石家莊鐵道大學
SHIJIAZHUANG TIEDAO UNIVERSITY

在线开放课程

MATLAB在科学研究中的应用

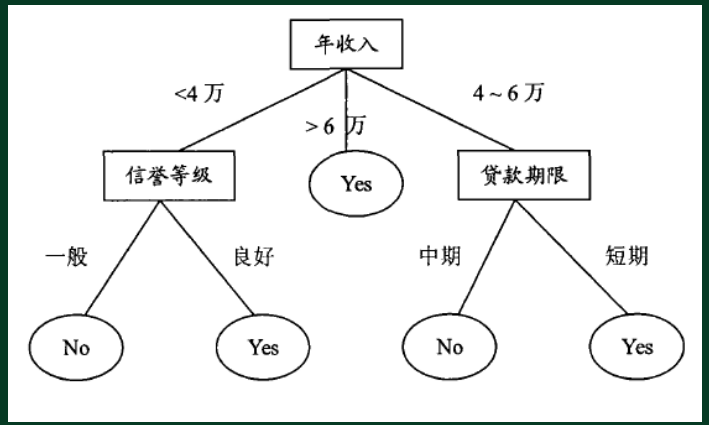
决策树

主讲 ： 卞建鹏

1、决策树概念

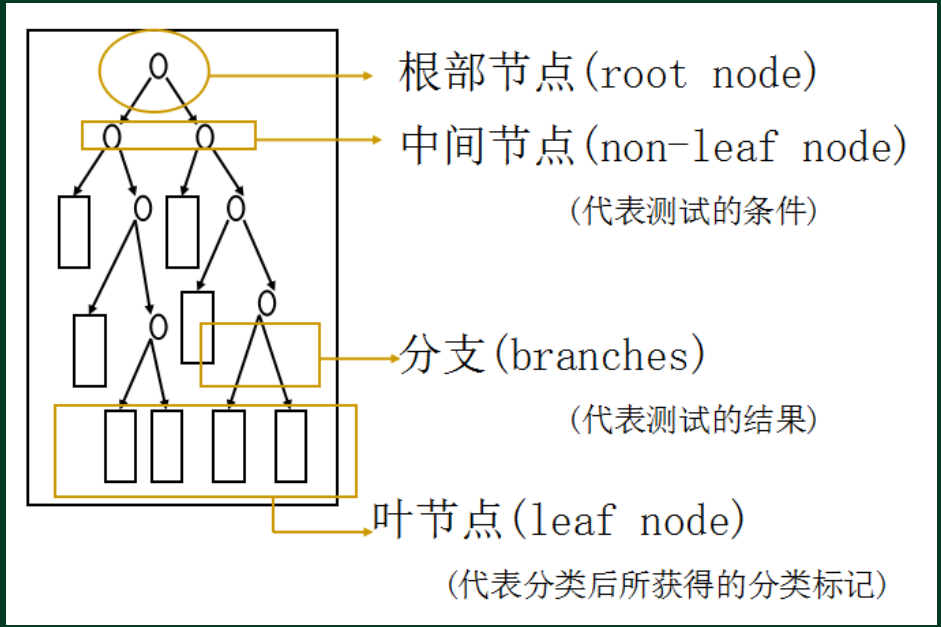
一种描述概念空间的有效归纳推理方法。基于决策树的学习方法可以进行不相关的多概念学习，具有简单快捷，已经在各个领域取得广泛应用。

客户年收入5万，信誉等级一般，申请贷款期限为短期，根据决策树判断是否应该为其发放贷款。



1、决策树概念

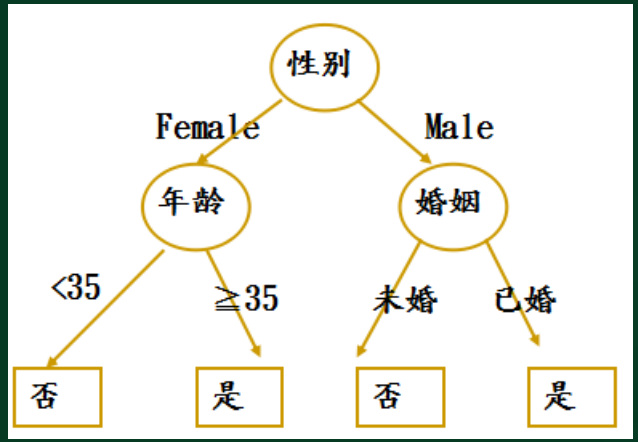
每个内部结构点表示在一个属性上的测试，每个分支表示一个测试输出，每个叶节点代表一个类别。



1、决策树概念

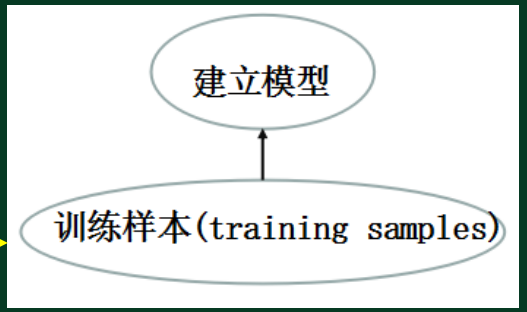
編號	性別	年齡	婚姻	家庭人數	購買RV房車
A0001	Male	45	未婚	1	是
A0002	Male	52	已婚	7	是
A0003	Female	38	已婚	5	是
A0004	Male	25	已婚	5	否
A0005	Female	48	已婚	4	是
A0006	Male	32	未婚	3	是
A0007	Female	65	已婚	4	否
A0008	Male	33	已婚	3	是
A0009	Male	45	已婚	4	是
A0010	Female	52	未婚	1	是
A0011	Male	38	未婚	1	否
…	…	…	…	…	…
Z0099	Male	22	未婚	4	是

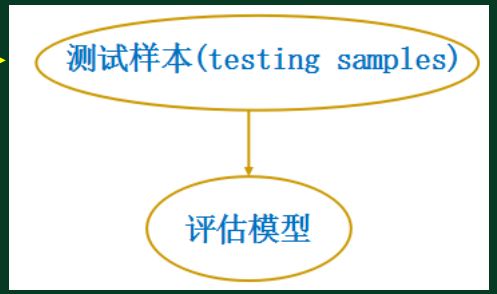
分类标记



1、决策树概念

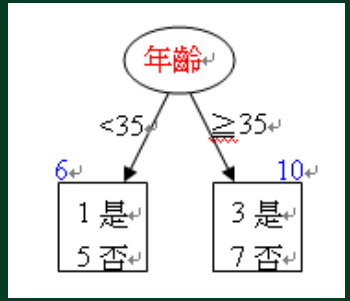
编号	性别	年龄	婚姻	家庭人数	购买RV房车
A0001	Male	45	未婚	1	是
A0002	Male	52	已婚	7	是
A0003	Female	38	已婚	5	是
A0004	Male	25	已婚	5	否
A0005	Female	48	已婚	4	是
A0006	Male	32	未婚	3	是
A0007	Female	65	已婚	4	否
A0008	Male	33	已婚	3	是
A0009	Male	45	已婚	4	是
A0010	Female	52	未婚	1	是
A0011	Male	38	未婚	1	否
…	…	…	…	…	…
Z0099	Male	22	未婚	4	是





1、决策树概念

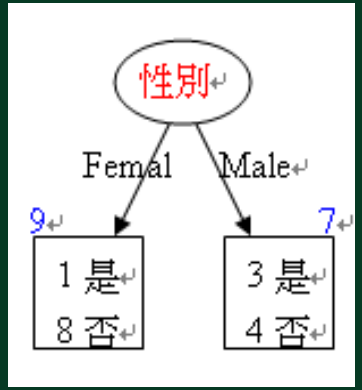
年龄	性别	家庭所得	購買RV房車
<35	Male	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	低所得	否
<35	Male	高所得	否
≧35	Female	低所得	否
<35	Female	低所得	否
<35	Female	高所得	是
≧35	Male	小康	是
<35	Male	高所得	否
≧35	Female	小康	否
<35	Male	低所得	否
≧35	Female	小康	否
≧35	Male	低所得	是
≧35	Male	小康	是
≧35	Female	低所得	否

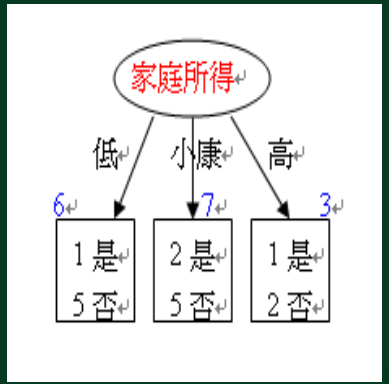


根部节点

中间节点

停止分支





2、ID3算法

Quinlan(1979)提出，以Shannon(1949)的信息论为依据。

ID3算法的属性选择度量就是使用信息增益，选择最高信息增益的属性作为决策树当前节点的测试属性。

信息论：若一事件有k种结果，对应的概率为P_i。则此事件发生后所得到的信息量I(Entropy)为：

$$I=-(p_1*log_2(p_1)+ p_2*log_2(p_2)+\ldots+ p_k*log_2(p_k))$$

设 $k=4 \rightarrow p_1=0.25, p_2=0.25, p_3=0.25, p_4=0.25$

$$I=-(.25*log_2(.25)*4)=2$$

设U为u个元组的集合，类别属性中的分类有m个，设u_i是分别属于这m个类的样本数，$\frac{u_i}{u}$是U中样本属于该分类的概率的估计值，那么对于这个给定的样本分类的信息熵是

$$I(u_1,u_2,\ldots,u_m)=-\sum_{i=1}^{m}\frac{u_i}{u}\log_2\frac{u_i}{u}$$

具有值域$\{a_1,a_2,\ldots,a_v\}$的属性A可以用来将U划分为子集$\{U_1,U_2,\ldots,U_v\}$，其中，U_j包含U中A值为a_j的那些样本，设U_j包含第i类给定样本分类的u_{ij}个样本。则根据A划分的期望信息称作A的熵为

$$E(A)=\sum_{j=1}^{v}\frac{u_{1j}+\ldots+u_{mj}}{u}I(u_{1j},\ldots,u_{mj})$$

根据A进行的划分获得的信息增益为

$$Gain(A)=I(u_1,u_2\ldots,u_m)-E(A)$$

年齡	性別	家庭所得	購買RV房車
<35	Male	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	低所得	否
<35	Male	高所得	否
≧35	Female	低所得	否
<35	Female	低所得	否
<35	Female	高所得	是
≧35	Male	小康	是
<35	Male	高所得	否
≧35	Female	小康	否
<35	Male	低所得	否
≧35	Female	小康	否
≧35	Male	低所得	是
≧35	Male	小康	是
≧35	Female	低所得	否

n=16

n_1=4 是

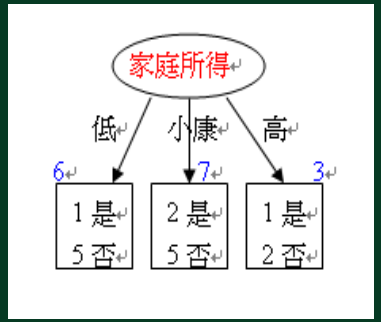
n_2=12 否

I(16,4)= - ((4/16)*$\log_2$(4/16)+(12/16)*$\log_2$(12/16))=0.8113

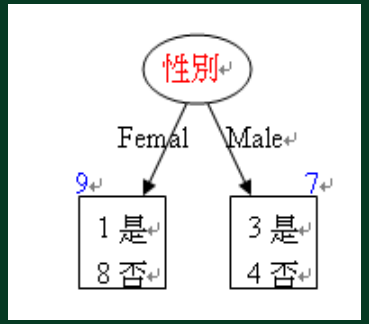
E(年龄)=(6/16)*I(6,1)+(10/16)*I(10,3)=0.7946

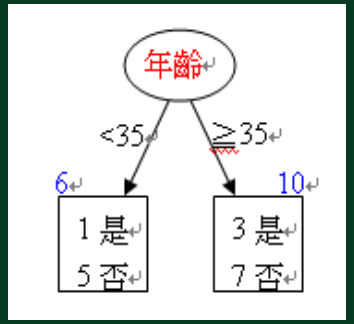
Gain(年龄)=I(16,4)-E(年龄)=0.0167

年齡	性別	家庭所得	購買RV房車
<35	Male	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	低所得	否
<35	Male	高所得	否
≧35	Female	低所得	否
<35	Female	低所得	否
<35	Female	高所得	是
≧35	Male	小康	是
<35	Male	高所得	否
≧35	Female	小康	否
<35	Male	低所得	否
≧35	Female	小康	否
≧35	Male	低所得	是
≧35	Male	小康	是
≧35	Female	低所得	否



Gain(家庭所得)=0.0177

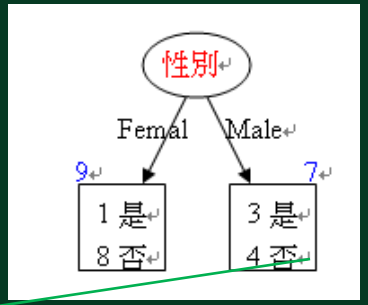




Gain(性別)=0.0972　Gain(年齡)=0.0167

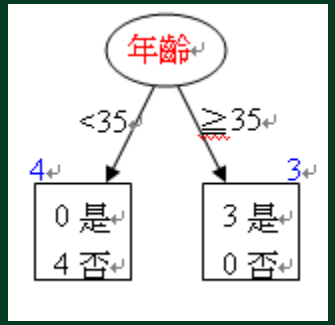
Max:作为第一个分类依据

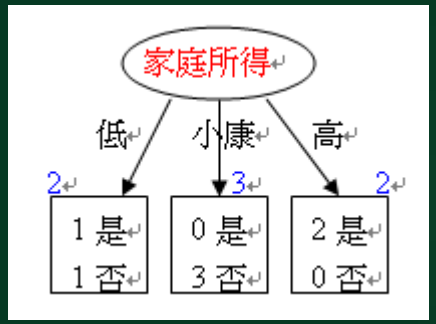
年齡	性別	家庭所得	購買RV房車
<35	Male	小康	否
<35	Male	低所得	否
<35	Male	高所得	否
<35	Male	高所得	否
≧35	Male	小康	是
≧35	Male	小康	是
≧35	Male	低所得	是



I(7,3)=-((3/7)*log2(3/7)+(4/7)*log2(4/7))=0.9852

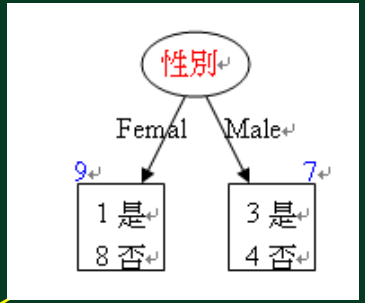
年齡	性別	家庭所得	購買RV房車
<35	Female	低所得	否
<35	Female	高所得	是
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	低所得	否
≧35	Female	低所得	否
≧35	Female	低所得	否





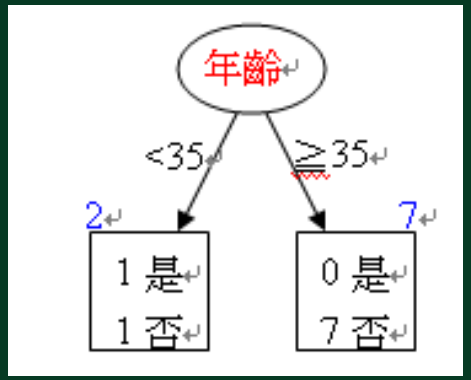
Gain(年齡)=0.9852　Gain(家庭所得)=0.688

年齡	性別	家庭所得	購買RV房車
<35	Male	小康	否
<35	Male	低所得	否
<35	Male	高所得	否
<35	Male	高所得	否
≧35	Male	小康	是
≧35	Male	小康	是
≧35	Male	低所得	是

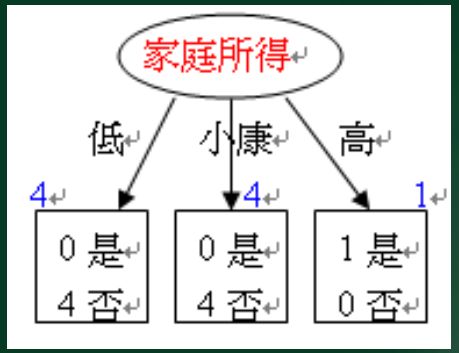


$I(9,1)=-((1/9)*\log_2(1/9)+(8/9)*\log_2(8/9))=0.5032$

年齡	性別	家庭所得	購買RV房車
<35	Female	低所得	否
<35	Female	高所得	是
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	低所得	否
≧35	Female	低所得	否
≧35	Female	低所得	否



Gain(年齡)=0.2222

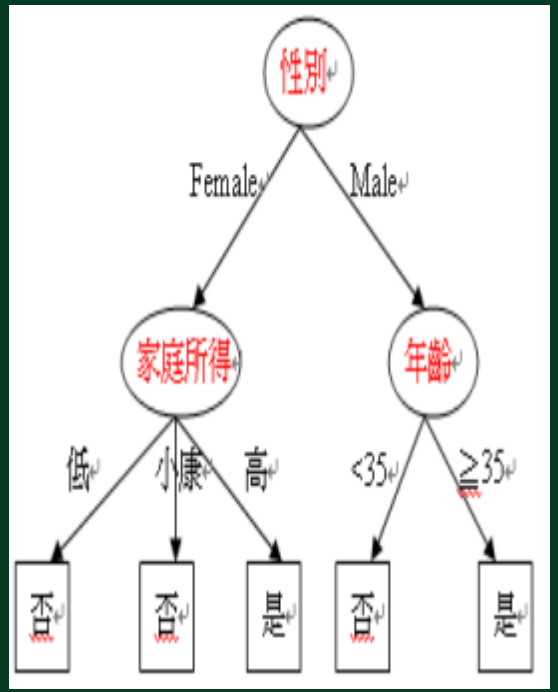


Gain(家庭所得)=0.5032

2、ID3算法

年齡	性別	家庭所得	購買RV房車
<35	Male	小康	否
≧35	Female	小康	否
≧35	Female	小康	否
≧35	Female	低所得	否
<35	Male	高所得	否
≧35	Female	低所得	否
<35	Female	低所得	否
<35	Female	高所得	是
≧35	Male	小康	是
<35	Male	高所得	否
≧35	Female	小康	否
<35	Male	低所得	否
≧35	Female	小康	否
≧35	Male	低所得	是
≧35	Male	小康	是
≧35	Female	低所得	否

决策树



3、MATLAB程序

构造决策树

t = classregtree(X,y)

t = classregtree(X,y,'Name',value)

X是样本特征值，y是样本类别，Name\value是成对出现的可选项。

t是得到的决策树模型。

如果y是确定的数值，得到的是回归树。

如果y是分类变量、字符数组或者字符串数组，得到的就是分类树。

3、MATLAB程序

画出决策树

view(t)

view(t,param1,val1,param2,val2,...)

决策树剪枝

t2 = prune(t1,'level',level)：剪掉t1中的后level层，0不剪枝，1表示最底层，2表示最深的两层，以此类推。

t2 = prune(t1,'nodes',nodes)：剪掉第nodes个分枝节点后的所有枝，如果nodes不是分枝节点就不会剪枝。

3、MATLAB程序

用决策树进行预测

yfit = eval(t,X)

yfit = eval(t,X,s)

[yfit,nodes] = eval(...)

[yfit,nodes,cnums] = eval(...)

t是决策树模型，X是预测样本，yfit是预测结果。s是剪枝选项，如果s是单个数值，就是几层剪枝。如果s是数值数组，那么返回一个矩阵，yfit(i)就是s(i)层剪枝的结果。nodes返回该样本所处的节点位置。cnums返回预测的类别号，1、2、3等。

4、程序实例

load fisheriris

classregtree(meas,species,'names',{'SL' 'SW' 'PL' 'PW'})

1	5.1000	3.5000	1.4000	0.2000	setosa
2	4.9000	3	1.4000	0.2000	setosa
3	4.7000	3.2000	1.3000	0.2000	setosa
4	4.6000	3.1000	1.5000	0.2000	setosa

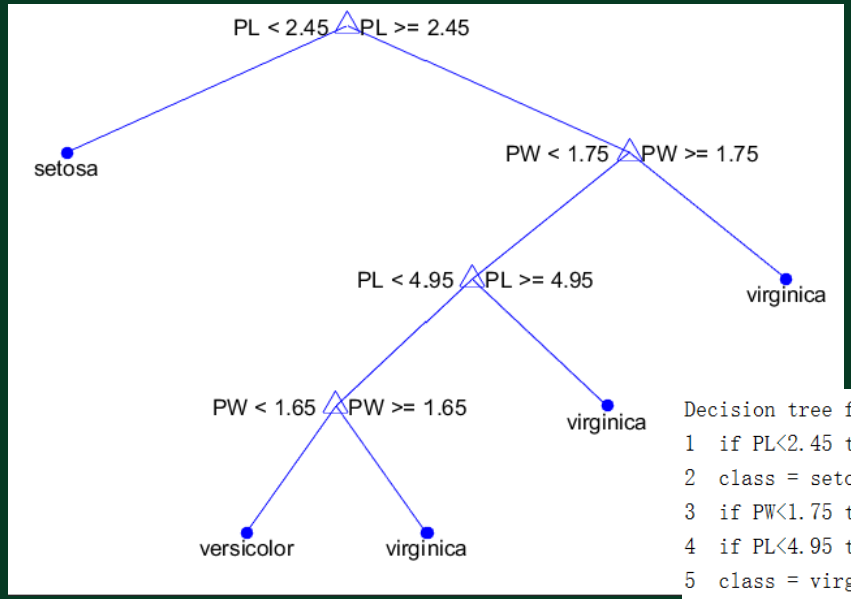
...

51	7	3.2000	4.7000	1.4000	versicolor
52	6.4000	3.2000	4.5000	1.5000	versicolor
53	6.9000	3.1000	4.9000	1.5000	versicolor
54	5.5000	2.3000	4	1.3000	versicolor

...

147	6.3000	2.5000	5	1.9000	virginica
148	6.5000	3	5.2000	2	virginica
149	6.2000	3.4000	5.4000	2.3000	virginica
150	5.9000	3	5.1000	1.8000	virginica

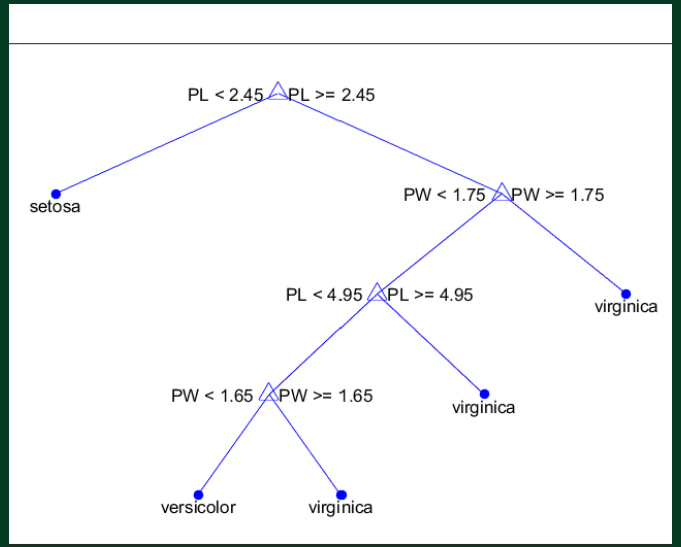
```
load fisheriris
classregtree(meas,species,'names',{'SL' 'SW' 'PL' 'PW'})
view(t)
```

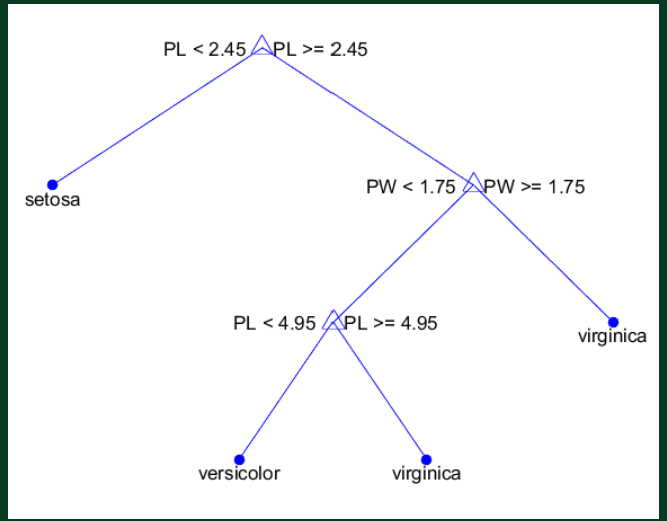



```
Decision tree for classification
1  if PL<2.45 then node 2 elseif PL>=2.45 then node 3 else setosa
2  class = setosa
3  if PW<1.75 then node 4 elseif PW>=1.75 then node 5 else versicolor
4  if PL<4.95 then node 6 elseif PL>=4.95 then node 7 else versicolor
5  class = virginica
6  if PW<1.65 then node 8 elseif PW>=1.65 then node 9 else versicolor
7  class = virginica
8  class = versicolor
9  class = virginica
```

4、程序实例


```
t2 = prune(t,'level',1)        %1表示最底层
t2.view
[yfit,nodes,cnums]=eval(t,meas);
```





小结

1. 决策树概念
2. ID3算法
3. 实例分析