**统计学习方法读书报告**

09017244 郑健雄

**本周进度:**统计学习方法第八章和第九章部分内容，负责8.2讲解。

**下周进度:**统计学习方法第九章和第十章。

1. **问题与解答**

会上讨论的问题：

1.(我提出) 梯度提升算法第一步初始化为什么要估计损失函数极小化的常数值？

讨论结果：之前提升树算法中，初始值取0并不影响算法的运算过程，但是在梯度提升算法中，第一，初始值是为了优化损失函数，而损失函数为平方损失，所以要找一个常数使整体损失极小化。第二，因为梯度提升算法计算残差需要计算梯度，而当初始值为0的时候，其计算出的梯度没什么意义，影响优化的过程。

2. (别人提出)为什么提升树模型中没有权值？

讨论结果： 经过讨论，得出的结论是是否添加权重要看具体的情况，一般来说因为分类树是直接使用Adaboost算法的，是会分配权重的，而回归树则是逐步拟合残差的过程，其误分类率很低，可以不添加权重，也可以添加权重，使后面的树的权重偏低，但是会增加迭代的次数。

3. (别人提出) 提升方法中分类器过多是否会出现过拟合状况？有正则化方法吗？

讨论结果：分类器过多的话，是有可能出现过拟合的，这和其他学习方法是类似的。对于这个问题，可以设定阈值，使得模型在某种精确度下停下来，或者引入惩罚项，使得后来得分类器影响尽可能小。

4. （别人提出）AdaBoost如何拓展至多分类问题？

讨论结果：Adaboost算法可以扩展至多分类问题，其本身是二分类算法，可以把多分类变成二分类，把一个类提出来，另外的类作为一个总类，进行二分类，重复这种步骤即可。

1. **下周计划安排**

下周计划阅读完第九章和第十章。

1. **读书收获**

第八章内容总结：

1. AdaBoost算法的思路是训练多个弱分类器，然后将其线性组合为强分类器，其具体做法是进行迭代构建分类器的过程，将上个过程正确分类的样本权值降低，而提高误分类样本的权值。在构建结束后，对于分类误差率低的分类器，给予高的权值，高的分类器，给予低的权值，然后线性组合。在训练过程中，样本点新的权值依赖于上一个旧的权值，通过计算得出新的权值。

2. Adaboost训练误差分析:

Adaboost算法的训练误差是以指数速率下降的，这是很有吸引力的性质，同时它也不需要提前设定误差率，其可以在算法过程中动态调整。

3.Adaboost的解释:

Adaboost算法可以理解为是前向分布算法的一种特例。其算法模型为加法模型，而损失函数为指数函数。这个解释说明了Adaboost算法的合理性。

4. 提升树模型:

提升树被认为是统计学习中性能最好的算法之一。

其一般有分类树，回归树和一般决策树。分类树的训练实际上使用Adaboost算法。回归树的训练思路是逐步拟合残差，这个类似泰勒展开的思想，用多个近似函数去表征一个完整函数。其回归有很好的效果。

对于指数损失函数和平方损失函数，残差是比较好计算的，但是对于一般损失函数而言，提升方法的优化比较困难。所以提出了近似的求残差方法，并给出了梯度提升算法，其思路是用梯度来近似替代残差，而其他步骤类似于提升算法，也是拟合残差。梯度提升算法有着较广泛的应用。

第八章收获：阅读本章内容时，发现学习方法中应用了很多高等数学的思想以及性质，并且其有着很好的效果，认识到曾经不知道如何使用的微积分有着非常广泛的应用。