实验一 ：集成学习资料查阅

概念

集成学习其实就是分类器集成，通过构建并结合多个学习器来完成学习任务。一般结构是：先产生一组“个体学习器”，再用某种策略将它们结合起来。结合策略主要有平均法、投票法和学习法。

集成学习的一般结构：先产生一组“个体学习器”，再用某种策略将它们结合起来。个体学习器通常由一个现有的学习算法从训练数据中产生，例如C4.5决策算法、BP神经网络算法等，此时集成中只包含同种类型的个体学习器，例如猫狗二分类实验使用的resnet34网络，这样的集成是“同质”的。同质集成中的个体学习器亦称为“基学习器”。相应的学习算法称为“基学习算法”。集成也可包含不同类型的个体学习器，例如，同时包含决策树和神经网络，这样的集成称为“异质”的。异质集成中的个体学习器由不同的学习算法生成，这时就不再有基学习算法，常称为“组件学习器”或直接称为个体学习器。

集成学习通过将多个学习器进行结合，常可获得比单一学习器更加显著的泛化性能。这对“弱学习器”尤为明显。因此集成学习的理论研究都是针对弱学习器进行的，而基学习器有时也被直接称为弱学习器。

但需注意的是，虽然从理论上说使用弱学习器集成足以获得很好的性能，但在实践中出于种种考虑，例如希望使用较少的个体学习器，或是重用一些常见学习器的一些经验等，人们往往会使用比较强的学习器。

集成学习原理

以本实验为例，基学习器为resnet34，每一个基学习器的准确率为54%，集成学习的结果通过设定投票法产生（即是阈值为60%），即“少数服从多数”，最后集成学习的结果为74%的精度。如图当基学习器的精度只有33.3%时，集成学习的效果变的更糟。所以集成学习的基学习器必须要有一定的精度，并且要有“多样性”，即学习器之间有差异。

分析原理：

考虑二分类问题 和真实函数 f ，假定基分类器的错误率为 ，即对每个基分类器 有：



假设集成通过简单投票法结合T个基分类器，若有超过半数的基分类器正确，则集成分类就正确：



假设分类器的错误率相互独立，则由Hoeffding 不等式可知，集成的错误率为：

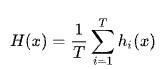
上式显示出，随着集成中个体分类器数目的增大，集成的错误率将指数级下降，最终趋向于零。

结合策略：

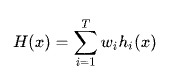
平均法

对于数值预测问题，通常使用的结合策略是平均法，也就是说，对于数值类的回归预测问题，通常使用的结合策略是平均法，也就是说，对于若干和弱学习器的输出进行平均得到最终的预测输出。

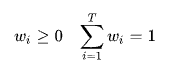
最简单的平均是算术平均，也就是说最终预测是



如果每个个体学习器有一个权重,则最终预测是：



其中，其中是个体学习器的权重，通常有



投票法

对于分类问题的预测，我们通常使用的是投票法。假设我们的预测类别是对于任意一个预测样本x，我们的T个弱学习器的预测结果分别是

最简单的投票法是相对多数投票法，也就是我们常说的[少数服从多数](https://baike.baidu.com/item/%E5%B0%91%E6%95%B0%E6%9C%8D%E4%BB%8E%E5%A4%9A%E6%95%B0/10350157" \t "_blank)，也就是T个弱学习器的对样本x的预测结果中，数量最多的类别为最终的分类类别。如果不止一个类别获得最高票，则随机选择一个做最终类别。

稍微复杂的投票法是绝对多数投票法，也就是我们常说的要票过半数。在相对多数投票法的基础上，不光要求获得最高票，还要求票过半数。否则会拒绝预测。

更加复杂的是加权投票法，和[加权平均法](https://baike.baidu.com/item/%E5%8A%A0%E6%9D%83%E5%B9%B3%E5%9D%87%E6%B3%95/10357914" \t "_blank)一样，每个弱学习器的分类票数要乘以一个权重，最终将各个类别的加权票数求和，最大的值对应的类别为最终类别。

学习法

对弱学习器的结果做平均或者投票，相对比较简单，但是可能学习误差较大，于是就有了学习法这种方法。对于学习法，代表方法是stacking，当使用stacking的结合策略时， 我们不是对弱学习器的结果做简单的逻辑处理，而是再加上一层学习器，也就是说，我们将训练集弱学习器的学习结果作为输入，将训练集的输出作为输出，重新训练一个学习器来得到最终结果。

在这种情况下，我们将弱学习器称为初级学习器，将用于结合的学习器称为次级学习器。对于测试集，我们首先用初级学习器预测一次，得到次级学习器的输入样本，再用次级学习器预测一次，得到最终的预测结果。

相关文献

* [郭宏，丁晓青.    汉字识别多分类器集成的新方法．](https://xueshu.baidu.com/usercenter/paper/show?paperid=41e47a18ce8cd95618285c2563ef4f5e&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&site=baike)《 清华大学学报(自然科学版) 》

zh

主要内容：

* [林晓帆，丁晓青.    独立分类器集成理论及其在字符识别中的应用．](https://xueshu.baidu.com/usercenter/paper/show?paperid=22e7e69a2554ced1bae827641d5bde40&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&site=baike)《 CNKI 》

主要内容：

实验二：是非

实验目的：为了构建出纯粹的是非网络，对训练集中的非猫但是特征类似猫的图片抑制其反向传播，使得网络只能识别猫的特征和非猫但特征和猫相似的物体（你仅限于狗）。

实验方法：一般认为网络的方向传播是神经网络的自学习过程，为了达到实验目的中学习猫的特征和类似猫的特征，通过softmax函数对实验目的量化，即，设猫的label=1 ， 其他的label = 0;

实验过程：

实验结果：