每一步以此来估计回归树叶结点区域以拟合残差的近似值，并利用线性搜索估计叶节点区域的值是损失函数最小化，然后更新回归树。