

刘建平Pinard

十年码农，对数学统计学，数据挖掘，机器学习，大数据平台，大数据平台应用开发，大数据可视化感兴趣。

[博客园](#) [首页](#) [新随笔](#) [联系](#) [订阅](#) [管理](#)

决策树算法原理(上)

决策树算法在机器学习中算是很经典的一个算法系列了。它既可以作为分类算法，也可以作为回归算法，同时也特别适合集成学习比如随机森林。本文就对决策树算法原理做一个总结，上篇对ID3，C4.5的算法思想做了总结，下篇重点对CART算法做一个详细的介绍。选择CART做重点介绍的原因是scikit-learn使用了优化版的CART算法作为其决策树算法的实现。

1. 决策树ID3算法的信息论基础

机器学习算法其实很古老，作为一个码农经常会不停的敲if, else if, else,其实就已经在用到决策树的思想了。只是你没有想过，有这么多条件，用哪个条件特征先做if, 哪个条件特征后做if比较优呢？怎么准确的定量选择这个标准就是决策树机器学习算法的关键了。1970年代，一个叫昆兰的大牛找到了用信息论中的熵来度量决策树的决策选择过程，方法一出，它的简洁和高效就引起了轰动，昆兰把这个算法叫做ID3。下面我们就看看ID3算法是怎么选择特征的。

首先，我们需要熟悉信息论中熵的概念。熵度量了事物的不确定性，越不确定的事物，它的熵就越大。具体的，随机变量X的熵的表达式如下：

$$H(X) = - \sum_{i=1}^n p_i \log p_i$$

其中n代表X的n种不同的离散取值。而 p_i 代表了X取值为i的概率，log为以2或者e为底的对数。举个例子，比如X有2个可能的取值，而这两个取值各为1/2时X的熵最大，此时X具有最大的不确定性。值为

$H(X) = -(\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2}) = \log 2$ 。如果一个值概率大于1/2，另一个值概率小于1/2，则不确定性减少，对应的熵也会减少。比如一个概率1/3，一个概率2/3，则对应熵为 $H(X) = -(\frac{1}{3} \log \frac{1}{3} + \frac{2}{3} \log \frac{2}{3}) = \log 3 - \frac{2}{3} \log 2 < \log 2$ 。

熟悉了一个变量X的熵，很容易推广到多个变量的联合熵，这里给出两个变量X和Y的联合熵表达式：

公告

★珠江追梦，饮岭南茶，恋鄂北家★

昵称：刘建平Pinard

园龄：1年5个月

粉丝：1150

关注：13

[+加关注](#)

2018年4月						
日	一	二	三	四	五	六
25	26	27	28	29	30	31
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	1	2	3	4	5

常用链接

[我的随笔](#)

[我的评论](#)

[我的参与](#)

[最新评论](#)

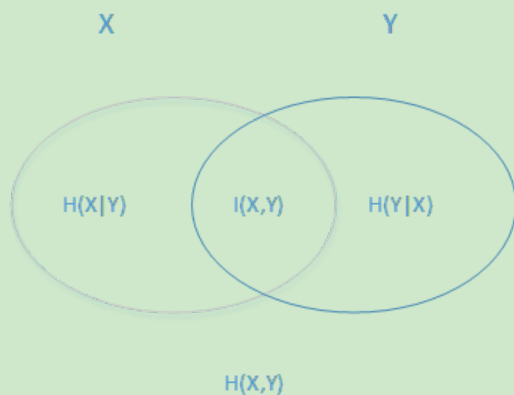
$$H(X,Y) = - \sum_{i=1}^n p(x_i, y_i) \log p(x_i, y_i)$$

有了联合熵，又可以得到条件熵的表达式 $H(X|Y)$ ，条件熵类似于条件概率，它度量了我们的 X 在知道 Y 以后剩下的不确定性。表达式如下：

$$H(X|Y) = - \sum_{i=1}^n p(x_i, y_i) \log p(x_i | y_i) = \sum_{j=1}^n p(y_j) H(X|y_j)$$

好吧，绕了一大圈，终于可以重新回到ID3算法了。我们刚才提到 $H(X)$ 度量了 X 的不确定性，条件熵 $H(X|Y)$ 度量了我们在知道 Y 以后 X 剩下的不确定性，那么 $H(X)-H(X|Y)$ 呢？从上面的描述大家可以看出，它度量了 X 在知道 Y 以后不确定性减少程度，这个度量我们在信息论中称为互信息，记为 $I(X,Y)$ 。在决策树ID3算法中叫做信息增益。ID3算法就是用信息增益来判断当前节点应该用什么特征来构建决策树。信息增益大，则越适合用来分类。

上面一堆概念，大家估计比较晕，用下面这个图很容易明白他们的关系。左边的椭圆代表 $H(X)$ ，右边的椭圆代表 $H(Y)$ ，中间重合的部分就是我们的互信息或者信息增益 $I(X,Y)$ ，左边的椭圆去掉重合部分就是 $H(X|Y)$ ，右边的椭圆去掉重合部分就是 $H(Y|X)$ 。两个椭圆的并就是 $H(X,Y)$ 。



2. 决策树ID3算法的思路

上面提到ID3算法就是用信息增益大小来判断当前节点应该用什么特征来构建决策树，用计算出的信息增益最大的特征来建立决策树的当前节点。这里我们举一个信息增益计算的具体的例子。比如我们有15个样本 D ，输出为0或者1。其中有9个输出为1，6个输出为0。样本中有个特征 A ，取值为 $A1$ ， $A2$ 和 $A3$ 。在取值为 $A1$ 的样本的输出中，有3个输出为1，2个输出为0，取值为 $A2$ 的样本输出中，2个输出为1，3个输出为0，在取值为 $A3$ 的样本中，4个输出为1，1个输出为0。

$$\text{样本} D \text{的熵: } H(D) = -\left(\frac{9}{15} \log_2 \frac{9}{15} + \frac{6}{15} \log_2 \frac{6}{15}\right) = 0.971$$

我的标签

随笔分类(101)

- 0040. 数学统计学(4)
- 0081. 机器学习(62)
- 0082. 深度学习(10)
- 0083. 自然语言处理(23)
- 0121. 大数据挖掘(1)
- 0122. 大数据平台(1)
- 0123. 大数据可视化

随笔档案(101)

- 2017年8月 (1)
- 2017年7月 (3)
- 2017年6月 (8)
- 2017年5月 (7)
- 2017年4月 (5)
- 2017年3月 (10)
- 2017年2月 (7)
- 2017年1月 (13)
- 2016年12月 (17)
- 2016年11月 (22)
- 2016年10月 (8)

常去的机器学习网站

52 NLP
Analytics Vidhya
机器学习库
机器学习路线图
深度学习进阶书

样本D在特征下的条件熵为: $H(D|A) = \frac{5}{15}H(D1) + \frac{5}{15}H(D2) + \frac{5}{15}H(D3)$

$$= -\frac{5}{15}(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}) - \frac{5}{15}(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}) - \frac{5}{15}(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}) = 0.888$$

对应的信息增益为 $I(D, A) = H(D) - H(D|A) = 0.083$

下面我们看看具体算法过程大概是怎么样的。

输入的是m个样本，样本输出集合为D，每个样本有n个离散特征，特征集合即为A，输出为决策树T。

算法的过程为：

- 1) 初始化信息增益的阈值 ϵ
- 2) 判断样本是否为同一类输出 D_i ，如果是则返回单节点树T。标记类别为 D_i
- 3) 判断特征是否为空，如果是则返回单节点树T，标记类别为样本中输出类别D实例数最多的类别。
- 4) 计算A中的各个特征（一共n个）对输出D的信息增益，选择信息增益最大的特征 A_g
- 5) 如果 A_g 的信息增益小于阈值 ϵ ，则返回单节点树T，标记类别为样本中输出类别D实例数最多的类别。
- 6) 否则，按特征 A_g 的不同取值 A_{gi} 将对应的样本输出D分成不同的类别 D_i 。每个类别产生一个子节点。对应特征值为 A_{gi} 。返回增加了节点的数T。
- 7) 对于所有的子节点，令 $D = D_i, A = A - \{A_g\}$ 递归调用2-6步，得到子树 T_i 并返回。

3. 决策树ID3算法的不足

ID3算法虽然提出了新思路，但是还是有很多值得改进的地方。

a) ID3没有考虑连续特征，比如长度，密度都是连续值，无法在ID3运用。这大大限制了ID3的用途。

b) ID3采用信息增益大的特征优先建立决策树的节点。很快就被发现，在相同条件下，取值比较多的特征比取值少的特征信息增益大。比如一个变量有2个值，各为1/2，另一个变量为3个值，各为1/3，其实他们都是完全不确定的变量，但是取3个值的比取2个值的信息增益大。如果校正这个问题呢？

c) ID3算法对于缺失值的情况没有做考虑

d) 没有考虑过拟合的问题

深度学习入门书

积分与排名

积分 - 303043

排名 - 608

阅读排行榜

1. 梯度下降（Gradient Descent）小结(103565)
2. 梯度提升树(GBDT)原理小结(51305)
3. 线性判别分析LDA原理总结(34105)
4. scikit-learn决策树算法类库使用小结(29434)
5. 谱聚类（spectral clustering）原理总结(22596)

评论排行榜

1. 梯度提升树(GBDT)原理小结(82)
2. 谱聚类（spectral clustering）原理总结(75)
3. 梯度下降（Gradient Descent）小结(65)
4. 卷积神经网络(CNN)反向传播算法(57)
5. 集成学习之Adaboost算法原理小结(50)

推荐排行榜

1. 梯度下降（Gradient Descent）小结(43)
2. 集成学习原理小结(15)
3. 奇异值分解(SVD)原理与在降维中的应用(15)
4. 卷积神经网络(CNN)反向传播算法(14)
5. 集成学习之Adaboost算法原理小结(13)

ID3 算法的作者昆兰基于上述不足，对ID3算法做了改进，这就是C4.5算法，也许你会问，为什么不叫ID4，ID5之类的名字呢？那是因为决策树太火爆，他的ID3一出来，别人二次创新，很快就占了ID4，ID5，所以他另辟蹊径，取名C4.0算法，后来的进化版为C4.5算法。下面我们就来聊下C4.5算法

4. 决策树C4.5算法的改进

上一节我们讲到ID3算法有四个主要的不足，一是不能处理连续特征，第二个就是用信息增益作为标准容易偏向于取值较多的特征，最后两个是缺失值处理的问题和过拟合问题。昆兰在C4.5算法中改进了上述4个问题。

对于第一个问题，不能处理连续特征，C4.5的思路是将连续的特征离散化。比如m个样本的连续特征A有m个，从小到大排列为 a_1, a_2, \dots, a_m ，则C4.5取相邻两样本值的平均数，一共取得m-1个划分点，其中第i个划分点 T_i 表示为：

$T_i = \frac{a_i + a_{i+1}}{2}$ 。对于这m-1个点，分别计算以该点作为二元分类点时的信息增益。选择信息增益最大的点作为该连续特征的二元离散分类点。比如取到的增益最大的点为 a_t ，则小于 a_t 的值为类别1，大于 a_t 的值为类别2，这样我们就做到了连续特征的离散化。要注意的是，与离散属性不同的是，如果当前节点为连续属性，则该属性后面还可以参与子节点的产生选择过程。

对于第二个问题，信息增益作为标准容易偏向于取值较多的特征的问题。我们引入一个信息增益比的变量 $I_R(X, Y)$ ，它是信息增益和特征熵的比值。表达式如下：

$$I_R(D, A) = \frac{I(A, D)}{H_A(D)}$$

其中D为样本特征输出的集合，A为样本特征，对于特征熵 $H_A(D)$ ，表达式如下：

$$H_A(D) = - \sum_{i=1}^n \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$$

其中n为特征A的类别数， D_i 为特征A的第i个取值对应的样本个数。D为样本个数。

特征数越多的特征对应的特征熵越大，它作为分母，可以校正信息增益容易偏向于取值较多的特征的问题。

对于第三个缺失值处理的问题，主要需要解决的是两个问题，一是在样本某些特征缺失的情况下选择划分的属性，二是选定了划分属性，对于在该属性上缺失特征的样本的处理。

对于第一个子问题，对于某一个有缺失特征值的特征A。C4.5的思路是将数据分成两部分，对每个样本设置一个权重（初始可以都为1），然后划分数据，一部分是有特征值A的数据D1，另一部分是没有特征A的数据D2。然后对于没有缺失特征A的数据集D1来和对应的A特征的各个特征值一起计算加权后的信息增益比，最后乘上一个系数，这个系数是无特征A缺失的样本加权后所占加权总样本的比例。

对于第二个子问题，可以将缺失特征的样本同时划分入所有的子节点，不过将该样本的权重按各个子节点样本的数量比例来分配。比如缺失特征A的样本a之前权重为1，特征A有3个特征值A1,A2,A3。3个特征值对应的无缺失A特征的样本个数为

2,3,4.则a同时划分入A1, A2, A3。对应权重调节为2/9,3/9, 4/9。

对于第4个问题, C4.5引入了正则化系数进行初步的剪枝。具体方法这里不讨论。下篇讲CART的时候会详细讨论剪枝的思路。

除了上面的4点, C4.5和ID的思路区别不大。

5. 决策树C4.5算法的不足与思考

C4.5虽然改进或者改善了ID3算法的几个主要的问题, 仍然有优化的空间。

1)由于决策树算法非常容易过拟合, 因此对于生成的决策树必须要进行剪枝。剪枝的算法有非常多, C4.5的剪枝方法有优化的空间。思路主要是两种, 一种是预剪枝, 即在生成决策树的时候就决定是否剪枝。另一个是后剪枝, 即先生成决策树, 再通过交叉验证来剪枝。后面在下篇讲CART树的时候我们会专门讲决策树的减枝思路, 主要采用的是后剪枝加上交叉验证选择最合适的决策树。

2)C4.5生成的是多叉树, 即一个父节点可以有多个节点。很多时候, 在计算机中二叉树模型会比多叉树运算效率高。如果采用二叉树, 可以提高效率。

3)C4.5只能用于分类, 如果能将决策树用于回归的话可以扩大它的使用范围。

4)C4.5由于使用了熵模型, 里面有大量的耗时的对数运算,如果是连续值还有大量的排序运算。如果能够加以模型简化可以减少运算强度但又不牺牲太多准确性的话, 那就更好了。

这4个问题在CART树里面部分加以了改进。所以目前如果不考虑集成学习话, 在普通的决策树算法里, CART算法算是比较优的算法了。scikit-learn的决策树使用的也是CART算法。在下篇里我们会重点聊下CART算法的主要改进思路, 上篇就到这里。下篇请看[决策树算法原理\(下\)](#)。

(欢迎转载, 转载请注明出处。欢迎沟通交流: pinard.liu@ericsson.com)

分类: [0081. 机器学习](#)

标签: [分类算法](#)

好文要顶

关注我

收藏该文

[刘建平Pinard](#)
[关注 - 13](#)
[粉丝 - 1150](#)

[+加关注](#)

« 上一篇: [机器学习算法的随机数据生成](#)
» 下一篇: [决策树算法原理\(下\)](#)

9

推荐

0

反对

posted @ 2016-11-10 15:54 刘建平Pinard 阅读(14643) 评论(23) 编辑 收藏

评论列表

- #1楼 2017-04-21 16:50 ShaunAgain

条件熵公式是不是写错?

支持(0) 反对(0)
- #2楼[楼主] 2017-04-22 08:28 刘建平Pinard

@ ShaunAgain
你好, 个人觉得应该没有错误。若有不对请帮忙具体指出

支持(0) 反对(0)
- #3楼 2017-04-28 08:13 ShaunAgain

@ 刘建平Pinard
是我搞错了, 起初没有看懂推导顺序。谢谢博主好资料。
https://en.wikipedia.org/wiki/Conditional_entropy

支持(0) 反对(0)
- #4楼 2017-08-24 22:21 曹真

这些都是博主原创的吗? 良心资料 感谢☺

支持(0) 反对(0)

#5楼[楼主] 2017-08-25 10:32 刘建平Pinard



@ 曹真

你好，是我根据自己整理的学习笔记和一些项目实践总结的。

支持(0) 反对(0)

#6楼 2017-09-23 20:23 Bitter-Coffe



楼主，C4.5处理连续值是相邻两个数的平均数而不是中位数，楼主笔误啦。。

支持(0) 反对(0)

#7楼 2017-09-23 21:32 Bitter-Coffe



博主，能具体说一下以信息增益作为评价标准偏向于选择较多的特征吗？谢谢博主！PS：最近跟随您的博客学习机器学习的相关知识，收益良多，再次感谢！！

支持(0) 反对(0)

#8楼[楼主] 2017-09-25 10:50 刘建平Pinard



@ Bitter-Coffe

的确是平均数，感谢指出错误，已经修改。

支持(0) 反对(0)

#9楼[楼主] 2017-09-25 11:44 刘建平Pinard



@ Bitter-Coffe

能帮到你很高兴，一起学习！

对于信息增益作为评价标准偏向于选择较多的特征，其实是很简单的。从信息增益的定义我们知道，它度量了输出在知道特征以后不确定性减少程度。

当特征类别数较多时候，这个不确定性减少程度一般会多一些，因为此时特征类别数较多，每个类别里的样本数量较少，样本更容易被划分的散落到各个特征类别，即样本不确定性变小。

当然，这个说法是针对大多数情况的。某一些极端的样本分布，有可能特征类别少的反而可能信息增益大。

支持(0) 反对(0)


#10楼 2017-09-29 09:53 Bitter-Coffe



@ 刘建平Pinard

讲的清楚明白，非常感谢！


支持(0) 反对(0)

#11楼 2017-10-31 09:56 周伟峰 



谢谢

支持(0) 反对(0)

#12楼 2017-12-25 17:27 jdnjj 





第二点中的例子是不是应该有9个1 6个0?

[@刘建平Pinard](#)

谢谢

支持(0) 反对(0)


#13楼[楼主 ] 2017-12-26 10:50 刘建平Pinard 



[@ jdnjj](#)

你好，的确写反了，感谢指正，已经修改。

支持(0) 反对(0)

#14楼 2018-01-07 14:47 JeaDong233 





[@ 刘建平Pinard](#)

老实您好，对于你评论区的这段话不是很理解。

（当特征类别数较多时候，这个不确定性减少程度一般会多一些，因为此时特征类别数较多，每个类别里的样本数量较少，样本更容易被划分的散落到各个特征类别，即样本不确定性变小。）

样本不确定性不应该是变大了么？比如特征A里面只有两个取值而且概率相等，和特征B里只有3个取值且概率相等。算出来的熵也是特征B的多。

支持(0) 反对(0)

#15楼[楼主 ] 2018-01-08 11:07 刘建平Pinard 




[@ JeaDong233](#)

你好，这里肯定是可以举出信息增益比信息增益好的具体例子的。

不过我们考虑的是大多数情况，有更多类别的特征更容易帮我们筛选出该特征和输出之间的对应关系，这样更容易被信息增益法选中。

支持(0) 反对(0)

#16楼 2018-01-22 11:44 小俊俊俊 



老师您好，请问在第二节中，算法实现过程，

第二步：判断样本是否为同一类输出D_i，是怎样判断的？是说所有的样本输出都是一样的吗？例如都是1或者都是0？

第三步：判断特征为空，是什么意思？

刚开始学习，望老师指教~谢谢~

支持(0) 反对(0)

#17楼[楼主] 2018-01-22 16:55 刘建平Pinard



@ 小俊俊俊

你好，第二步判断的是每个训练样本的输出类别，所以都是已知的。如果所有的输出类别都是一样的，就像你说的都是1或者0，那么该样本集为同一类别输出。

第三步是因为离散特征在每次递归分裂子树后会从特征集合 A 中删除，这样在树做递归分裂的时候 A 集合里的特征会变少（具体在第7步更新 A ），当 A 里所有的特征都被使用完后，该递归调用结束，子树生成完成。

支持(0) 反对(0)

#18楼 2018-01-22 18:51 小俊俊俊



@ 刘建平Pinard

如果该样本集为同一类别，就不再执行后边操作，对吗？

支持(0) 反对(0)

#19楼 2018-01-23 08:53 小俊俊俊



@ 刘建平Pinard

嗯嗯，都明白了，谢谢老师！

支持(0) 反对(0)

#20楼[楼主] 2018-01-23 10:33 刘建平Pinard



@ 小俊俊俊

是的，如果该样本集为同一类别，那么就没有预测价值啦。因为所有的样本输入它都会认为是同一个输出，不需要预测了，不再执行后面的操作。

支持(0) 反对(0)

#21楼 2018-03-28 11:13 博客之游者



老师您好，请教一个问题。在C4.5算法的改进中针对第三个问题的改进有点疑问，对于第一个子问题中那个设置权重的作用是什么，如果初始都设置权重为1，岂不是等同于不设置权重，这个权重后期怎么变化？还有最后一个系数，那个系数是不是就是指无特征值样本占总样本容量的比例？

支持(0) 反对(0)

#22楼[楼主] 2018-03-29 16:08 刘建平Pinard



@ 博客之游者

你好。

这个权重是可以自己调节的，如果各个样本的重要性没有特别需求，那么就都为1.也就是这个权重后面并不起实际作用。但是如果各个样本的权重重要性不一样，那么这个值就不是都为1。

最后那个系数，如果你的样本权重都是1，那么就是有特征值样本占总样本容量的比例。但是如果你前面的权重不是都为1，那么这里就是加权系数，即有特征值样本权重和占总样本权重和的比例。

支持(0) 反对(0)

#23楼 2018-03-29 16:10 博客之游者



@ 刘建平Pinard

好的，明白了，谢谢老师！

支持(0) 反对(0)

[刷新评论](#) [刷新页面](#) [返回顶部](#)



注册用户登录后才能发表评论，请 [登录](#) 或 [注册](#)，[访问网站首页](#)。

最新IT新闻:

- Spotify上市首日开盘价165.9美元，较参考股价上涨
 - 胡玮炜深夜发声证实美团收购摩拜 否认创始团队出局
 - Spotify早期投资者说绝不出售股票，认为公司价值千亿美元
 - 特斯拉回应外界质疑：现金流充足无需再融资
 - 走出领英后，沈博阳的目标是打造租房领域的超级独角兽
- » 更多新闻...

最新知识库文章:

- 写给自学者的入门指南
 - 和程序员谈恋爱
 - 学会学习
 - 优秀技术人的管理陷阱
 - 作为一个程序员，数学对你到底有多重要
- » 更多知识库文章...