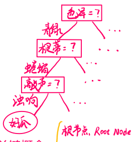
决策树

2023年11月14日 9:47

基本流程



关键概念
- 根节点, Root Node
- 内部节点, Interal Nodes
- 叶节点, Leaf Nodes

问题：先后顺序，何时停止

划分选择

信息增益 / 决策树ID3训练算法

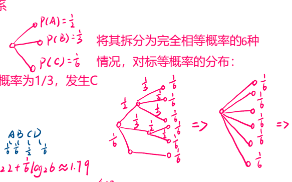
信息与熵的概念与度量 / 信息熵 / 信息增益

概念

- 熵　对一种事物的不确定性称为熵　（概率和信息熵是随机事件的一体两面）
  - 概率
    - 某事有多大可能性发生
    - 表征了事件的确定性
  - 信息熵
    - 某事有多少种可能性
    - 表征了事件的不确定性
- 信息
  - 消除不确定的事物
  - 作用
    - 调整概率
    - 排除干扰
    - 确定情况
- 噪音　不能消除某人对某件事情不确定性的事物
- 数据　噪音+信息

熵如何量化

- 参照一个不确定的事件用作单位　例：抛硬币的不確定性
  - 概率：50%正面，50%反面；相当于猜一次抛硬币的不确定性，记为1bit
  - 投硬币次数与结果不确定性呈指数关系
- 等概率 / 均匀分布
  - 8个等概率的不确定情况，相当于抛8个硬币($2^3=8$)，熵为3bit
  - 假设有m种等概率的不同情况，那么熵为 $\log_2 m$
- 每种情况概率不相等 / 一般分布　计算公式 $Ent(D)=-\sum_{k=1}^{|y|} p_k \log_2 p_k$
  - 例：假设发生A的概率为1/2，发生B的概率为1/3，发生C的概率为1/6，该如何求熵?
  - $P(A)=\frac{1}{2}$，$P(B)=\frac{1}{3}$，$P(C)=\frac{1}{6}$　将其拆分为完全相等概率的6种情况，对标等概率的分布：



$\Rightarrow \log_2 6$ 有 $\frac{1}{2}(\log_2 6-\log_2 3)+\frac{1}{3}(\log_2 6-\log_2 2)+\frac{1}{6}(\log_2 6-\log_2 1)=\frac{1}{2}\log_2\frac{6}{3}+\frac{1}{3}\log_2\frac{6}{2}+\frac{1}{6}\log_2\frac{6}{1}=\frac{1}{2}\log_2\frac{2}{1}+\frac{1}{3}\log_2\frac{3}{1}+\frac{1}{6}\log_2\frac{6}{1}$

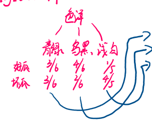
信息如何量化　得知信息前后，熵的差额就是信息的量

- 假设我并不知道选择题是ABCD哪个选项时的熵 $\log_2 4=2$ ($m=4$)
- 有人告诉我C有一半概率正确后的熵 $\frac{1}{6}\log_2 6+\frac{1}{6}\log_2 6+\frac{1}{2}\log_2 2+\frac{1}{6}\log_2 6\approx 1.79$　(A B C D: $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{2}$ $\frac{1}{6}$)
- 信息“C有一半概率正确”：$2-1.79=0.21$

例：西瓜数据集

编号	A.色泽	B.根蒂	C.敲声	D.纹理	E.脐部	D.触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑		…				…
⋮	…						…
17	浅白						否

色泽	青绿	乌黑	浅白
正例	3	4	1
反例	3	2	4



$Ent(D_1)=\frac{3}{6}\log_2\frac{6}{3}+\frac{3}{6}\log_2\frac{6}{3}=1$

$Ent(D_2)=\frac{4}{6}\log_2\frac{6}{4}+\frac{2}{6}\log_2\frac{6}{2}=0.918$

$Ent(D_3)=\frac{1}{5}\log_2\frac{5}{1}+\frac{4}{5}\log_2\frac{5}{4}=0.722$

目标：分类是好瓜还是坏瓜。共17个样本，其中8好9坏，属于一般分布，则总的熵为 $Ent(D)=\frac{8}{17}\log_2\frac{17}{8}+\frac{9}{17}\log_2\frac{17}{9}\approx 0.998$

若想使得机器基于信息增益生成的决策树具有最佳的分类效果，应当按照以下步骤执行：

1）欲确定第一次分类的属性，应当计算各属性下所有属性值的熵。以“色泽”为例，计算出不同色泽下的熵分别为 $Ent(D_1),Ent(D_2),Ent(D_3)$，则色泽的熵就是它们的熵的概率和，即 $[\frac{6}{17}Ent(D_1)+\frac{6}{17}Ent(D_2)+\frac{5}{17}Ent(D_3)]$，将其与总熵Ent(D)作差得到色泽的信息增益为0.109。以同样的方式求出其它属性的信息增益，最终得到纹理的信息增益=0.381最大，则第一轮应先分类纹理。

2）欲确定第一次分类后不同属性值的下一轮分类的属性，注意需要对第一次分类后不同属性值的各样本集分别分析。以色泽清晰的样本为例，数据总量为9，则计算其它不同属性的熵时只需要针对这8个数据，其余的数据忽视即可。若这9个样本中正例有7个，反例有2个，则未分类时的总熵为 $Ent(D')=\frac{7}{9}\log_2\frac{9}{7}+\frac{2}{9}\log_2\frac{9}{2}=0.764$。最终算得此时信息增益最大的属性是根蒂，即对于色泽清晰的样本，它们下一轮分类的属性是根蒂。同理，若样本的色泽是稍糊，则只针对色泽稍糊的样本计算它们其他属性的信息增益，最终得到信息增益最大的是触感，即对于色泽稍糊的样本，下一轮分类的属性是触感。

增益率 / 决策树C4.5训练算法

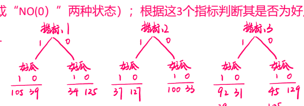
- 实际上，信息增益准则对可取值数目较多的属性有所偏好，为减少这种偏好可能带来的不利影响，可使用增益率来选择最优划分属性 $Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)}$
- 其中 $IV(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$ 称为属性a的固有值；属性a的可能取值数目越多，IV(a)值通常越大
- 对可取数目较少的属性有所偏好
  - 一般先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益最高的

基尼指数 / 决策树CART训练算法 / Classification and Regression

分类树：基尼指数最小准则

数据（各项指标只有“YES(1)”或“NO(0)”两种状态）；根据这3个指标判断其是否为好瓜（由于存在缺失值，3个类别统计后总数可能不同）

指标1	指标2	指标3	好瓜
0	0	1	1
1	1	0	1
1	0	Nan	0
⋮	⋮	…	



发现即使确定某项指数有或没有后，依然无法准确判断它是好瓜还是坏瓜，即不纯的(impurity)

同是正例：$\frac{105}{105+39}\times\frac{105}{105+39}$，同是反例：$\frac{39}{105+39}\times\frac{39}{105+39}$，则同一类别的概率则是同为正、反例概率和

纯度计算：以指标1为例，随机取两个样本，它们是同一类别的概率为 $(\frac{39}{105+39})^2+(\frac{105}{105+39})^2$，则不同的概率为 $1-(\frac{39}{105+39})^2-(\frac{105}{105+39})^2$。该值越大，说明这两个样本越可能不同，即纯度是下降的

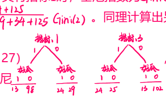
反之，若该值越小，则基尼系数越小，纯度越高。数据集D的纯度可用基尼值来度量 $Gini(D)=\sum_{k=1}^{|y|}\sum_{k'\neq k}p_k p_{k'}=1-\sum_{k=1}^{|y|}p_k^2$

公式
- $Gini(p)=\sum_{k=1}^{k}p_k(1-p_k)=1-\sum_{k=1}^{k}p_k^2$
- pk表示选中的样本属于k类别的概率，则这个样本被错分的概率为(1-pk)
- 样本集合有k个类别，一个随机选中的样本可以属于这k个类别中的任意一个，对类别进行加和
- 当为二分类时，Gini(p)=2p(1-p)

例：基于基尼指数构建决策树

1）要求根节点，首先需要计算3个指标的基尼指数。以指标1为例，样本具有指标1时，基尼指数为 $Gini(1)=1-(\frac{105}{105+39})^2-(\frac{39}{105+39})^2$，不具有指标1时，基尼指数为 $Gini(2)=1-(\frac{34}{34+125})^2-(\frac{125}{34+125})^2$ 则指标1的基尼指数就是它们的加权平均值 $\frac{105+39}{105+39+34+125}Gini(1)+\frac{34+125}{105+39+34+125}Gini(2)$。同理计算出另外两个指标的基尼指数，得到指标2的基尼指数=0.36最小，所以以指标2做第一次分叉。

2）对于第二次及以上分叉，以具有指标2的样本为例（m+=37，m-=127）



，首先计算出此时的基尼指数 $Gini(D')=1-(\frac{37}{37+127})^2-(\frac{127}{37+127})^2=0.35$，再计算具有指标2的样本，其指标1和指标3的基尼指数。最终得到指标3的基尼指数=0.29比指标1的小，并且小于Gini(D')，即不纯度下降。所以对于具有指标2的样本，下一次以指标3分叉。

3）注意，若分叉后得到的基尼指数大于分叉前的基尼指数，即纯度下降，就不应该分支。

剪枝处理

概述

目的：对付过拟合

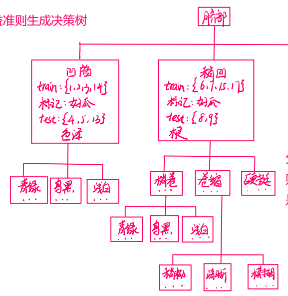
数据集

编号	A-F	好瓜
1-8	…	是
9-17	…	否

划分出训练集和验证集（该样本中以测试集充当训练集）
其中训练集有5个正例，5个反例；验证集有3个正例，4个反例

训练集
信息增益准则生成决策树

训练集与验证集在决策树上的情况



分类后的训练集样本若m+≥m-，则标记该分类为好瓜，否则为坏瓜；即通过剪枝将该属性变成叶节点时，该标记值就是机器最终判断结果

预剪枝

自上而下，若分完后，精度上升，则应该分；若精度不变或下降，则没必要分支，应剪枝

train: $m^+=5 \geq m^-=5$
标记: 好瓜
test: $m^+=3$, $m^-=4$
脐部

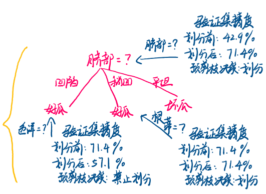
对于脐部
- 分类前，test精度 Accuracy $=\frac{3}{3+4}=42.9\%$　(Accuracy $=\frac{\text{正确预测的样本}}{\text{所有样本}}$)
- 分类后　test1: {4,5,13}(对应“凹陷”)，test2: {8,9}，test3: {11,12}
  test精度 $=\frac{5}{7}=71.4\%>42.9\%$，即分类后精度上升，应该分叉

将部分数据划分到“凹陷”后，对于色泽
- 分之前：5/7=71.4%
- 分之后，若有：

青绿	乌黑	浅白
train: {1}	train: {2,3}	train: {14}
标记: 好瓜	标记: 好瓜	标记: 坏瓜
test: {4,13}	test: ∅	test: {5}

由于将5号误分为坏瓜（之前分类正确，分支后分类错误）
即精度下降到4/7=57.1%，应剪枝

最终预剪枝决策树



后剪枝

对比预剪枝能剪则剪，后剪枝自下而上，能保留的都保留
1）自下而上计算精度: 3/7
2）把①剪掉后，精度: 4/7>3/7，该剪枝

train: {5:5} ($\frac{m^+}{m^-}$)
好
test: {3:4}
脐部

后剪枝

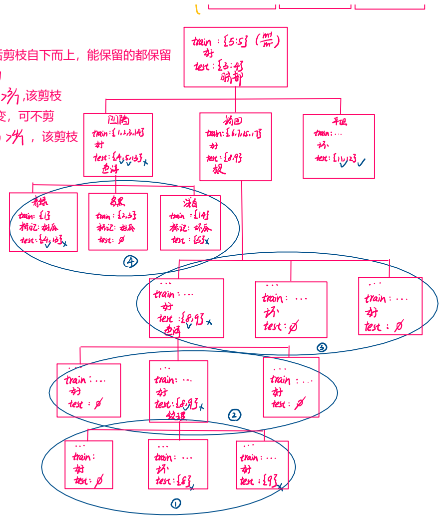
对比预剪枝能剪则剪，后剪枝自下而上，能保留的都保留

1）自下而上计算精度：3/7

2）把①剪掉后，精度：4/7 > 3/7，该剪枝

3）把②剪掉后，精度不变，可不剪

4）把④剪掉后，精度：5/7 > 4/7，该剪枝



连续与缺失值

连续值处理

二分法，使得信息增益最大的阈值

对于集合D，按照不同的属性进行分类，这些属性的属性值均为连续值

例

根據糖濃度对西瓜數據集分類

1）按照浓度从小到大的顺序给数据集样本排序

2）通过计算信息增益，查找每一次分类的最佳阈值，

缺失值处理

继续以剪枝处理的西瓜数据集D为例，即数据集为

编号	A-F	好瓜
1-8	...	是
9-17	...	否

其中，1-8是正例，9-17是反例。

1）以色泽属性为例，缺失了1号，5号，13号在色泽上的数据，即无缺失值的样例子集$\tilde{D}$中只有14个样例，计算分类前的信息熵$Ent(\tilde{D})=\frac{6}{14}\log_2\frac{14}{6}+\frac{8}{14}\log_2\frac{14}{8}\approx 0.985$

2）令$\tilde{D}^1, \tilde{D}^2, \tilde{D}^3$ 分别表示在色泽属性上取值为青绿，乌黑和浅白的样本子集，有$Ent(\tilde{D}^1)=\frac{2}{4}\log_2\frac{4}{2}+\frac{2}{4}\log_2\frac{4}{2}\approx 1.0$，$Ent(\tilde{D}^2)=\frac{4}{6}\log_2\frac{6}{4}+\frac{2}{6}\log_2\frac{6}{2}\approx 0.918$，$Ent(\tilde{D}^3)=\frac{4}{4}\log_2\frac{4}{4}=0$

3）因此，样本子集$\tilde{D}$上色泽的信息增益为$Gain(\tilde{D},色泽)=Ent(\tilde{D})-\sum_{v=1}^{3}\tilde{r}_v Ent(\tilde{D}^v)=0.985-(\frac{4}{14}\times Ent(\tilde{D}^1)+\frac{6}{14}\times Ent(\tilde{D}^2)+\frac{4}{14}\times Ent(\tilde{D}^3))=0.306$

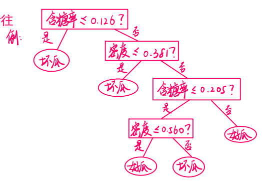
4）于是，样本集D上色泽属性的信息增益为$Gain(D,色泽)=\rho\times Gain(\tilde{D},色泽)=\frac{14}{17}\times 0.306=0.252$，类似地可计算出其它属性在D上的信息增益

对缺失值的惩罚系数：$len(\frac{\tilde{D}}{D})$

多变量决策树

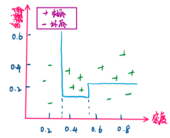
对于正常基于连续值生成的决策树，它们不同的属性最后往往呈线性的关系，因为每次的取值都是一次二分类

例：

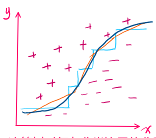


关于密度和含糖率生成的决策树

⇒

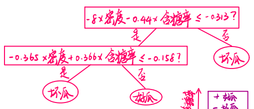


决策树对应的分类边界



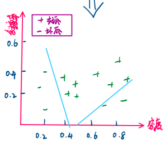
决策树对复杂分类边界的分段近似

多变量决策树



⇓

多变量决策树的分界点不再是对某个属性，而是对属性的线性组合进行一种测试



多变量决策树对应的分类边界