# ML实验报告：随机森林

霍超凡

2019 年 12 月 21 日

## 1 实验内容

利用随机森林进行鸢尾花数据集的分类。

## 2 实验要求

对鸢尾花数据集进行分类，准确率越高越好。并将算法思想以及实验结果记录到实验报告里。向助教老师演示实验，并提交实验代码和实验报告。

## 3 实验环境

Jupyter Notebook + Python3

## 4 实验原理

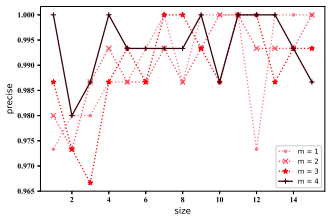
随机森林使用集成学习中的Bagging策略，在从训练集中随机采样得到的不同的子数据集中分别构建不同的决策树，最终对未知样本的预测由多棵树投票得到。随机森林加入随机划分和随机采样，减少了过拟合的风险，性能得到极大提升。

### 树的构建过程

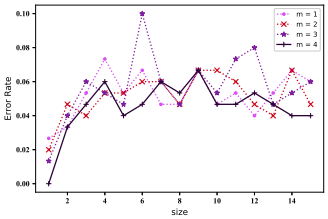
1. 从原始数据集中有放回抽样得到不同训练集，这些训练集将被用来构建决策树。
2. 在构建树的过程中，随机从属性集中选取一个子集，划分节点的最优属性将从这个子集中产生。
3. 不采用剪枝策略，随性生长。

### 包外估计

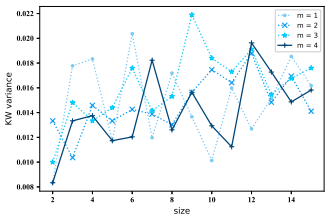
在构建树的过程中采取有放回的采样方法，约有三分之一的数据没有被用来训练决策树，或者反过来，对于原始数据的每一个样本，约有占比三分之一的树在训练过程中没有使用这个样本，这些没有被用于训练的样本将作为验证集评估决策树的性能。其具体计算过程中如下：从原始数据集的第一个样本开始遍历，使用在训练过程中没有用到这个样本的决策树预测这个样本的类别，这个样本的最终类别由不同决策树投票决定，计算被错误分类的样本树占总样本数的比例，这个比例就是使用包外估计得到的错误率。




(a) 精度




(b) 错误率




(c) 相关度

图 1: 设置不同随机森林中树的数目和随机扰动中属性子集的大小m，构建不同随机森林，并测试随机森林的相关数据。图1(a)是随机森林在原始数据的正确率，实线的m等于总属性个数，意味着不添加任何属性扰动，不添加属性扰动构建的随机森林的性能似乎比添加扰动要好；图1(b)是使用包外估计得到的错误率，添加属性扰动似乎具有较高的错误率；图1(c) 是随机森林的KW方差，添加扰动的随机森林中树与树之间的差异性不如没有添加属性扰动。以上结果均存在随机性，和我们所预想的不同。

### 树之间的相关性衡量

有两个因素会影响到随机森林的性能，

- 随机森林中树与树之间的相关性，树之间的相关性越高，随机森林的错误率越高。
- 随机森林中单棵树的性能，随机森林中每棵树的错误率越小，整体错误率越小。

为衡量随机森林中树与树之间的相关性，使用KW方差

$$KW = \frac{1}{NL^2}\sum_{i=1}^{N} l_i(L - l_i)$$

其中$L$是随机森林中树的数量，$N$是原始训练集的样本总数，$l_i$是将第$i$个样本分错的树的个数。

## 5 实验结果

以下实验结论均是针对鸢尾花数据集，不具有一般性，仅供参考。

（1）增加随机森林中树的数目的确会提高随机森林的性能，在鸢尾花数据集上树的数目取8 左右，过多的数目不会再提高性能，因为已经达到100%的准确度。

（2）加入属性扰动后，相比不加入扰动提高了树之间的差异度，但是使用包外估计得到的错误率略微提高，在整个原始数据集的精度越稍稍降低。选取子集中属性的个数似乎对结果影响不大。

（3）在构建树的过程中，随机选取不同的划分准测，得到的随机森林中树与树之间的差异性不如添加属性扰动。

表 1: 添加不同扰动所构建随机森林性能对比

| | 单棵树 | 数据扰动 | 数据+属性扰动 | 数据+划分准测扰动 | 数据+属性+划分准则扰动 |
|---|---|---|---|---|---|
| 正确率 | 0.973 | **1.000** | 0.986 | **1.000** | 0.993 |
| 错误率 | **0.026** | 0.040 | 0.046 | 0.080 | 0.046 |
| 相关度 | **1.000** | 0.013 | 0.011 | 0.021 | 0.012 |