机器学习及其MATLAB实现—从基础到实践 第8课

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

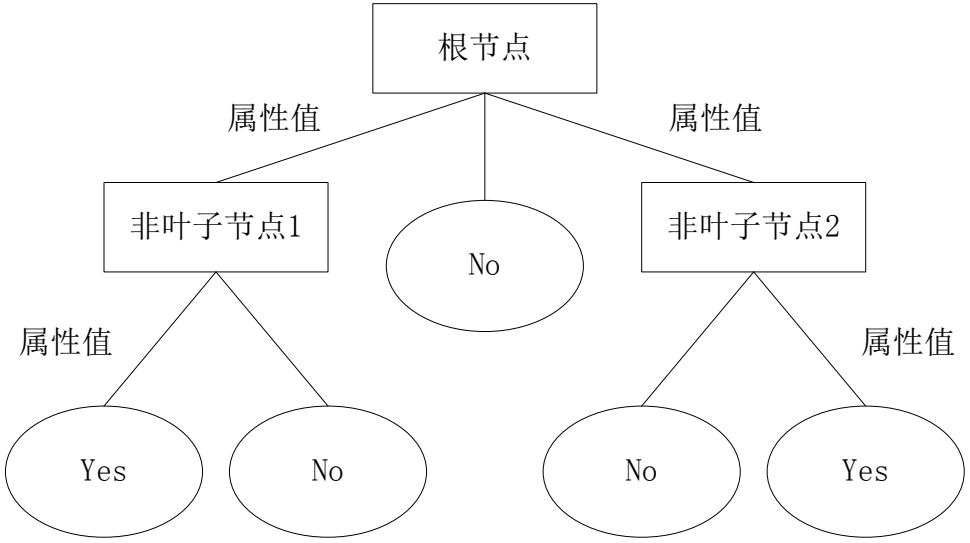
http://edu.dataguru.cn

课程目录

- 第一课　MATLAB入门基础
- 第二课　MATLAB进阶与提高
- 第三课　BP神经网络
- 第四课　RBF、GRNN和PNN神经网络
- 第五课　竞争神经网络与SOM神经网络
- 第六课　支持向量机（Support Vector Machine, SVM）
- 第七课　极限学习机（Extreme Learning Machine, ELM）
- **第八课　决策树与随机森林**
- 第九课　遗传算法（Genetic Algorithm, GA）
- 第十课　粒子群优化（Particle Swarm Optimization, PSO）算法
- 第十一课　蚁群算法（Ant Colony Algorithm, ACA）
- 第十二课　模拟退火算法（Simulated Annealing, SA）
- 第十三课　降维与特征选择

决策树原理概述

- 决策树通过把样本实例从根节点排列到某个叶子节点来对其进行分类。树上的**每个非叶子节点代表对一个属性取值的测试**，其分支就代表测试的每个结果；而**树上的每个叶子节点均代表一个分类的类别**，树的最高层节点是根节点。
- 简单地说，决策树就是一个类似流程图的树形结构，采用**自顶向下的递归方式**，从树的根节点开始，在它的内部节点上进行属性值的测试比较，然后按照给定实例的属性值确定对应的分支，最后在决策树的叶子节点得到结论。这个过程在**以新的节点为根的子树上重复**。



决策树原理概述

- **ID3算法**

到目前为止，已经有很多种决策树生成算法，但是国际上最有影响力的示例学习算法首推 J. R. Quinlan 的 ID3（Iterative Dichotomic version 3）算法。Quinlan 的首创性工作主要是在决策树的学习算法中引入信息论中互信息的概念，他将其称作信息增益（information gain），以之作为属性选择的标准。

为了精确地定义信息增益，我们先定义信息论中广泛使用的一个度量标准，称为熵（entropy），它刻画了任意样例集的纯度（purity）。

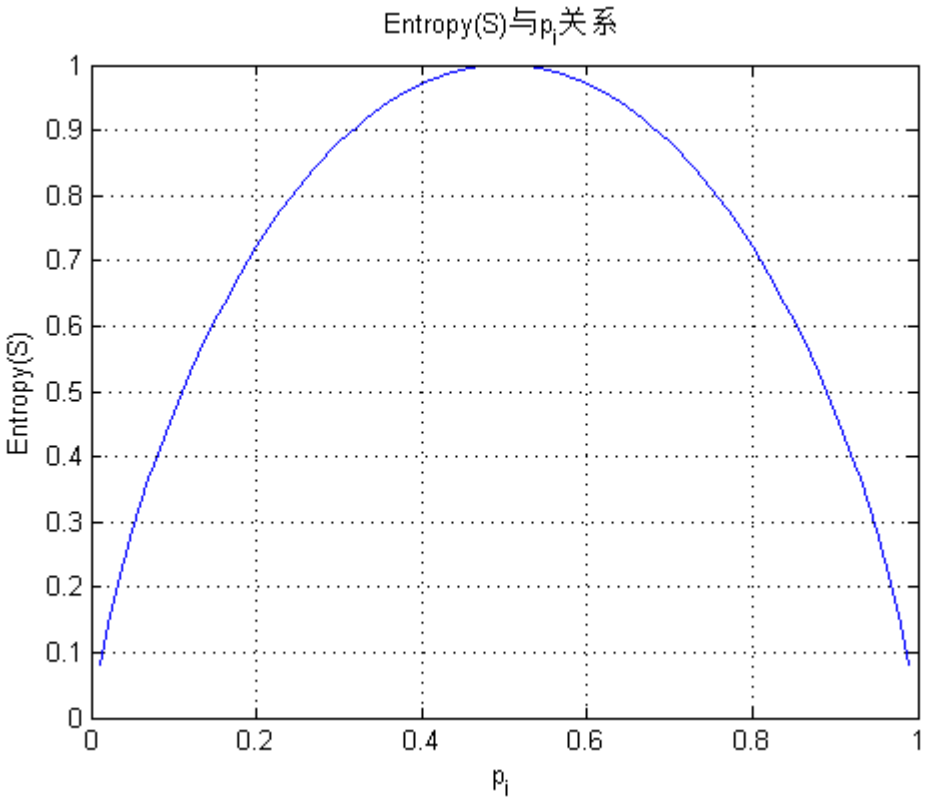
如果目标属性具有 c 个不同的值，那么集合 S 相对于 c 个状态的分类的熵定义为：

$$Entropy(S)=\sum_{i=1}^{c}-p_i\log_2 p_i \tag{28-1}$$

其中，p_i为子集合中第 i 个属性值的样本数所占的比例。

由上式可以得到：若集合 S 中的所有样本均属于同一类，则 $Entropy(S)=0$；若两个类别的样本数不相等，则 $Entropy(S)\in(0,1)$。

特殊地，若集合 S 为布尔型集合，即集合 S 中的所有样本属于两个不同的类别，则有以下关系成立：若两个类别的样本数相等，则 $Entropy(S)=1$。



决策树原理概述

- **ID3算法**

已经有了熵作为衡量训练样例集合纯度的标准，信息增益 $Gain(S,A)$ 的定义为：

$$Gain(S,A)=Entropy(S)-\sum_{v\in V(A)}\frac{|S_v|}{|S|}Entropy(S_v)$$

其中，$V(A)$ 是属性 A 的值域；S_v 是集合 S 中在属性 A 上值等于 v 的子集。

引入信息增益的概念后，下面将详细介绍 ID3 算法的基本流程，不妨设 Examples 为训练样本集合，Attribute list 为候选属性集合。

（1） 创建决策树的根节点 N；

（2） 若所有样本均属于同一类别 C，则返回 N 作为一个叶子节点，并标志为 C 类别；

（3） 若 Attribute list 为空，则返回 N 作为一个叶子节点，并标志为该节点所含样本中类别最多的类别；

（4） 计算 Attribute list 中各个候选属性的信息增益，选择最大的信息增益对应的属性 Attribute*，标记为根节点 N；

（5） 根据属性 Attribute*值域中的每个值 V_i，从根节点 N 产生相应的一个分支，并记 S_i 为 Examples 集合中满足 Attribute* = V_i 条件的样本子集合；

（6） 若 S_i 为空，则将相应的叶子节点标志为 Examples 样本集合中类别最多的类别；否则，将属性 Attribute*从 Attribute list 中删除，返回（1），递归创建子树。

决策树原理概述

- **C4.5算法**

针对 ID3 算法存在的一些缺点，许多学者包括 Quinlan 都做了大量的研究。C4.5 算法便是 ID3 算法的改进算法，其相比于 ID3 改进的地方主要有：

（1） 用信息增益率（gain ratio）来选择属性。

信息增益率是用信息增益和分裂信息量（split information）共同定义的，如下所示：

$$GainRatio(S,A)=\frac{Gain(S,A)}{SplitInformation(S,A)}$$

其中，分裂信息量的定义为：

$$SplitInformation(S,A)=-\sum_{i=1}^{c}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|}$$

采用信息增益率作为选择分支属性的标准，克服了 ID3 算法中信息增益选择属性时偏向选择取值多的属性的不足。

（2） 树的剪枝。

剪枝方法是用来处理过拟合问题而提出的。剪枝一般分两种方法：先剪枝和后剪枝。

先剪枝方法通过提前停止树的构造，比如决定在某个节点不再分裂，而对树进行剪枝。一旦停止，该节点就变为叶子节点，该叶子节点可以取它所包含的子集中类别最多的类作为节点的类别。

后剪枝的基本思路是对完全成长的树进行剪枝，通过删除节点的分支，并用叶子节点进行替换，叶子节点一般用子集中最频繁的类别进行标记。

C4.5 算法采用悲观剪枝法（Pessimistic Pruning）是 Quinlan 在 1987 年提出的，属于后剪枝方法的一种。它使用训练集生成决策树，并用训练集进行剪枝，不需要独立的剪枝集。悲观剪枝法的基本思路是：若使用叶子节点代替原来的子树后，误差率能够下降，则就用该叶子节点代替原来的子树。

决策树原理概述

优点

- 决策树易于理解和实现。 人们在通过解释后都有能力去理解决策树所表达的意义。
- 对于决策树，数据的准备往往是简单或者是不必要的。其他的技术往往要求先把数据归一化，比如去掉多余的或者空白的属性。
- 能够同时处理数据型和常规型属性。 其他的技术往往要求数据属性的单一。
- 是一个白盒模型。如果给定一个观察的模型，那么根据所产生的决策树很容易推出相应的逻辑表达式。

缺点

- 对于各类别样本数量不一致的数据，在决策树当中信息增益的结果偏向于那些具有更多数值的特征。
- 决策树内部节点的判别具有明确性，这种明确性可能会带来误导。

随机森林概述

- **Bootstrap抽样**

设集合 S 中含有 n 个不同的样本 $\{x_1, x_2, ..., x_n\}$，若每次有放回地从集合 S 中抽取一个样本，一共抽取 n 次，形成新的集合 S^*，则集合 S^* 中不包含某个样本 $x_i (i = 1, 2, ..., n)$ 的概率为

$$p = \left(1 - \frac{1}{n}\right)^n$$

当 $n \to \infty$ 时，有

$$\lim_{n\to\infty} p = \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx 0.368$$

因此，虽然新集合 S^* 的样本总数与原集合 S 的样本总数相等（都为 n），但是新集合 S^* 中可能包含了重复的样本（有放回抽取）。若除去重复的样本，新集合 S^* 中仅包含了原集合 S 中约 1 - 0.368 = 63.2%的样本。

随机森林概述

- **Bagging算法**

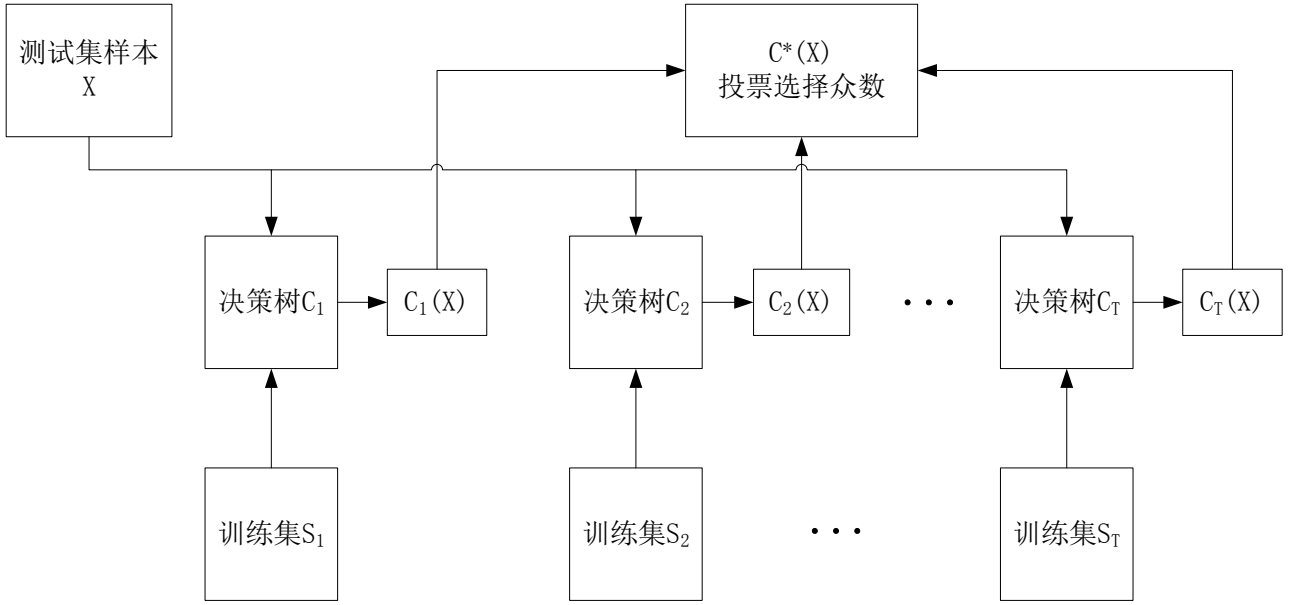
Bagging（Bootstrap aggregating 的缩写）算法是最早的集成学习算法，具体的步骤可以描述为：

（1） 利用 Bootstrap 方法重采样，随机产生 T 个训练集 $S_1, S_2, \ldots, S_T$；

（2） 利用每个训练集，生成对应的决策树 $C_1, C_2, \ldots, C_T$；

（3） 对于测试集样本 X，利用每个决策树进行测试，得到对应的类别 $C_1(X), C_2(X), \ldots, C_T(X)$；

（4） 采用投票的方法，将 T 个决策树中输出最多的类别作为测试集样本 X 所属的类别。



随机森林概述

- **随机森林算法**

随机森林算法与 Bagging 算法类似，均是基于 Bootstrap 方法重采样，产生多个训练集。不同的是，随机森林算法在构建决策树的时候，采用了随机选取分裂属性集的方法。详细的随机森林算法流程如下所示：

不妨设样本的属性个数为 M ， m 为大于零且小于 M 的整数。

（1） 利用 Bootstrap 方法重采样，随机产生 T 个训练集 $S_1, S_2, \ldots, S_T$ ；

（2） 利用每个训练集，生成对应的决策树 $C_1, C_2, \ldots, C_T$ ；在每个非叶子节点（内部节点）上选择属性前，从 M 个属性中随机抽取 m 个属性作为当前节点的分裂属性集，并以这 m 个属性中最好的分裂方式对该节点进行分裂（一般而言，在整个森林的生长过程中， m 的值维持不变）。

（3） 每棵树都完整成长，而不进行剪枝。

（4） 对于测试集样本 X ，利用每个决策树进行测试，得到对应的类别 $C_1(X), C_2(X), \ldots, C_T(X)$ ；

（5） 采用投票的方法，将 T 个决策树中输出最多的类别作为测试集样本 X 所属的类别。

重点函数解读

- ClassificationTree.fit
- view
- predict
- classRF_train
- classRF_predict
- TreeBagger

案例分析

决策树-乳腺癌诊断

随机森林-乳腺癌诊断

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间