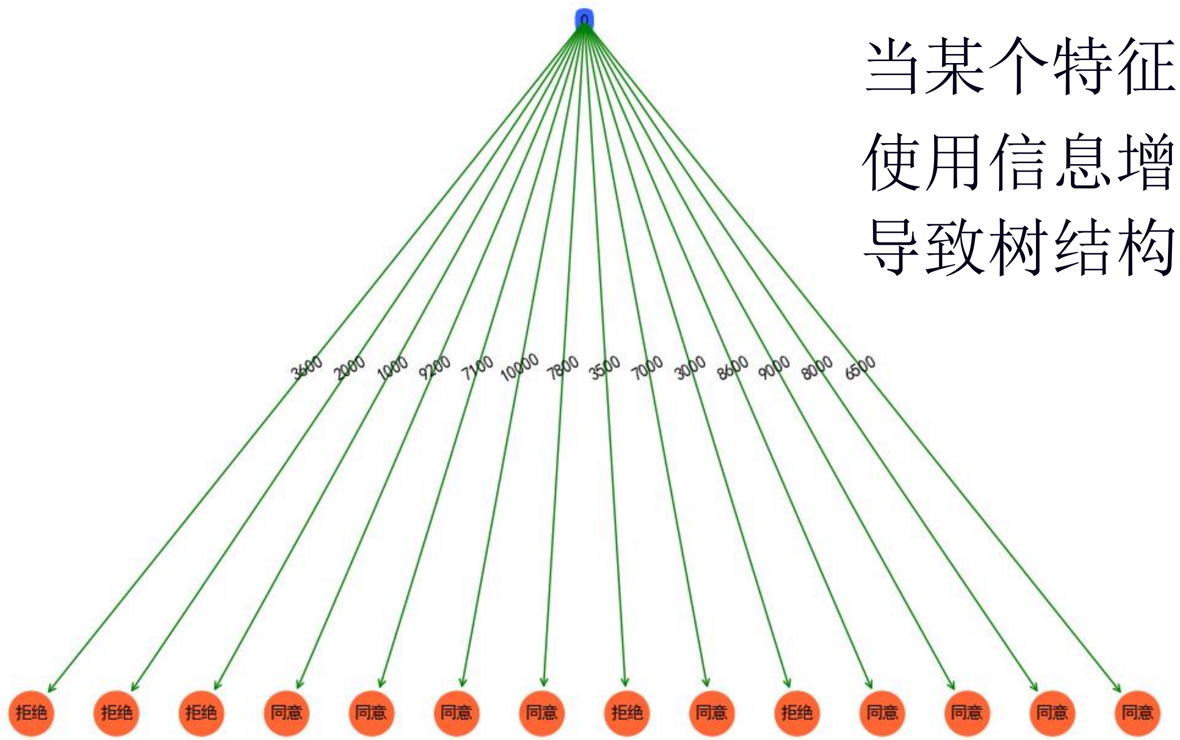
本章授课内容

河北师范大学软件学院
Software College of Hebei Normal University

1. ID3算法的不足
2. 特征熵与相对信息增益
3. C4.5算法建树与预测
4. 离散化与二分方法
5. 实例讲解

1.1 ID3算法的不足

假设现有的数据样本多出来一维特征：
表格贷款数据.xlsx, 工资(分类多)
此时利用ID3算法进行建树会有什么样的效果？



当某个特征的分类情况过多时，
使用信息增益会优先考虑该特征，
导致树结构**扁胖**

2 特征熵与相对信息增益

减小因为特征空间长度对划分造成的影响：使用相对信息增益

信息增益(绝对信息增益)$:\mathrm{IG}(x^i)=E(D)-E(D\mid x^i)$

相对信息增益：$\mathrm{IG\_Ratio}(x^i)=\dfrac{E(D)-E(D\mid x^i)}{E(x^i)}$

特征熵 $E(x^i)$:基于第i维特征的信息熵

$$
\begin{array}{ccc|c}
x^{(1)} & x^{(2)} & x^{(3)} & y
\end{array}
$$

$$
\mathrm{D}=\left|\begin{array}{ccc|c}
x_1^{(1)} & x_1^{(2)} & x_1^{(3)} & y^1 \\
x_2^{(1)} & x_2^{(2)} & x_2^{(3)} & y^2 \\
x_3^{(1)} & x_3^{(2)} & x_3^{(3)} & y^3 \\
x_4^{(1)} & x_4^{(2)} & x_4^{(3)} & y^4
\end{array}\right|
$$

假设 $\left|\begin{array}{l} x_1^{(1)}=\alpha_1 \\ x_2^{(1)}=\alpha_2 \\ x_3^{(1)}=\alpha_2 \\ x_4^{(1)}=\alpha_2 \end{array}\right|$

$$E(x^i)=-\frac{1}{4}\log_2(\frac{1}{4})-\frac{3}{4}\log_2(\frac{3}{4})$$

2 特征熵与相对信息增益的实现：

信息熵(经验熵)的实现:calcShannonEnt

第i维特征的特征熵：calcShannonEnt(data,col=i)

相对信息增益的实现:getBestSplit

第一层循环:(对i维度)

　　第二层循环:(对特征空间)

　　　　……通过叠加求得条件熵currentEnt

　　IG=baseEnt-currentEnt

　　……求bestIG和bestIndex

返回 bestIndex

xiEnt=calcShannonEnt(data,col=i)

IG_Ratio=(baseEnt-currentEnt)/xiEnt

3 C4.5算法的实现

创建tree.py

1. calcShannonEnt(data , col=-1) #计算总体的信息熵
2. splitData(data,index,value) #对样本集按某个特征及其取值划分后的子样本集
3. getBestSplit(data) #获得信息增益最大的特征
4. toLeafNode(labelList) #将决策点变成叶子节点
5. createTree(data)   #建树
6. predict(tree , example) #对单个样本进行类别预测

4 连续特征的离散化

对于连续型特征如何处理：

连续特征的离散化就是采取各种方法将连续的区间划分为小的区间，并将这些连续的小区间与离散的值关联起来

连续特性离散化的问题本质：决定选择多少个分割点和确定分割点的位置

最常见的，最简单的：等步长离散

离散化的优点：

减少连续特征空间的长度

离散化的数据更易于理解，使用和解释

很多算法不适用与连续型特征

可以也有效的克服数据中的隐藏缺陷(极端值)，使模型更加稳定

4.1 离散化处理的一般过程

1. 对连续型特征进行排序
2. 初步确定连续特征的划分断点

---

3. 按照给定的判断标准继续分割断点或合并断点
4. 如果第三步得到判定标准的终止条件，则终止整个连续特征离散化的过程，否则继续按第三步执行

- 离散化方法
  - 无监督
    - 等步长离散化方法
    - 等频率离散化方法
    - 聚类离散化方法
  - 有监督
    - 基于熵的离散化方法

4.2 无监督离散化方法

1. 等步长离散化方法
   将数据均匀的划分成n等份，每份的间距相等
2. 等频率离散化方法
   将数据均匀的划分成n等份，每份中包含的样本数相同

简单但不稳定：
等步长方容易受到异常点的影响
等频避免了上述问题，但会出现相同值被分入不同箱的情况

3. 聚类离散化方法：例如k_means 算法将连续型特征划分成簇

4.3 有监督的离散化方法

基于熵的离散化方法: (标签是类别，离散的)

使用类别标签计算和确定分割点，自顶向下的分裂方法

1. 计算总体的熵
2. 把**每个值**看做分割点，将数据分为两部分，在多种可能的方法中寻找一种产生最小熵的分法
3. 在分成的两个区间中，继续寻找获取最小熵的分法
4. 如果达到用户指定的个数，或者熵减少不是很显著，则停止分裂

例如：example.txt

c				
c			b	
a		c	b	
a	b	b	a	
a	b	a	a	c
a	a	a	a	b
2	5	9	14	19

4. 3 有监督的离散化方法

$\{'a':10,'b':6,'c':4\}$

首先计算总体样本(基于标签)的熵：

$$\text{ent(D)=-(}\frac{10}{20}\log_2(\frac{10}{20})+\frac{6}{20}\log_2(\frac{6}{20})+\frac{4}{20}\log_2(\frac{4}{20}))$$

假设利用(x<5和x≥5)作为标准

将连续特征进行分割

左：$\{'a':4,'c':2\}$

右：$\{'a':6,'b':6,'c':2\}$

c				
c			b	
a		c	b	
a	b	b	a	
a	b	a	a	c
a	a	a	a	b
2	5	9	14	19

4.3 有监督的离散化方法

分割后得到的两部分可以分别求熵，
加权平均求和得条件熵，
最后求得信息增益(遍历完所有的情况，取增益最大的分割值)

$$ent(D\,|\,x<5)=-(\frac{4}{6}\log_2(\frac{4}{6})+\frac{2}{6}\log_2(\frac{2}{6}))=0.918$$

$$ent(D\,|\,x\geq 5)=-(\frac{6}{14}\log_2(\frac{6}{14})+\frac{6}{14}\log_2(\frac{6}{14})+\frac{2}{14}\log_2(\frac{2}{14}))=1.45$$

$$ent(D\,|\,x,5)=\frac{6}{20}\times ent(D\,|\,x<5)+\frac{14}{20}\times ent(D\,|\,x\geq 5)=1.290$$

$$IG(D\,|\,x,5)=ent(D)-ent(D\,|\,x,5)=1.485-1.290$$

4.4 离散化与二类问题

有监督的离散化的一次分裂过程，可以看做是对连续特征空间进行二分类

$$\mathrm{D}_l = \{D \mid x^i < a_1\} \qquad \mathrm{D}_r = \{D \mid x^i \geq a_1\}$$

a_1 称为“二分标准”

此时问题就转化为如何确定最优“二分标准”

二分标准的选取方法

1. 最简单：将当前数据集中出现的每个值作为二分标准备选空间
2. 最常见：定步长：计算量小，不够稳定
3. 最稳定：将当前数据集中出现的相邻两个值的平均值作为二分标准备选空间

4.4 离散化与二类问题

若 x^i 在当前数据集中有 m 个取值，不妨设为
并且满足 $\mu_1 < \cdots < \mu_m$（若不然，进行一次排序即可），
那么依次选择 $p_1 \cdots p_v$ 作为二分标准并选出最好的

1. 依次选择 $p_1 = \mu_1 \cdots p_m = \mu_m$ 作为二分标准
2. $p_1 \cdots p_v$ 构成等差数列，并满足：$p_1 = \mu_1$ $\quad p_v = \mu_m$
3. 依次选择 $p_1 = \dfrac{\mu_1 + \mu_2}{2} \cdots p_{m-1} = \dfrac{\mu_{m-1} + \mu_m}{2}$ 作为二分标准

当二类问题与决策树结合起来，那对于连续型特征，决策节点的选取就变成：确定特征与最优“二分标准”；
每次分割将数据集分为：左子集和右子集两部分

4.4 离散化与二类问题

相应的可以使用二类问题处理离散型的特征，例如：

$$x^i \in \{a_1, \cdots, a_m\};$$

$$D_l = \{D \mid x^i \boxed{=} a_j\}, D_r = \{D \mid x^i \boxed{\neq} a_j\}; j = 1, \cdots, m$$

如何将离散型的二分类修改成与连续型二分类相同的形式：
独热编码：

从一个特征变为m个特征

x^i	$x^{i\_a_1}, x^{i\_a_2}, \cdots, x^{i\_a_m}$
$x_1^i = a_m$	$0 \quad 0 \quad \cdots \quad 1$
$\vdots \quad \in \{a_1, \cdots, a_m\}$	$\vdots$
$x_n^i = a_2$	$0 \quad 1 \quad \cdots \quad 0$

经过独热处理后，离散特征可以更方便的进行二分类