本周学习：AdaBoost

集成方法：

AdaBoost算法和GBDT(Gradient Boost Decision Tree，梯度提升决策树)算法是基于Boosting思想的机器学习算法。在Boosting思想中是通过对样本进行不同的赋值，对错误学习的样本的权重设置的较大，这样，在后续的学习中集中处理难学的样本，最终得到一系列的预测结果，每个预测结果有一个权重，较大的权重表示该预测效果较好。

AdaBoost思想：

  AdaBoost算法是基于Boosting思想的机器学习算法，其中AdaBoost是Adaptive Boosting的缩写，AdaBoost是一种迭代型的算法，其核心思想是针对同一个训练集训练不同的学习算法，即弱学习算法，然后将这些弱学习算法集合起来，构造一个更强的最终学习算法。

构造强学习算法的步骤：

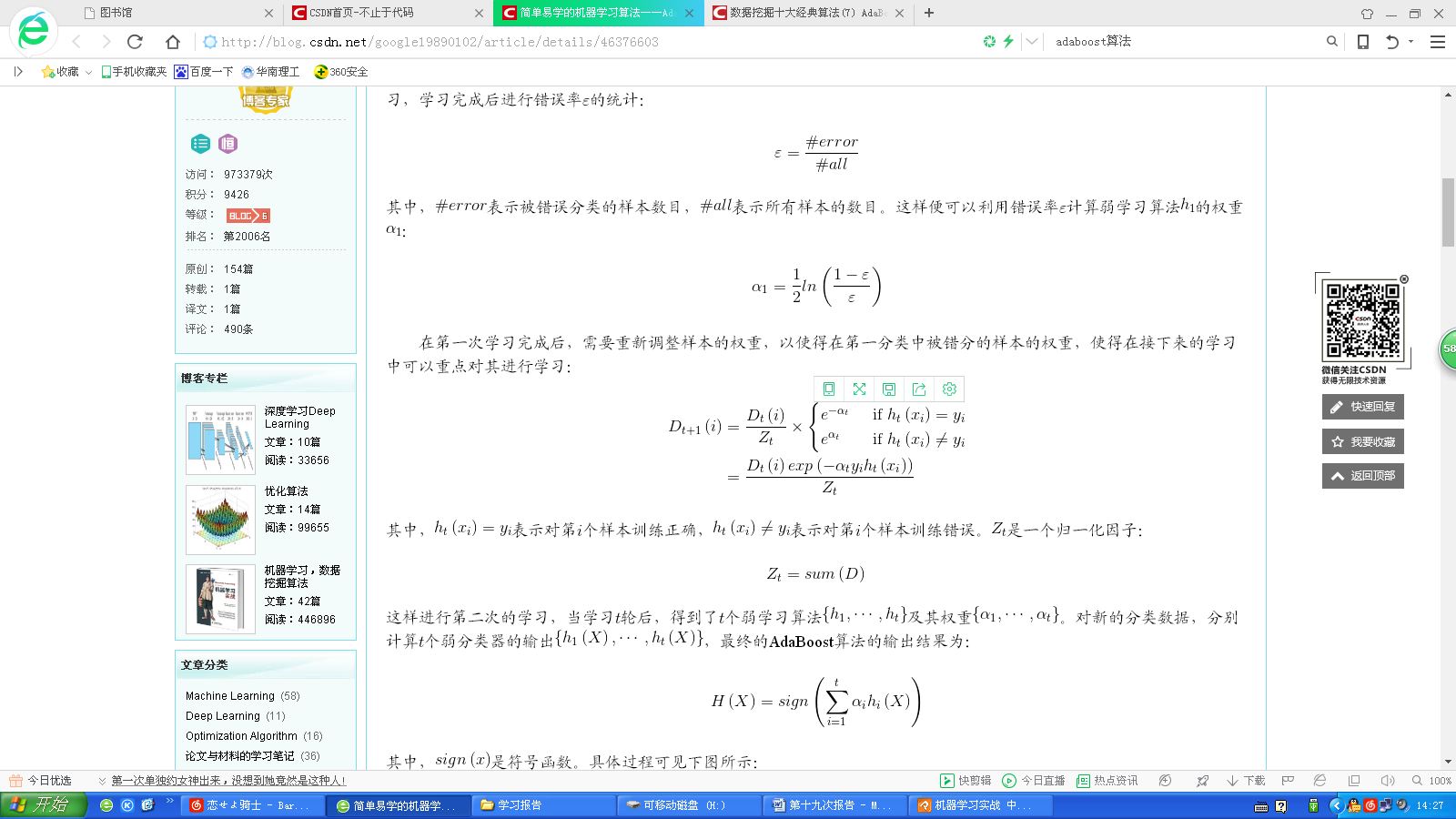
先选定一个弱方法（即分类效果不是非常好的算法），然后利用同一个数据集来训练弱方法来提升弱方法的性能。先设定两个权重，一个是训练集的样本权重，记为向量D，另一个是每一个弱方法的权重，记为向量α。

设数据集为{（x1,y1），（x2，y2）……（xn，yn）}，初始化样本的权重为1/n。

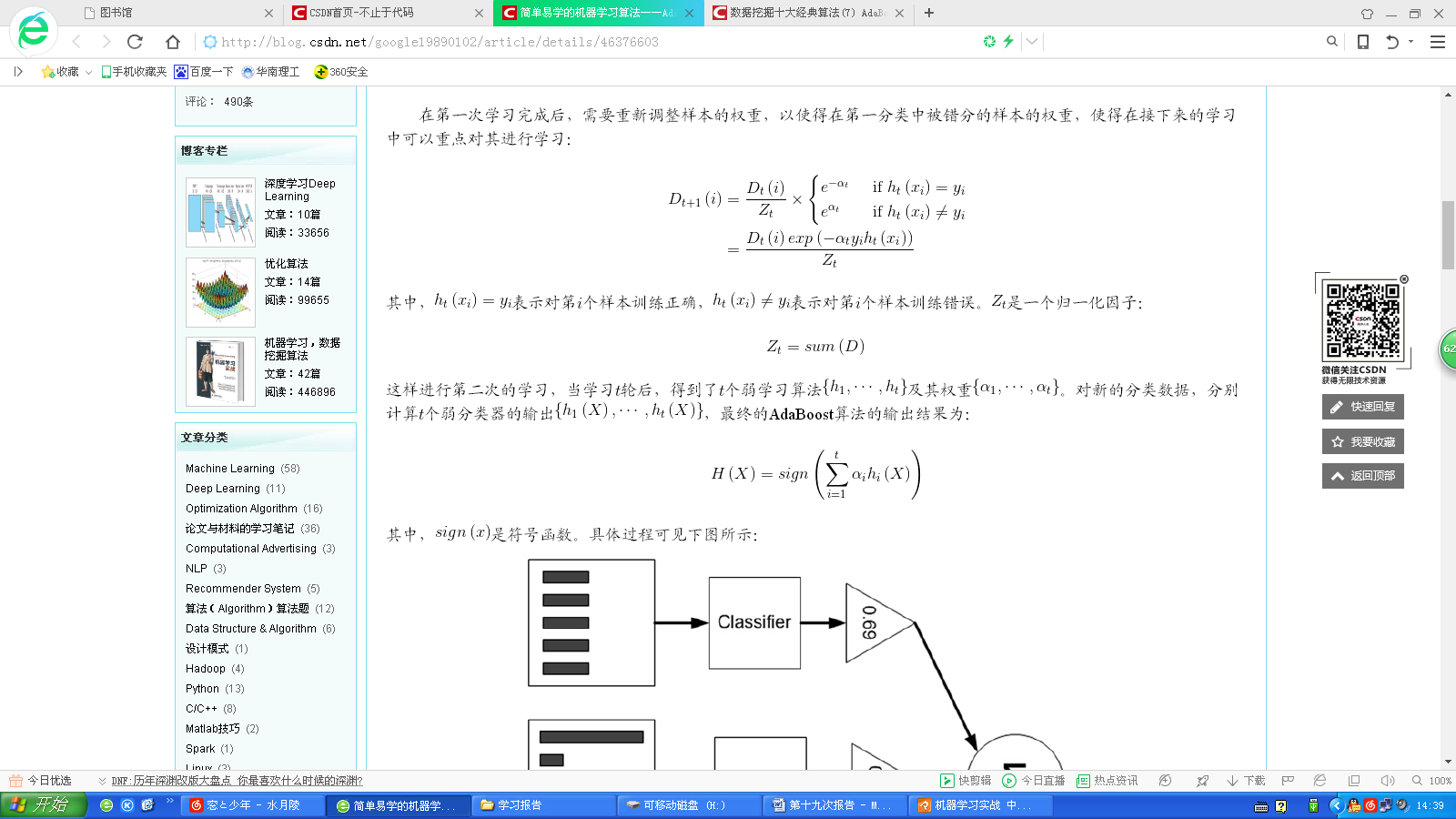
利用第一个弱方法h1训练（分类目标是分类错误率与样本权重有关），训练结束后对错误率ε进行统计：

ε=num\_erro/num\_all，表示分类错误的样本数目除以总的数目

那么对弱方法h1的权重的计算如下（错误越多，方法权重越小）：

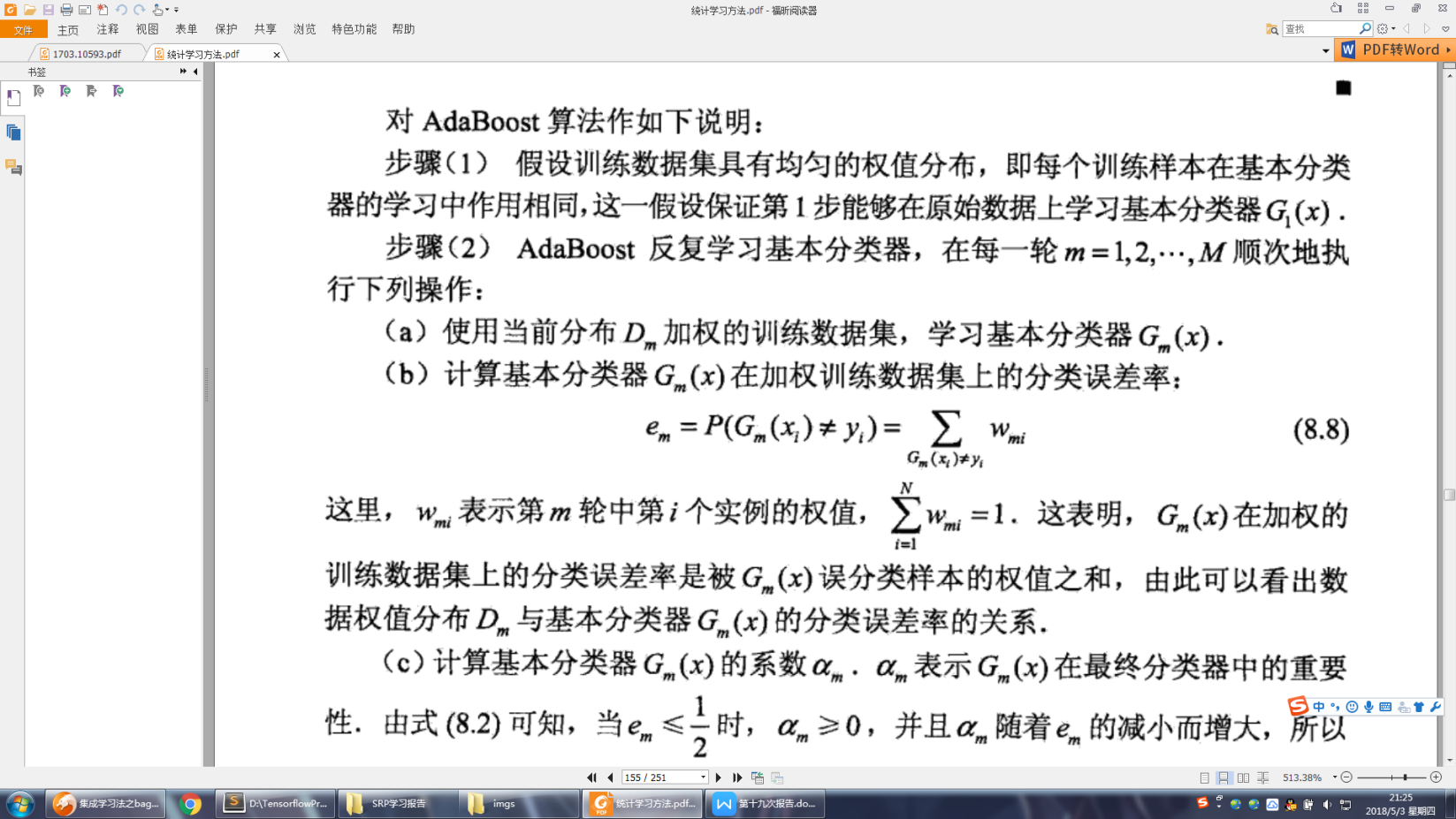


 在第一次学习完成后，需要重新调整样本的权重，以使得在第一次分类中被错分的样本的权重最大，使得在接下来的学习中可以重点对其进行学习：

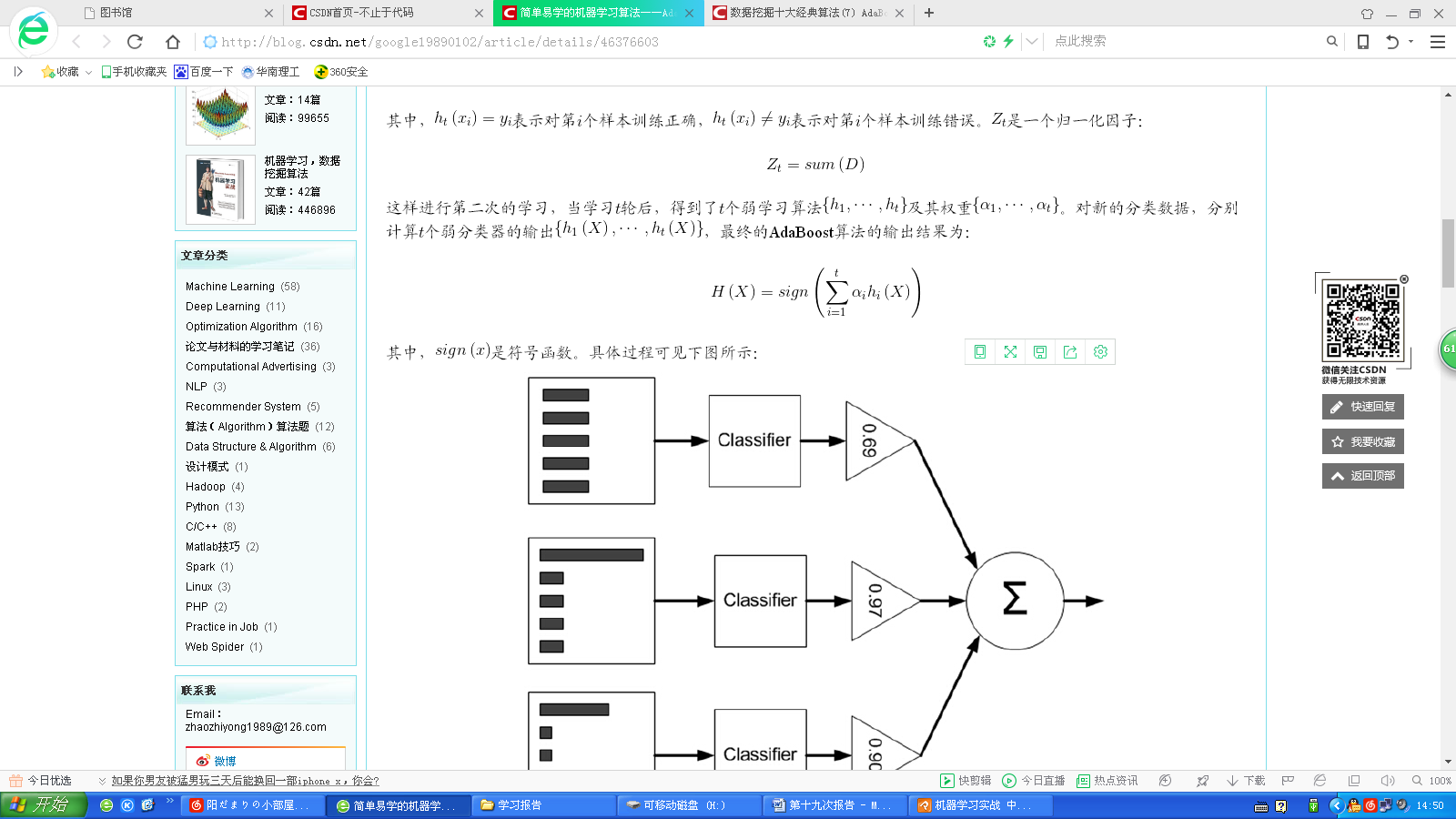


其中，Zt是归一化因子，使得总和为1，Zt=sum（D）

接下来的分类错误率与样本权重有关，即



重复进行这个过程直至进行了t轮弱方法的学习，那么就得到了t个弱学习算法{h1，h2…..ht}及其权重{α1，α2,……αt}。对新的分类数据，分别计算个弱分类器的输出，最终的AdaBoost算法的输出结果为：

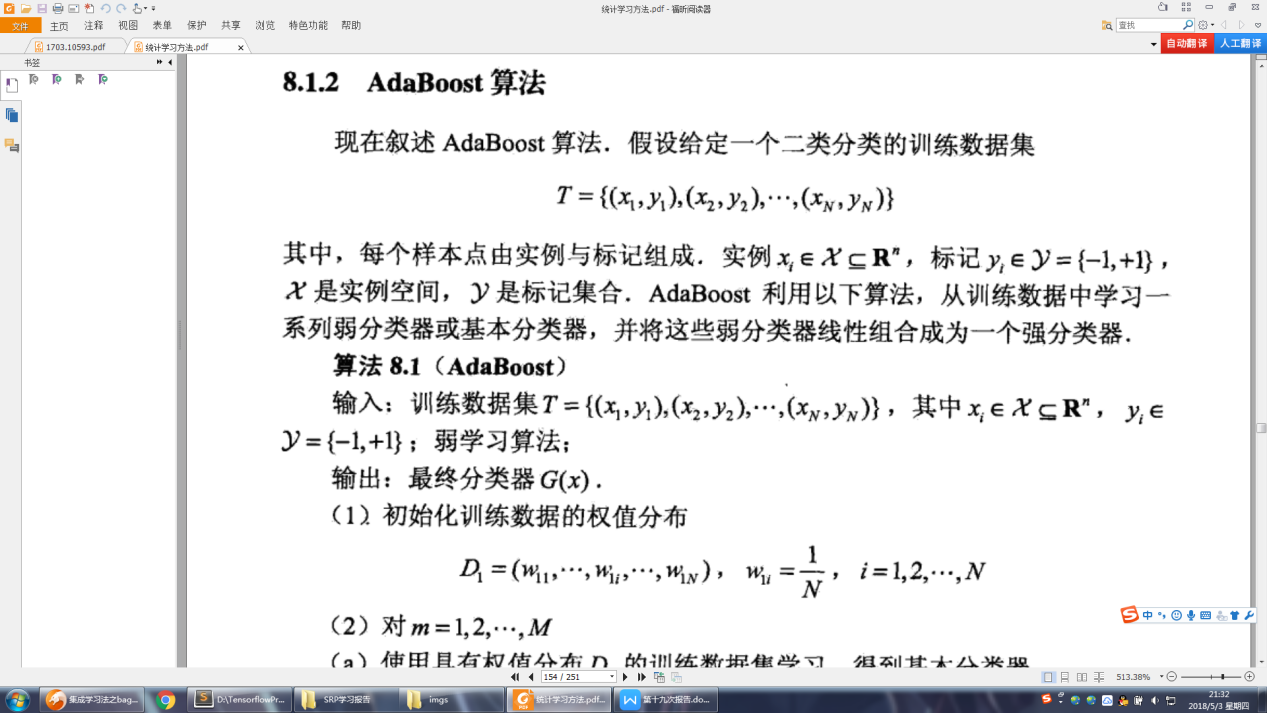


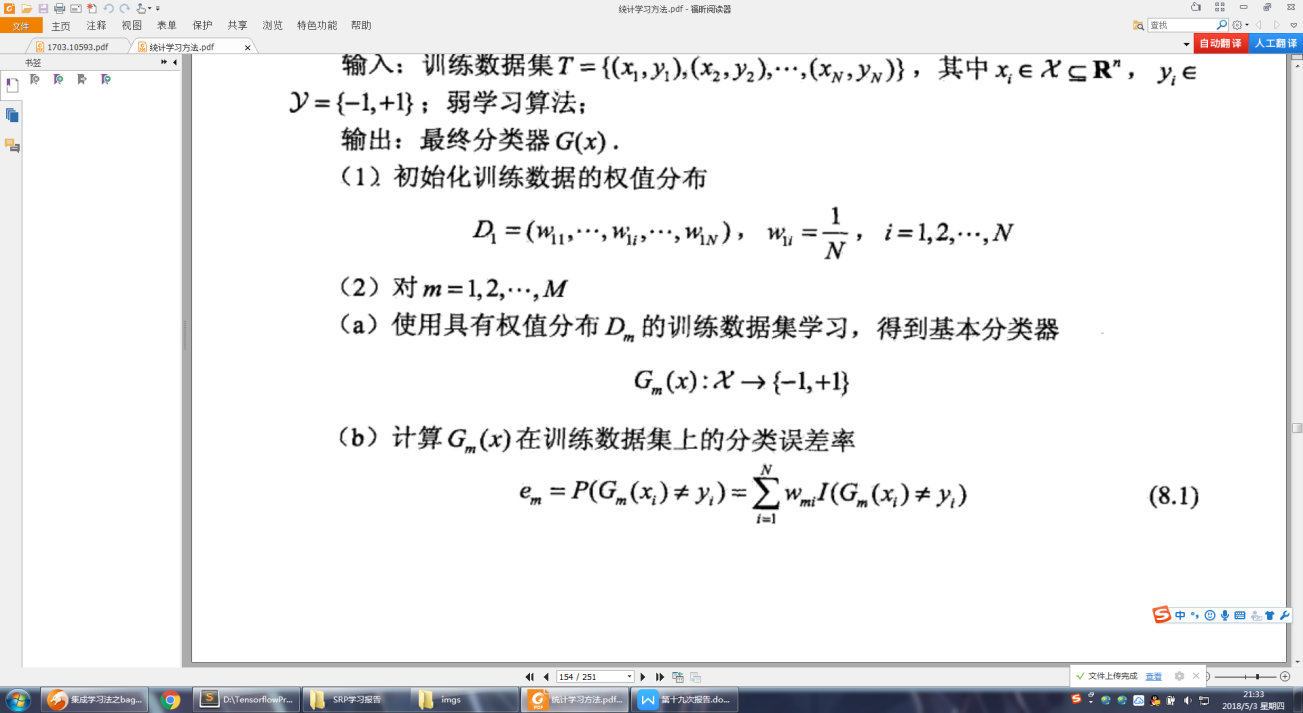
优点：泛化错误率低，易编码，可以应用到大部分分类器中，无需参数调整

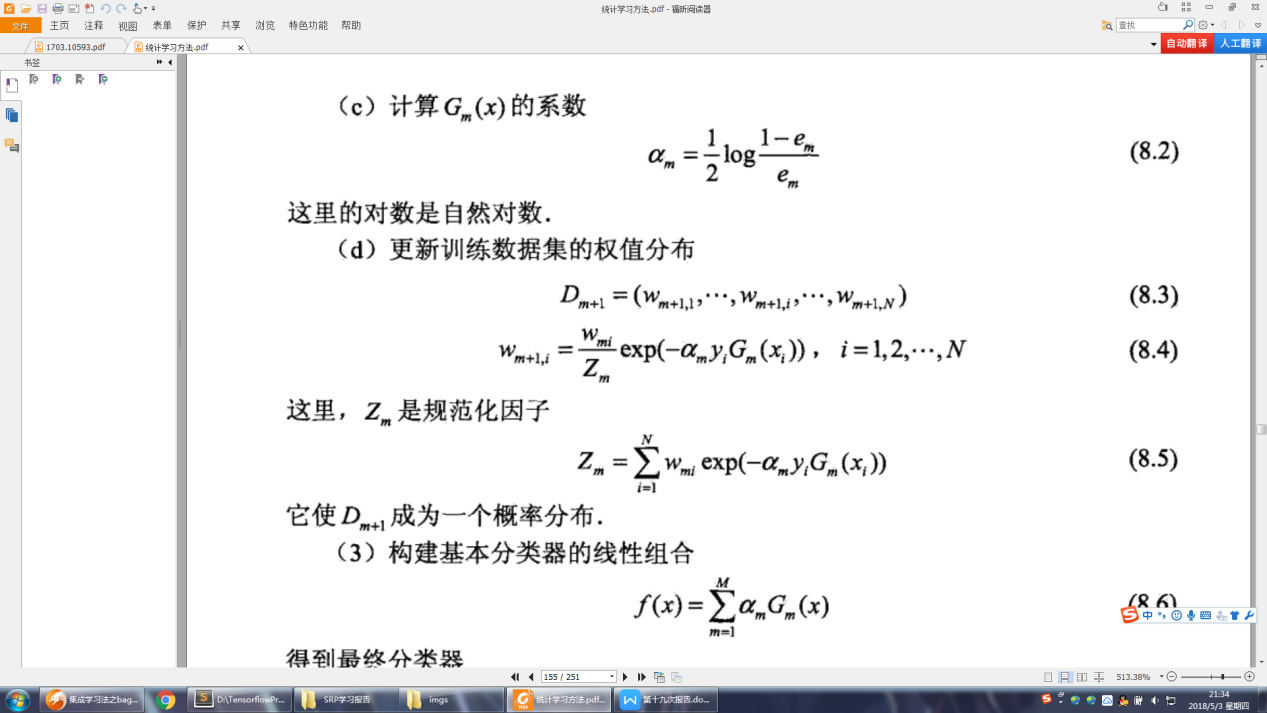
缺点：对离群点敏感

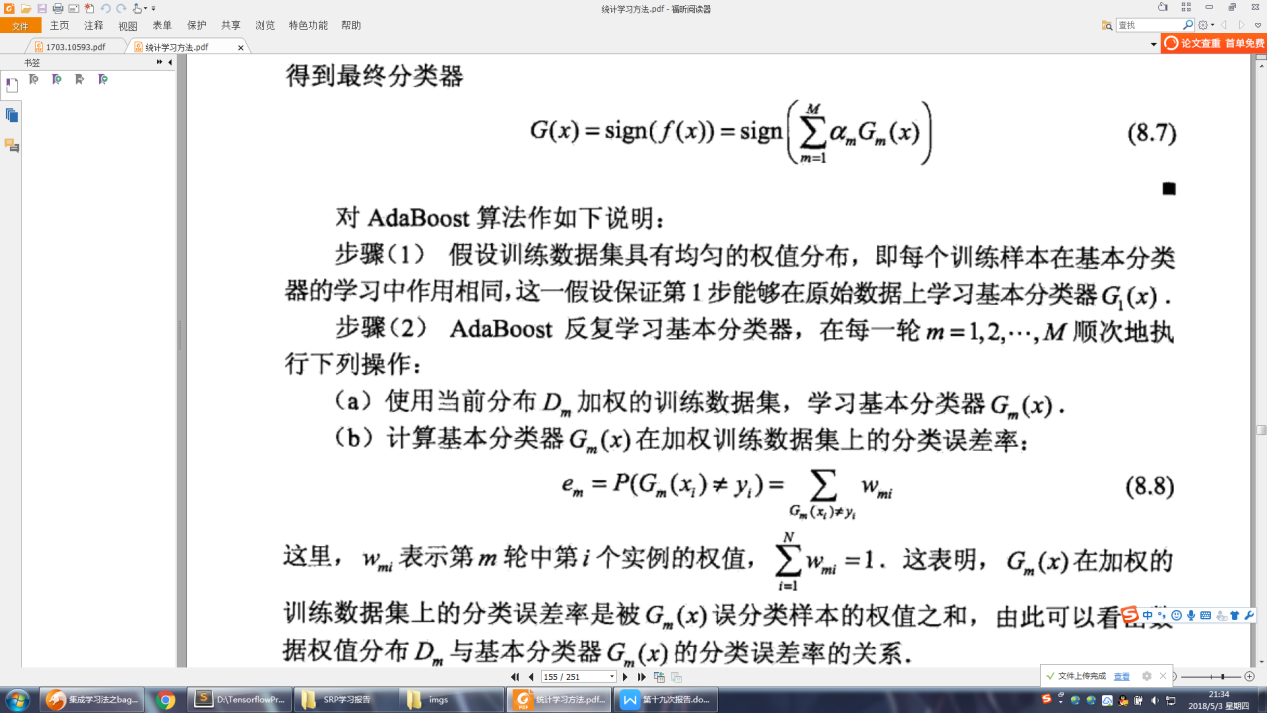
即达到了三个臭皮匠胜过诸葛亮的目标！

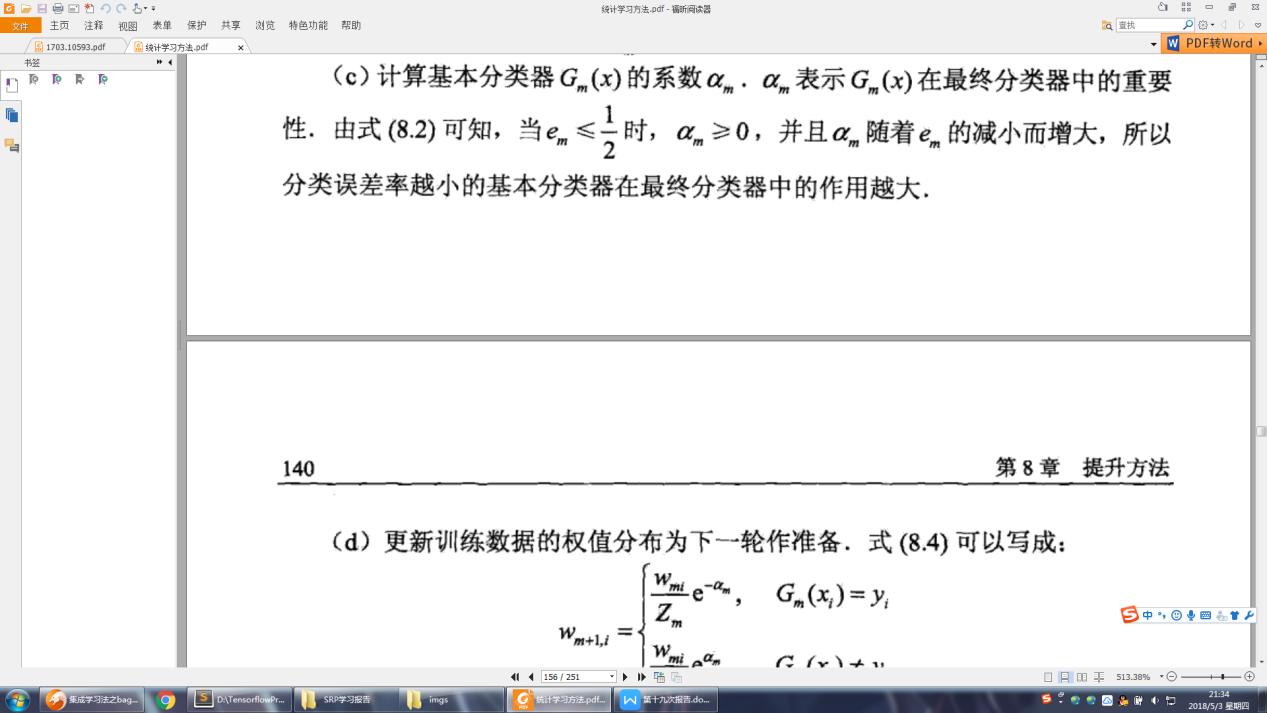
附李航的《统计学习方法》讲解

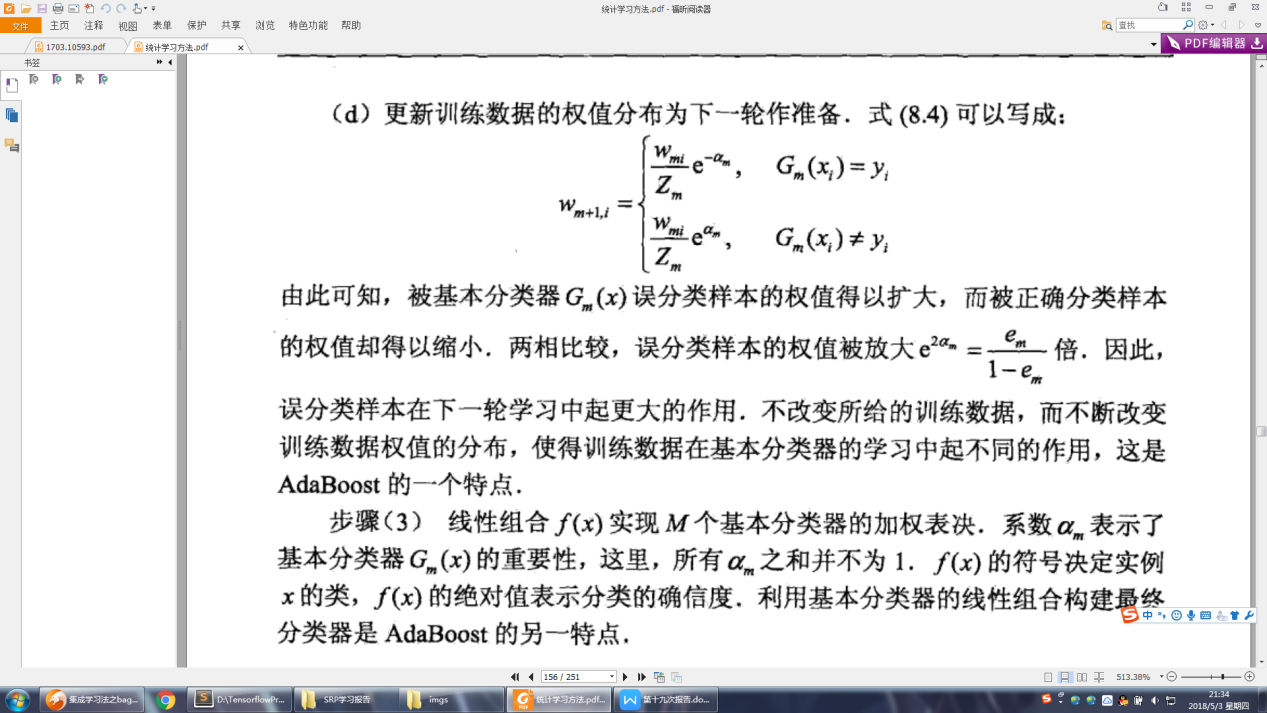












参考资料：

[1].CSDN博主**[zhiyong\_will](http://my.csdn.net/google19890102" \t "_blank)** <http://blog.csdn.net/google19890102/article/details/46376603>

[2].机器学习实战 第七章