机器学习数据分析和建模中级教程系列讲座-第4讲

**决策树(Decision Tree)(下)**

主讲人：元俊

**预备知识**

数据结构 基础

**上一讲回顾**

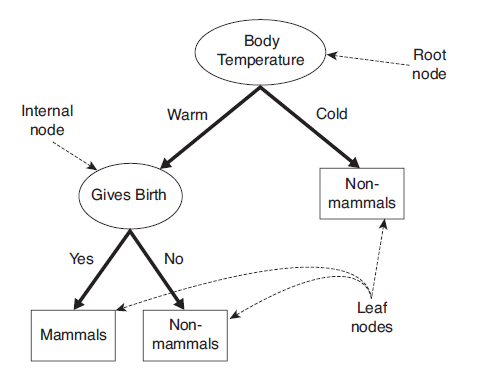
混沌的量化—>信息熵的定义

**决策树的定义回顾**

**根节点**（root node），表示第一个特征属性，只有出边没有入边；

**内部节点**（internal node），表示特征属性，有一条入边至少两条出边

**叶子节点**（leaf node），表示类别，只有一条入边没有出边。



**ID3/C4.5/CART**

决策树（decision tree）算法基于特征属性进行分类，其主要的优点：模型具有可读性，计算量小，分类速度快。

决策树算法包括了由Quinlan提出的ID3与C4.5，Leo Breiman等提出的CART。

其中，C4.5是基于ID3的，对分裂属性的目标函数做出了改进。

**How To?**

* 如何选择较优的特征属性进行分裂？每一次特征属性的分裂，相当于对训练数据集进行再划分，对应于一次决策树的生长。ID3算法定义了信息熵增益来进行特征选择。
* 什么时候应该停止分裂？有两种自然情况应该停止分裂，一是该节点对应的所有样本记录均属于同一类别，二是该节点对应的所有样本的特征属性值均相等。但除此之外，是不是还应该其他情况停止分裂呢？

**ID3的问题？**

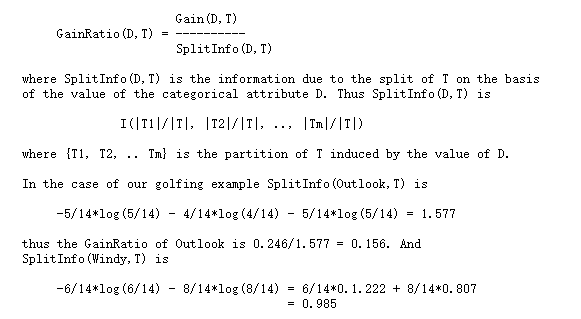
假如有身份证这一列？

ID3倾向于会把每一条数据都分成一个单独的列！

**C4.5的改进**

**增加了对连续量的支持**

信息增益率，加大对**ID3的“***多子多福”***不良副作用的惩罚**

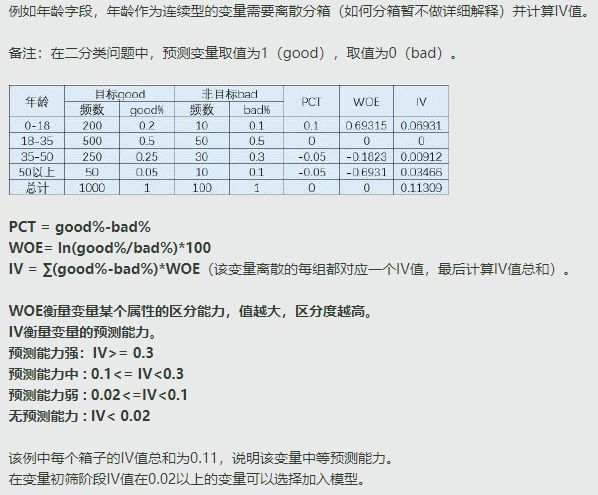


**进阶**

**如何筛选变量？WOE/IV**

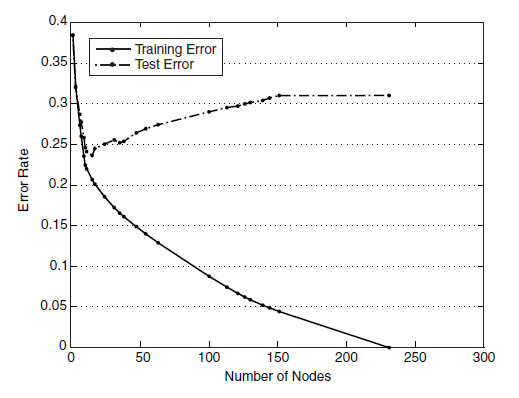
IV的全称是Information Value，中文意思是信息价值，或者信息量。

我们在用逻辑回归、决策树等模型方法构建分类模型时，经常需要对自变量进行筛选。比如我们有200个候选自变量，通常情况下，不会直接把200个变量直接放到模型中去进行拟合训练，而是会用一些方法，从这200个自变量中挑选一些出来，放进模型，形成入模变量列表。那么我们怎么去挑选入模变量呢？



**进阶2**

**过拟合问题**



* 剪枝（Pruning）

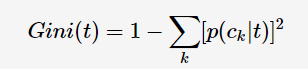
预剪枝（划分节点的时候有没有必要继续下去，合不合算？）

后剪枝（从底层内部节点开始逐步向上递归评估，拿测试集评估）

**CART（Classification And Regression Tree）**

分类与回归树（Classification and Regression Trees, CART）是由四人帮Leo Breiman, Jerome Friedman, Richard Olshen与Charles Stone于1984年提出，既可用于分类也可用于回归。本文将主要介绍用于分类的CART。CART被称为数据挖掘领域内里程碑式的算法。

CART采用Gini指数来度量分裂时的不纯度，之所以采用Gini指数，是因为较于熵而言其计算速度更快一些。



CART的剪枝方法



考虑C(T)表示决策树的训练误差，α为调节参数，|T|为模型的复杂度

小知识：

1945年，经过长达16小时的入学考试后，Breiman 成功进入加州理工学院，主修物理，并因成绩优异而获得奖学金。然而，到了大学高年级，由于在校园里到处充斥着理工知识而渐生厌倦，感觉 Caltech 就像一个“科学修道院”（scientific monastery）。终于，他在大学最后一年的四门物理课中都得了“D”.

Leo Breiman是一个兴趣广泛的人，他不仅是专业的统计学家和概率学家，还在其他方面也取得了很多成就。他在Catskills当过服务员，在Merchant Marine当过洗碗工，同时他是一名探寻过热带雨林核心地带的背包客，是一群来自墨西哥农村孩子的慈爱父亲，是Santa Monica学校董事会的主席，是他美丽小屋的建筑师，还是一个技艺高超的雕刻家。

1993年，Breiman 从伯克利退休。但或许谁也没有想到，Breiman 学术生涯的高潮才刚刚开始。在研究中，Breiman 发现，使用线性模型选择变量的子集（subset selection in linear regression）是一个很不稳定的过程；只要数据稍微变化，所选变量就可能很不相同。对于决策树也如此，只要数据稍有变化，所生成的决策树可能就大相径庭。

于是，Breiman 想，能否将这些不同的模型结果进行平均，以得到更好的预测效果。但如何扰动数据呢？Breiman想到了有放回的自助抽样（bootstrap）。首先，从原始样本得到大量的自助样本（bootstrap sample），分别用 CART 算法来估计决策树，然后再将这些决策树进行聚合平均（比如，平均1000棵决策树）。这就是Breiman (1996)提出的 “bagging” 算法（bootstrap aggregating），该文目前在 Google Scholar的引用量已超过 2 万。

2001年，Breiman 进一步提出了 “随机森林”（Random Forest）的算法。与Bagging 不同，在决策树的每个节点，仅随机选取部分变量进行分裂，以降低不同决策树之间的相关性。结果发现，Random Forest的预测效果比 Bagging 又有了进一步的提高，这使得随机森林成为机器学习，特别是数据科学的主要算法之一。

**CART的衍生算法**

Bagging/Random Forest

**扩展阅读**

天普大学ID3-C4.5的比较<https://cis.temple.edu/~giorgio/cis587/readings/id3-c45.html>

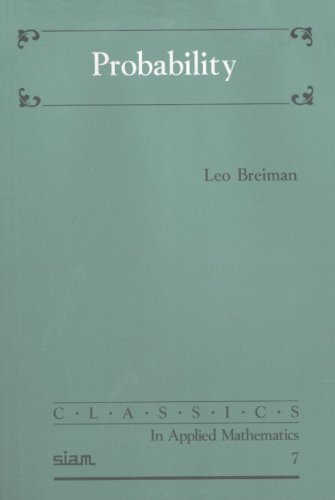
Berkeley统计系Leo Breiman发表的论文集合<https://www.stat.berkeley.edu/~breiman/papers.html>

Leo Breiman的访谈（中文版）<https://www.sohu.com/a/120433200_500658> 里面提到了Leo Breiman的自传和传奇经历

WOE和IV的定义

<http://ucanalytics.com/blogs/information-value-and-weight-of-evidencebanking-case/>

Leo Breiman的概率论



**代码附录：**

Git地址：<https://github.com/Yao-lab001/machine_learning_lecture.git>

运行环境：Python 3.7.6 or anaconda 2020.02

预装环境：matplotlib 3.1.3

WOE和IV计算的demo excel参见课件附件

**下一讲预告**

朴素贝叶斯

