决策树归纳法实验报告

17231067 张瑞达

1. 实验目的
2. 理解为什么我们使用基于熵的测度来构造决策树。

因为熵可以反映一个问题的不确定性，通过某种规则可以使得问题的不确定性越来越小，从而解决问题。

1. 了解在构建决策树的过程中如何使用信息增益来选择属性

遍历所有的属性，分别计算，使用这个属性划分数据集前后信息熵的变化值，然后选择信息熵变化幅度最大的那个属性，来优先作为划分数据集（分裂节点）的依据。**即选择信息增益最大的特征作为分裂节点。**

1. 理解决策树与一组规则的等价性。

可以将决策树看做一组if-then规则的集合，由决策树的根节点到叶节点的每一条路径对应一条规则。

1. 为了理解为什么我们有时需要修剪树以及我们如何修剪树？基于我们修剪决策树的方法。

使用决策树模型拟合数据时，容易造成过拟合。解决过拟合的方法是对决策树进行剪枝处理。剪枝有两种思路：前剪枝和后剪枝。

前剪枝：前剪枝是在构造决策树的同时进行剪枝。

①在决策树的构建过程中，设定一个阈值，信息熵减小的数量小于这个阈值，即使还可以继续降低信息熵，也停止继续创建分支。

②限制叶子节点的样本个数，当样本个数小于一定的阈值时，不再继续创建分支。

③设定一个高度，当达构建的树达到那个高度的时候，停止建立决策树。

后剪枝：构建完成后，对拥有同样父节点的一组节点进行检查，判断如果将其合并，信息熵的增加量是否小于某一阈值。如果小于阈值，则这一组节点可以合并为一个节点。

5. 了解软决策树的概念，以及为什么它们是经典决策树的重要扩展。

软决策树引入了模糊集合以及隶属度的概念，不同于硬决策树1或者0的判断，软决策树得到的是对模糊集合的隶属度，即样本对于模糊集合的属于程度。软决策树相比于硬决策树泛化能力更强，可解释性更好。

1. 实验内容与分析

阶段1：

根据下例完成后续阶段1的实验：

|  |  |  |  |
| --- | --- | --- | --- |
| 季节 | 时间已过8点 | 风力情况 | 赖床 |
| 春 | 0 | 小风 | 1 |
| 冬 | 0 | 无风 | 1 |
| 秋 | 1 | 小风 | 1 |
| 冬 | 0 | 无风 | 1 |
| 夏 | 0 | 小风 | 1 |
| 冬 | 1 | 小风 | 1 |
| 冬 | 0 | 大风 | 1 |
| 冬 | 0 | 无风 | 1 |
| 春 | 1 | 无风 | 0 |
| 夏 | 1 | 大风 | 0 |
| 夏 | 0 | 大风 | 0 |
| 秋 | 1 | 小风 | 0 |

这是小明一年中12天的赖床情况，有三个属性会影响到结果：季节、风力情况、是否已过8点。其中0表示否定，1表示肯定。

1. 实现代码来计算每个属性的信息熵。
   1. 季节的信息熵：整理表格中的信息得到

|  |  |  |  |
| --- | --- | --- | --- |
|  | 赖床 | 不赖床 | 合计 |
| 春 | 1 | 1 | 2 |
| 夏 | 1 | 2 | 3 |
| 秋 | 1 | 1 | 2 |
| 冬 | 5 | 0 | 5 |
| 合计 | 8 | 4 | 12 |

代码如下：

Eg1=-1/2\*log2(1/2)-1/2\*log2(1/2)

Eg2=-2/3\*log2(2/3)-1/3\*log2(1/3)

Eg3=-1/2\*log2(1/2)-1/2\*log2(1/2)

Eg4=0

info\_1=1/6\*Eg1+1/4\*Eg2+1/6\*Eg3+5/12\*Eg4

运行结果：

Eg1 = 1

Eg2 = 0.9183

Eg3 = 1

Eg4 = 0

info\_1 = 0.5629

* 1. 时间是否已过8点的信息熵：整理表格得到：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 赖床 | 不赖床 | 合计 |
| 已过8点 | 2 | 3 | 5 |
| 未过8点 | 6 | 1 | 7 |
| 合计 | 8 | 4 | 12 |

代码如下：

Eg1=-2/5\*log2(2/5)-3/5\*log2(3/5)

Eg2=-6/7\*log2(6/7)-1/7\*log2(1/7)

info\_2=5/12\*Eg1+7/12\*Eg2

运行结果：

Eg1 = 0.9710

Eg2 = 0.5917

info\_2 = 0.7497

* 1. 风力情况的信息熵：整理表格得到：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 赖床 | 不赖床 | 合计 |
| 小风 | 4 | 1 | 5 |
| 无风 | 3 | 1 | 4 |
| 大风 | 1 | 2 | 3 |
| 合计 | 8 | 4 | 12 |

代码如下：

Eg1=-1/5\*log2(1/5)-4/5\*log2(4/5)

Eg2=-3/4\*log2(3/4)-1/4\*log2(1/4)

Eg3=-1/3\*log2(1/3)-2/3\*log2(2/3)

info\_3=5/12\*Eg1+4/12\*Eg2+3/12\*Eg3

运行结果：

Eg1 = 0.7219

Eg2 = 0.8113

Eg3 = 0.9183

info\_3 = 0.8008

1. 从给定的分类问题中选择信息量最大的属性

根据（1）的计算结果，以季节为分类依据时信息熵最小，即信息量最大。

1. 找到一个合适的数据结构来表示一个决策树。基于第3.2节中讨论的原理，使用信息增益引导启发法。

整体熵：D=-1/3\*log2(1/3)-2/3\*log2(2/3)= 0.9183

季节信息增益：info\_1(D)=0.9183-0.5629=0.3554

时间信息增益：info\_2(D)=0.9183-0.7497=0.1686

风力信息增益：info\_3(D)=0.9183-0.8008=0.1175

1. 第一步以季节为分类依据，春、夏、秋需要进一步分类，冬可以确定赖床。
2. 对于春：

整体熵：D=-1/2\*log2(1/2)-1/2\*log2(1/2)= 1

时间信息增益：info\_1(D)=1-0=1

风力信息增益：info\_2(D)=1-0=1

所以任选时间或风力为分类依据，这里选择时间，可以确定超过8点不赖床，未超过8点赖床。

1. 对于夏

整体熵：D=-1/3\*log2(1/3)-2/3\*log2(2/3)= 0.9183

时间信息增益：info\_1(D)=0.9183-0.6667=0.2516

风力信息增益：info\_2(D)=0.9183-0=0.9183

所以选择风力为分类依据，可以确定超过小风时赖床，大风时不赖床。

1. 对于秋

整体熵：D=-1/2\*log2(1/2)-1/2\*log2(1/2)= 1

时间信息增益：info\_1(D)=1-1=0

风力信息增益：info\_2(D)=1-0=1

所以选择风力为分类依据，可以确定超过小风时赖床，大风时不赖床。

季节

春 夏 秋 冬

风力

风力

时间

冬

已过8点 未过8点 小风 大风

赖床

赖床

不赖床

赖床

不赖床

赖床

不赖床

阶段2：

1. 现在考虑连续属性或混合属性(包括连续属性和离散属性)的情况，我们如何处理决策树？你能提出一些离散化的方法吗？

将连续属性离散化以后按照离散属性的方法处理决策树。

离散化方法：数值积分法①一阶向前差分

②一阶向后差分

③双线性变换法

④修正双线性变换法

1. 树的大小和模型的精确度之间是否存在权衡？在紧凑性和性能方面是否存在一棵最优树？
   1. 一般来说存在，随着决策树层数的不断增加，信息熵越来越小，问题的不确定性越来越低，从而模型更加精确。但是决策树过大可能会带来过拟合的问题，为了避免过拟合有时需要放弃一定精确度的要求。
   2. 存在，决策树扩展时，设定一个阈值，如果信息熵的增量小于这一阈值则停止扩展；决策树扩展完成后，通过后剪枝的方法在能够容忍的范围内将一些节点合并，此时的决策树为最优解。
2. 对于一个数据元素，经典的决策树给出了一个硬界限来决定遵循哪个分支，您能否提出一种“软方法”来增强决策树的健壮性？

划分为若干个模糊集合，例如：抛弃成绩>90，成绩<90的分类依据；采取成绩好，成绩差的分类依据。最终得到的也不再是1或0的结果，而是对于不同模糊集合的隶属度，此时决定遵循哪个分支更加自由。

（4）与天真的贝叶斯方法相比，决策树学习的优点和缺点是什么？

优点：①不需要知道属性之间的内在联系，易于操作。

②可视化强，易于理解。

③适合高维数据。

缺点：①忽略了属性之间的内在联系。

②容易过拟合。

1. 总结

通过本次实验，熟悉了决策树的构建过程；体会到了决策树的优点和缺点；了解了软决策树的概念和优点；知道了控制决策树的大小的基本依据；明白了剪枝的原因和方法。