

一、概述

进行一个分类问题的实例分析，根据汽车的特征评估质量。在这里我们使用的是随机森林分类器。

二、内容

1 随机森林的生成

前面提到，随机森林中有许多的分类树。我们要将一个输入样本进行分类，我们需要将输入样本输入到每棵树中进行分类。打个形象的比喻：森林中召开会议，讨论某个动物到底是老鼠还是松鼠，每棵树都要独立地发表自己对这个问题的看法，也就是每棵树都要投票。该动物到底是老鼠还是松鼠，要依据投票情况来确定，获得票数最多的类别就是森林的分类结果。森林中的每棵树都是独立的，99.9%不相关的树做出的预测结果涵盖所有的情况，这些预测结果将会彼此抵消。少数优秀的树的预测结果将会超脱于芸芸“噪音”，做出一个好的预测。将若干个弱分类器的分类结果进行投票选择，从而组成一个强分类器，这就是随机森林bagging的思想

2、小知识点

将数据转换编码：需要仔细读代码，了解里面参数的含义。在python3转换中需要二维数组，及时数据只有一维也要转换为二维，不然会报错。这点在后面对单一数据点分类有一个完整的示例。

超参数：超参数是在开始学习前就设置好的参数，不能通过训练得到。比如本例子中的分类器个数（`n_estimators`）、深度（`max_depth`）。在最后我们进行超参调节，不同的超参训练出结果的对比。

参考博客：

朴素贝叶斯：<https://www.cnblogs.com/liuyihai/p/8309019.html>