大数据科学与应用技术

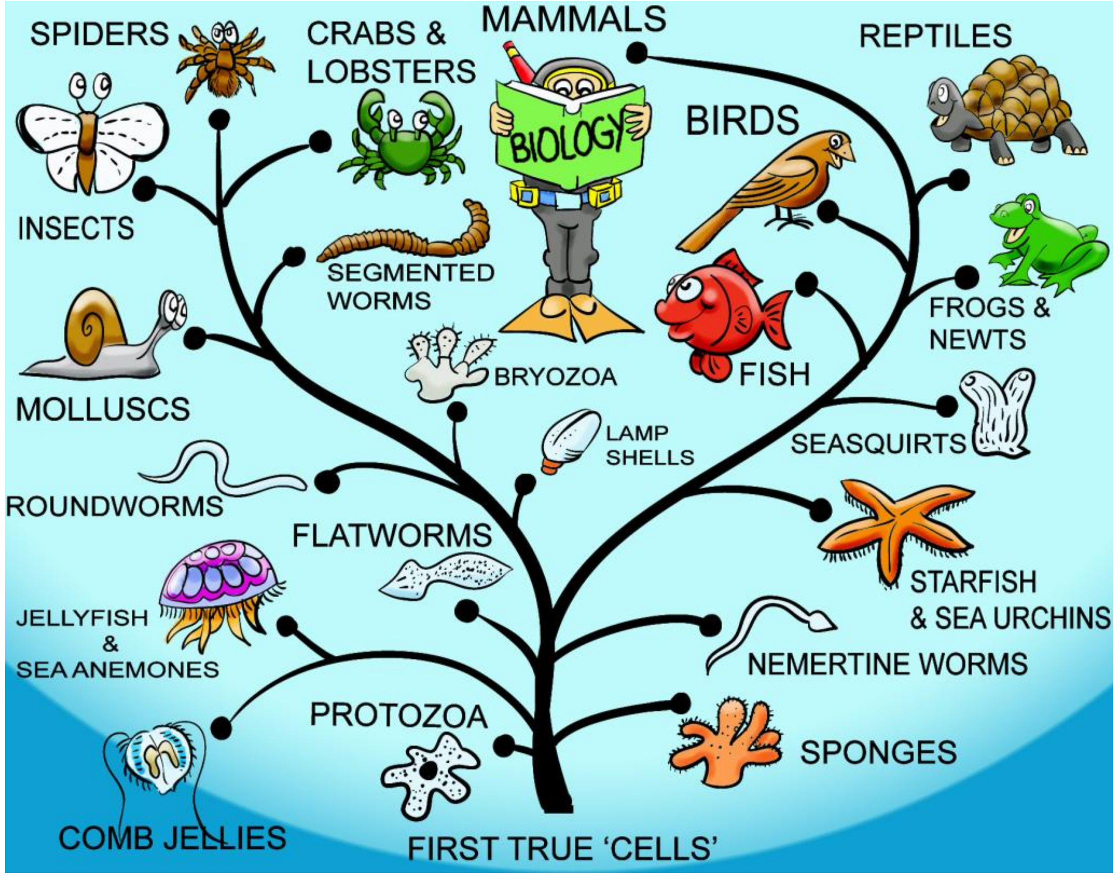
Lecturer：焦在滨（ Zaibin JIAO ）

Email：jiaozaibin@mail.xjtu.edu.cn

Overview

- Naïve Bayes Classifier
- Decision Tree Model

Thomas Bayes

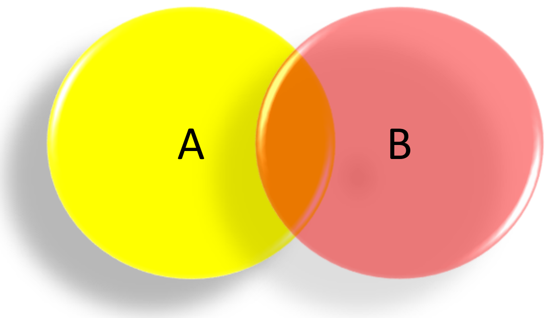


Evolution Tree

Bayes Theorem

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$



Likelihood of evidence B if A is true

Prior probability

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Posterior probability of A given the evidence B

Prior probability that evidence B is true

Fish Example

- Salmon vs. Tuna
- Grab a fish at random.
- $P(\omega_1)=P(\omega_2)$
- $P(\omega_1)>P(\omega_2)$
- Additional information

$$P(\omega_i \mid x) = \frac{P(x \mid \omega_i)P(\omega_i)}{P(x)}$$

Shooting Example

- Probability of Kill
  - P(A): 0.6
  - P(B): 0.5
- The target is killed with:
  - One shoot from A
  - One shoot from B
- What is the probability that it is shot down by A?
  - C: The target is killed.

$$P(A \mid C) = \frac{P(C \mid A)P(A)}{P(C)} = \frac{1 \times 0.6}{0.6 \times 0.5 + 0.4 \times 0.5 + 0.6 \times 0.5} = \frac{3}{4}$$

Cancer Example

- ω_1: Cancer; ω_2: Normal
- $P(\omega_1)=0.008$; $P(\omega_2)=0.992$
- Lab Test Outcomes: + vs. –
- $P(+|\omega_1)=0.98$; $P(-|\omega_1)=0.02$
- $P(+|\omega_2)=0.03$; $P(-|\omega_2)=0.97$
- Now someone has a positive test result…
- Is he/she doomed?

Cancer Example

$$P(\omega_1 | +) \propto P(+ | \omega_1) P(\omega_1) = 0.98 \times 0.008 = 0.0078$$

$$P(\omega_2 | +) \propto P(+ | \omega_2) P(\omega_2) = 0.03 \times 0.992 = 0.0298$$

$$P(\omega_1 | +) < P(\omega_2 | +)$$

$$P(\omega_1 | +) = \frac{0.0078}{0.0078 + 0.0298} = 0.21 >> P(\omega_1)$$

Headache & Flu Example

- H="Having a Fever"
- F="Coming down with COVID-19"
- P(H)=1/10; P(F)=1/40;   P(H|F)=1/2
- What does this mean?
- One day you wake up with a fever …
- Since 50% COVID-19 cases are associated with fever …
- I must have a 50-50 chance of coming down with COVID-19!

Headache & Flu Example

The truth is …

$$P(F|H) = \frac{P(H|F)P(F)}{P(H)} = \frac{1/2 \times 1/40}{1/10} = \frac{1}{8}$$

Naïve Bayes Classifier

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} P(\omega_i \mid a_1, a_2, \ldots, a_n)$$

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} \frac{P(a_1, a_2, \ldots, a_n \mid \omega_i) P(\omega_i)}{P(a_1, a_2, \ldots, a_n)}$$

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} \underline{P(a_1, a_2, \ldots, a_n \mid \omega_i)} P(\omega_i)$$

Conditionally Independent

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} P(\omega_i) \prod_j P(a_j \mid \omega_i)$$

MAP: Maximum A Posterior

Independence

$$P(A \cap B) = P(A)P(B|A) + P(B|A) = P(B)$$

$$\Downarrow$$

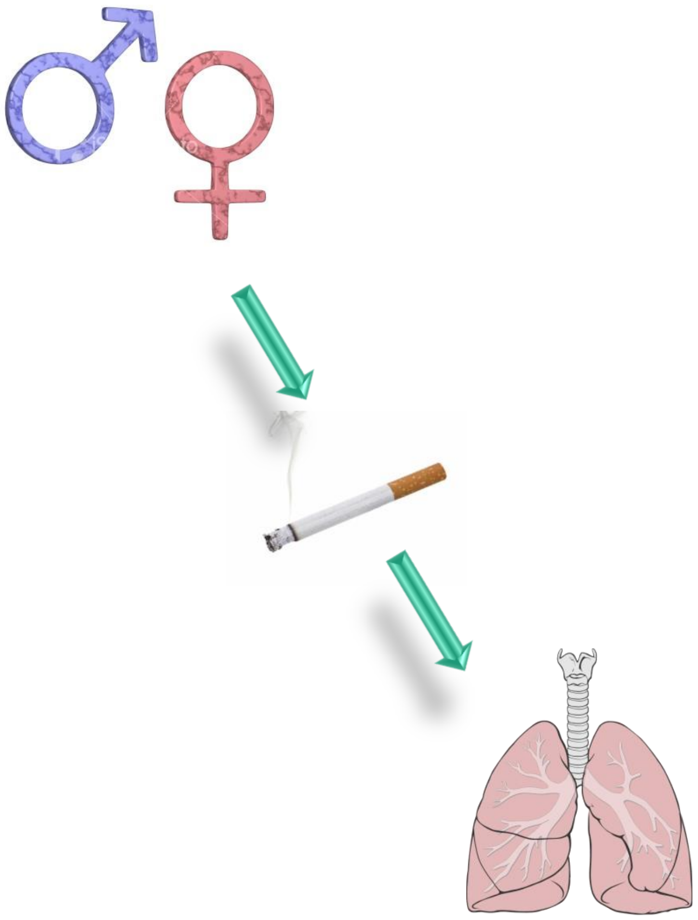
$$P(A \cap B) = P(A)P(B)$$

Conditionally Independent

$$P(A, B|G) = \underline{P(A|G)}P(B|G) \Longleftrightarrow \underline{P(A|G, B)} = P(A|G)$$

$$P(A, B \mid G) = P(A, B, G) / P(G) = P(A \mid B, G) \times P(B, G) / P(G)$$
$$= \underline{P(A \mid B, G)} \times P(B \mid G)$$

Conditional Independence



$P(Cancer|Male) = 65/100{,}000$

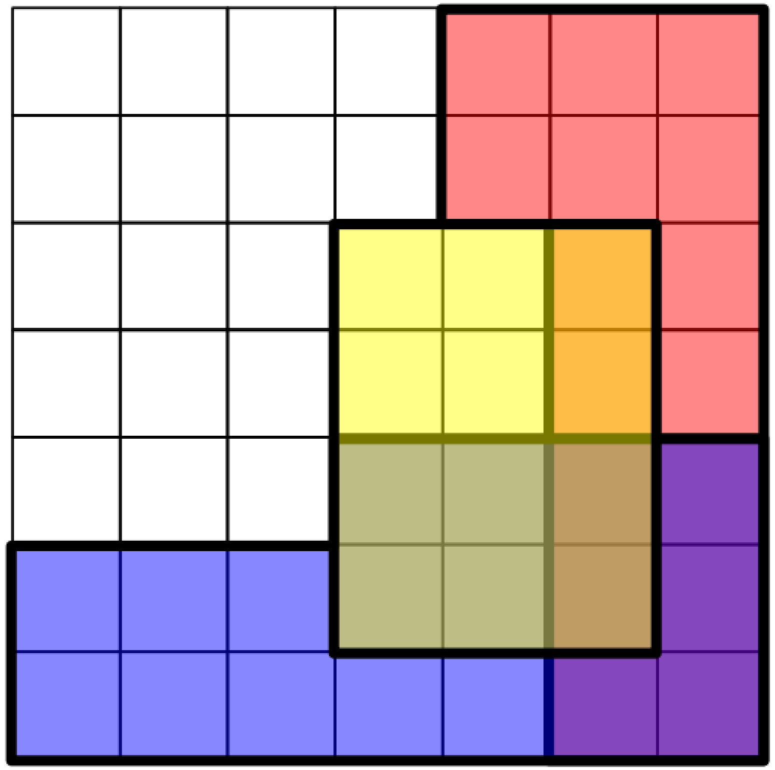
$P(Cancer|Female) = 48/100{,}000$

- Are the two events Male/Female and Cancer independent?
- Assume smoking is the sole contributing factor to cancer.

Conditionally Independent

$$P(Cancer|Male, Smoking) = P(Cancer|Smoking)$$

Conditional Independence



$P(R \cap B) = 6/49$

$P(R) = 16/49$

$P(B) = 18/49$

$P(R \cap B) \neq P(R)P(B)$

Not Independent

$P(R \cap B|Y) = 1/6$

$P(R|Y) = 1/3$

$P(B|Y) = 1/2$

$P(R \cap B|Y) = P(R|Y)P(B|Y)$

Conditionally Independent

Conditional Independence

- Two coins: fair vs. biased (two-headed)
- Select one coin at random and toss twice.
- A: First coin toss is head.
- B: Second coin toss is head.
- C: You selected the fair coin.

$$P(A) = P(B) = 0.5 \times 0.5 + 0.5 \times 1.0 = 0.75$$

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C)+P(A|\neg C)P(\neg C)} = \frac{0.5\times0.5}{0.5\times0.5+1\times0.5} = \frac{1}{3}$$

$$P(B|A) = \frac{1}{3} \times 0.5 + \frac{2}{3} \times 1.0 = \frac{5}{6} \neq P(B)$$ Not Independent

$$\underline{P(B|A,C) = P(B|C) = 0.5}$$ Conditionally Independent

Independent ≠ Uncorrelated

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X \sigma_Y}$$

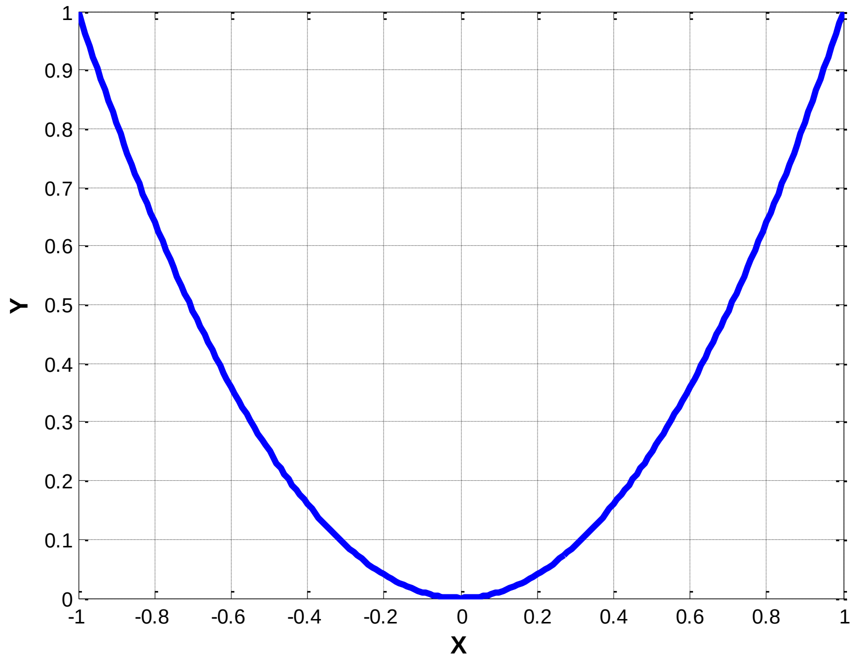
$$X \in [-1,\ 1]$$

$$Y = X^2$$

X	Y
1	1
0.5	0.25
0.2	0.04
0	0
-0.2	0.04
-0.5	0.25
-1	1

Cov (X,Y)=0 → X and Y are uncorrelated.

However, Y is completely determined by X.



Estimating $P(\alpha_j|\omega_i)$

α_1	α_2	α_3	ω
	+		ω_1
			ω_2
	-		ω_1
	+		ω_1
			ω_2

$$P(\omega_1) = 3/5; \qquad P(\omega_2) = 2/5$$

$$P(a_2 = '+' | \omega_1) = 2/3$$

$$P(a_2 = '-' | \omega_1) = 1/3$$

Laplace Smoothing

$$P(a_{jk} | \omega_i) = \frac{|a_j = a_{jk} \wedge \omega = \omega_i| + 1}{|\omega = \omega_i| + |a_j|}$$

How about continuous variables?

Tennis Example

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Tennis Example

Given :

$< \text{Outlook} = sunny, \text{Temperature} = cool, \text{Humidity} = high, \text{Wind} = strong >$

Predict :

$PlayTennis\ (yes\ or\ no)$

$Bayes\ Solution:$

$P(PlayTennis = yes) = 9/14$

$P(PlayTennis = no) = 5/14$

$P(Wind = strong \mid PlayTennis = yes) = 3/9$

$P(Wind = strong \mid PlayTennis = no) = 3/5$

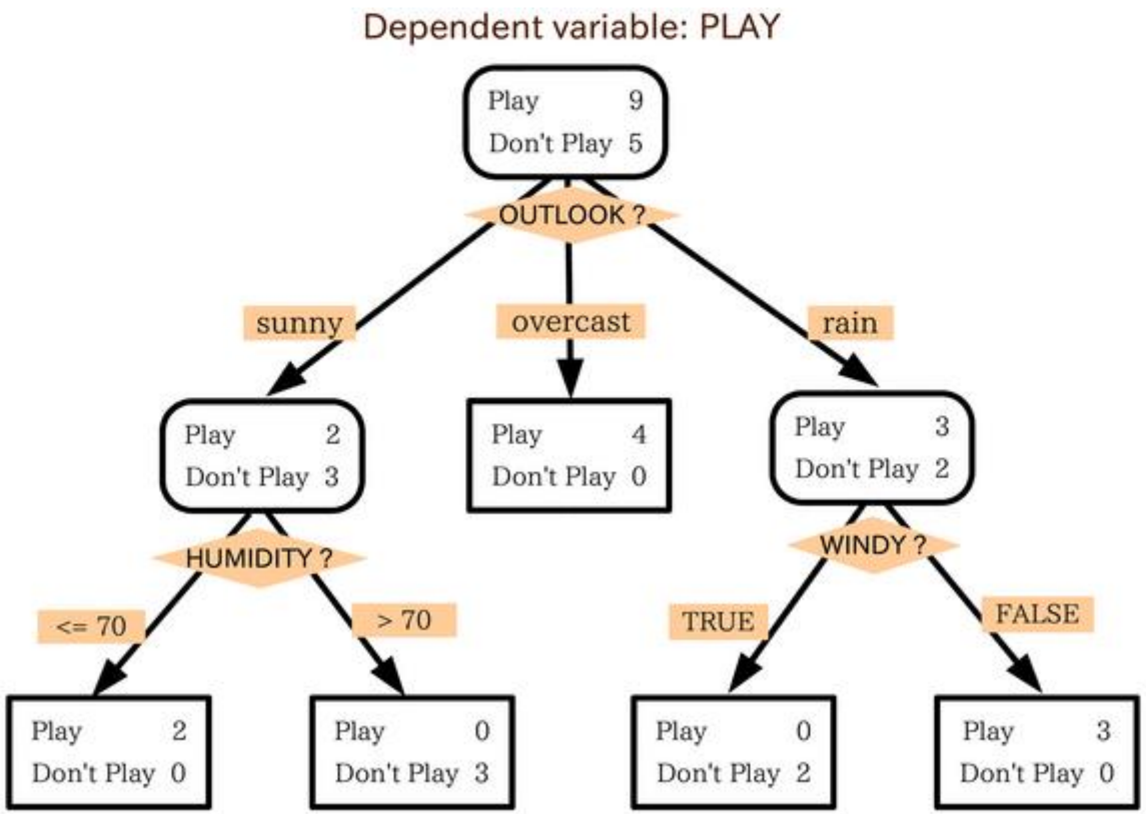
...

$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$

$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = 0.0206$

The conclusion is not to play tennis with probability: $\frac{0.0206}{0.0206 + 0.0053} = 0.795$

Decision Making



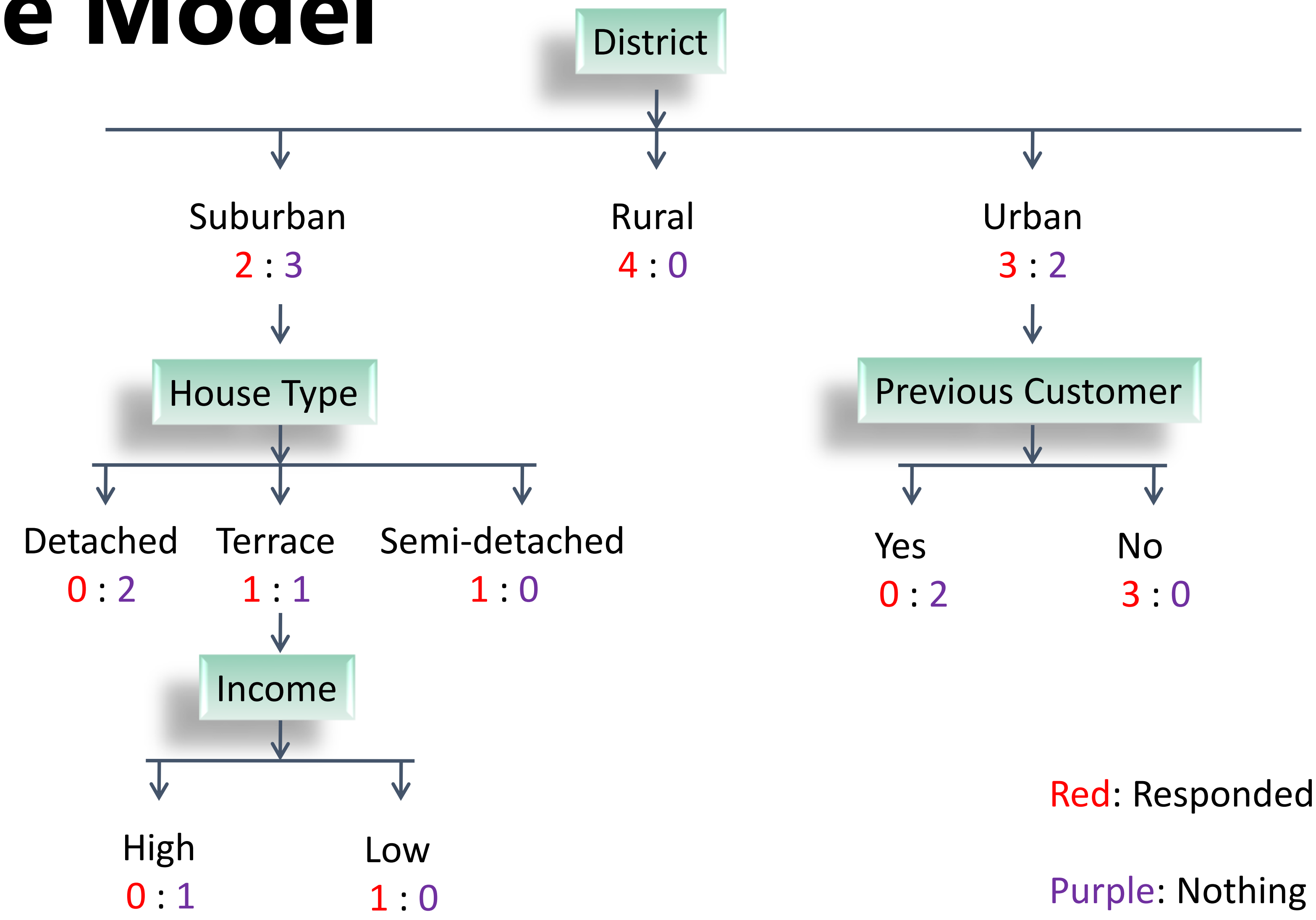
A Survey Dataset

District	House Type	Income	Previous Customer	Outcome
Suburban	Detached	High	No	Nothing
Suburban	Detached	High	Yes	Nothing
Rural	Detached	High	No	Responded
Urban	Semi-detached	High	No	Responded
Urban	Semi-detached	Low	No	Responded
Urban	Semi-detached	Low	Yes	Nothing
Rural	Semi-detached	Low	Yes	Responded
Suburban	Terrace	High	No	Nothing
Suburban	Semi-detached	Low	No	Responded
Urban	Terrace	Low	No	Responded
Suburban	Terrace	Low	Yes	Responded
Rural	Terrace	High	Yes	Responded
Rural	Detached	Low	No	Responded
Urban	Terrace	High	Yes	Nothing

A Survey Dataset

- Given the data collected from a promotion activity.
  - Could be tens of thousands of such records.
- Can we find any interesting patterns?
  - All rural households responded …
- To find out which factors most strongly affect a household’s response to a promotion.
  - Better understanding of potential customers
- Need a classifier to examine the underlying relationships and make future predictions.
- Send promotion brochures to selected households next time.
  - Targeted Marketing

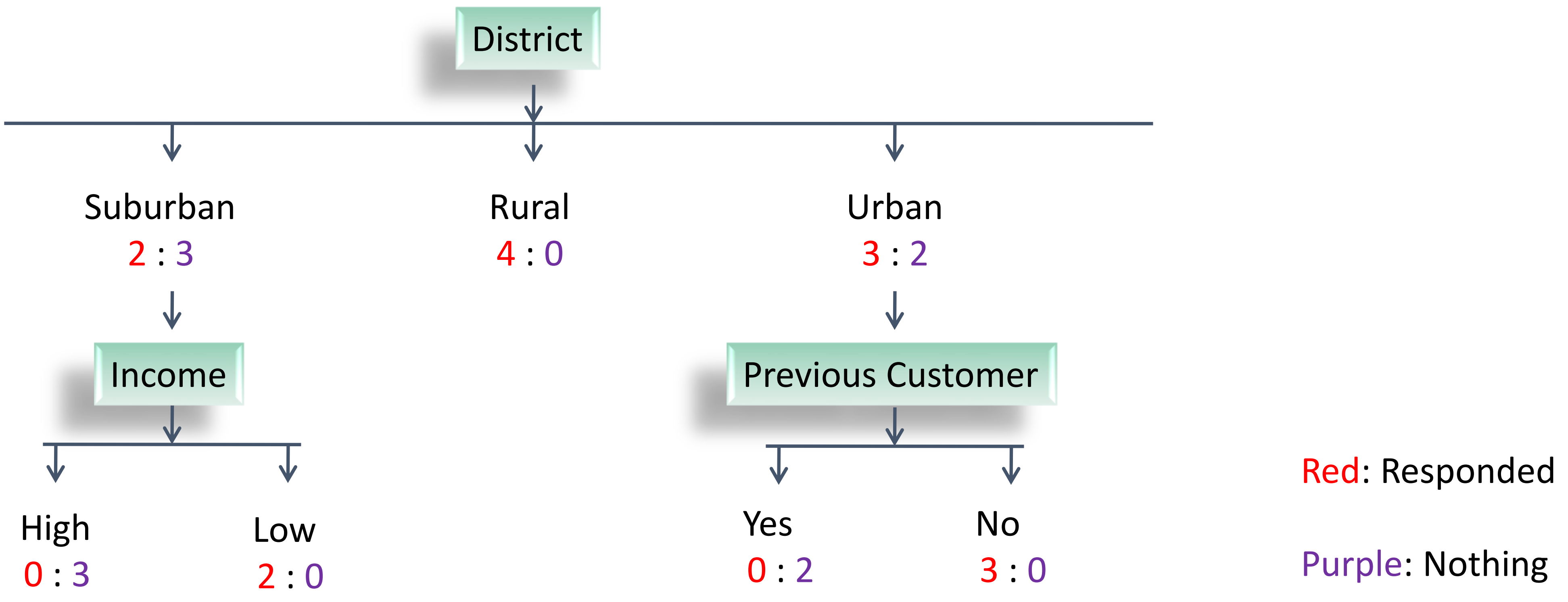
A Tree Model



Red: Responded

Purple: Nothing

Another Tree Model



Some Notes …

- Rules can be easily extracted from the built tree.
  - (District = Rural) → (Outcome = Responded)
  - (District = Urban) AND (Previous Customer = Yes) → (Outcome = Nothing)
- One dataset, many possible trees
- Occam's Razor
  - The term *razor* refers to the act of shaving away unnecessary assumptions to get to the simplest explanation.
  - “When you have two competing theories that make exactly the same predictions, the simpler one is the better.”
  - “The explanation of any phenomenon should make as few assumptions as possible, eliminating those making no difference in the observable predictions of the explanatory hypothesis or theory.”
- Simpler trees are generally preferred.

ID3

- How to build a shortest tree from a dataset?
- Iterative Dichotomizer 3
- **Ross Quinlan**: http://www.rulequest.com/
- One of the most influential Decision Trees models
- Top-down, greedy search through the space of possible decision trees
- Since we want to construct short trees …
- It is better to put certain attributes higher up the tree.
- Some attributes split the data more purely than others.
- Their values correspond more consistently with the class labels.
- Need to have some sort of measure to compare candidate attributes.

Entropy

$$Entropy(S) = -\sum_{i=1}^{C} p_i \log(p_i)$$

p_i: the proportion of instances in the dataset that take the i^{th} target value

$$S = [9/14\,(responses), 5/14\,(no\ responses)]$$

$$Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in A}\frac{|S_v|}{|S|} Entropy(S_v)$$

S_v: the subset of S where attribute A takes the value v.

Attribute Selection

$$Gain(S, District) = Entropy(S) - \frac{5}{14} Entropy(S_{District=Suburban})$$

$$-\frac{5}{14} Entropy(S_{District=Urban}) - \frac{4}{14} Entropy(S_{District=Rural})$$

$$= 0.940 - \frac{5}{14} \cdot 0.971 - \frac{5}{14} \cdot 0.971 - \frac{4}{14} \cdot 0 = \color{red}{0.247}$$

$$Gain(S, Income) = Entropy(S) - \frac{7}{14} Entropy\left(S_{Income=High}\right)$$

$$-\frac{7}{14} Entropy(S_{Income=Low})$$

$$= 0.940 - \frac{7}{14} \cdot 0.9852 - \frac{7}{14} \cdot 0.5917 = \color{red}{0.152}$$

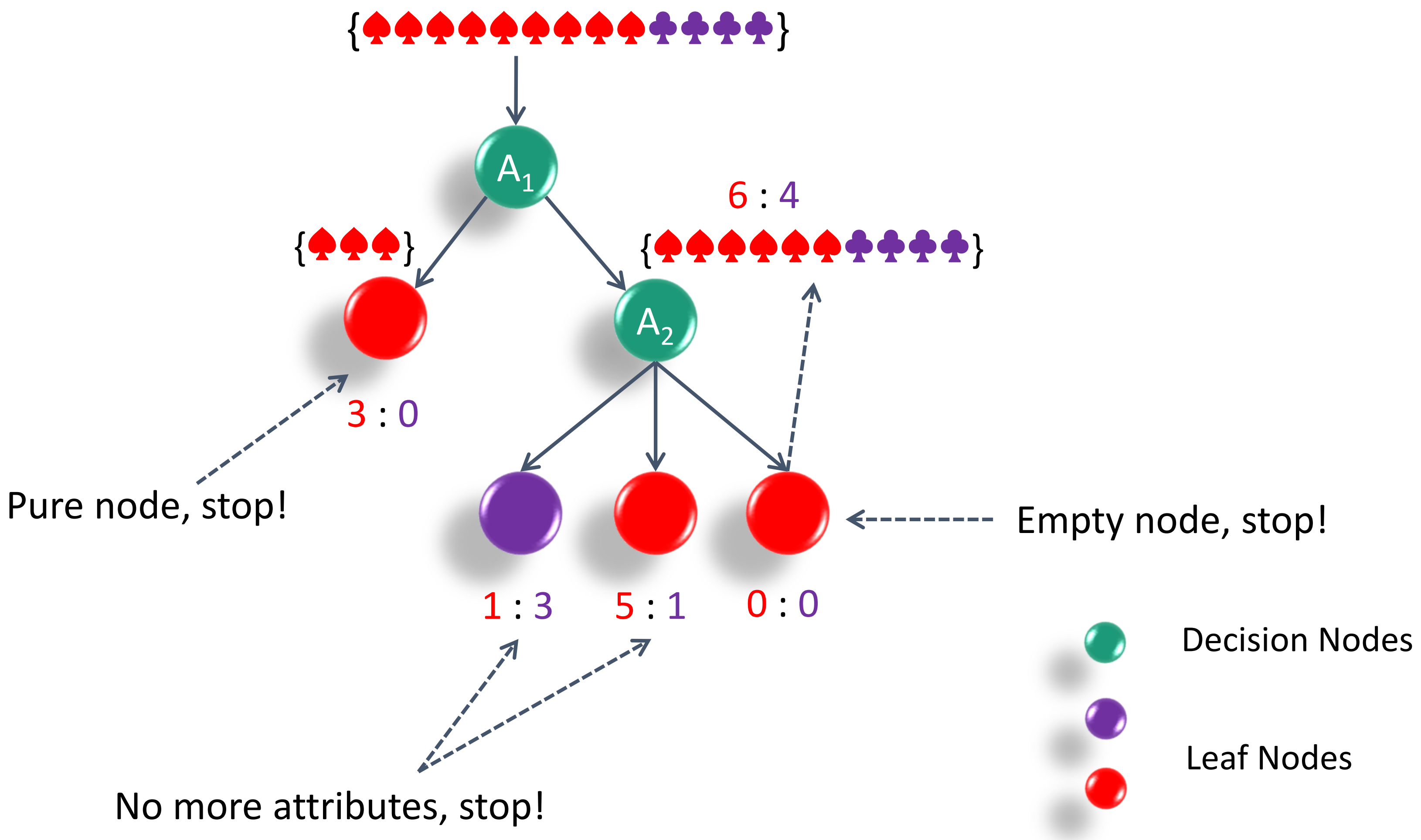
Overfitting

- It is possible to create a separate rule for each training sample.
  - Perfect Training Accuracy vs. Overfitting
  - Random Noise, Insufficient Samples
- We want to capture the general underlying functions or trends.
- Definition
  - Given a hypothesis space H, a hypothesis $h \in H$ is said to overfit the training data if there exists some alternative hypothesis $h' \in H$, such as h has smaller error than h' over the training samples, but h' has a smaller error than h over the entire distribution of instances.
- Solutions
  - Stop growing the tree earlier.
  - Allow the tree to overfit the data and then post-prune the tree.

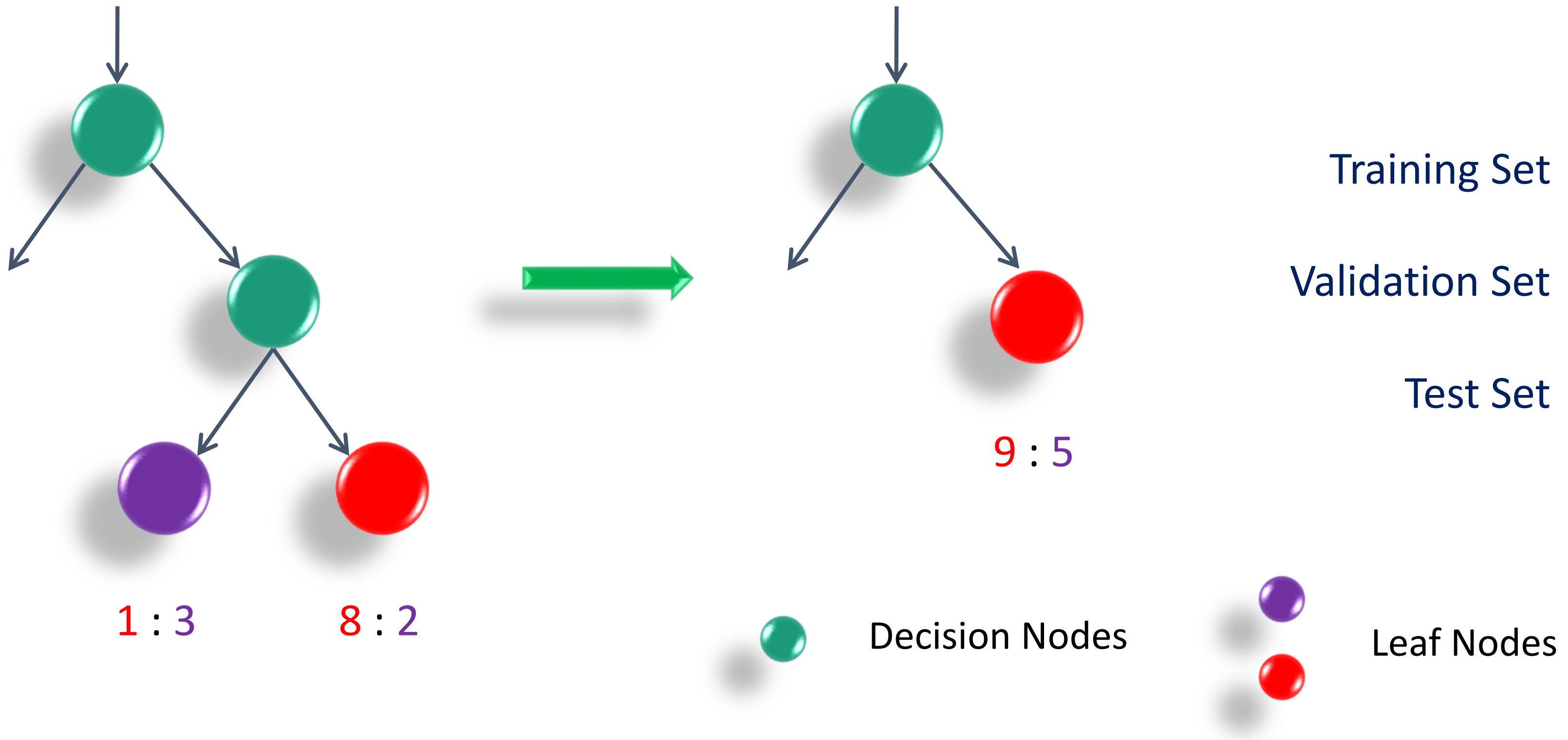
ID3 Framework

- **ID3(Examples, Target_attribute, Attributes)**
- Create a *Root* node for the tree.
- If *Examples* have the same target attribute T, return *Root* with label=T.
- If *Attributes* is empty, return *Root* with label=the most common value of *Target_attribute* in *Examples*.
- A ← the attribute from *Attributes* that best classifies *Examples*.
- The decision attribute for *Root* ← A.
- For each possible value v_i of A
  - Add a new tree branch below *Root*, corresponding to A= v_i.
  - Let *Examples* (v_i) be the subset of Examples that have value v_i for A.
  - If *Examples* (v_i) is empty
    - Below this new branch add a leaf node with label=the most common value of *Target_attribute* in *Examples*.
  - Else below this new branch add the subtree
    - **ID3(Examples(v_i), Target_attribute, Attributes-{A})**
- Return *Root*

ID3 Framework



Pruning（剪枝）



Training Set

Validation Set

Test Set

剪枝(pruning)是决策树算法对付过拟合的主要手段。

一、基于测试集剪枝的基本策略

预剪枝(prepruning)：在决策树生成过程中，对每一个节点在划分前先进行估计，若当前节点的划分不能带来决策树泛化性能的提升，则停止划分，并标记当前节点为叶节点。

后剪枝(post-pruning)：先生成一颗完整的决策树，然后自底向上对非叶节点进行考察，若将该节点对应的子树替换为叶节点能带来决策树泛化性能提升，则将该子树替换为叶节点。

如何判断剪枝泛化后性能是否提升？ 留出法

表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

训练集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	[illegible]	[illegible]	稍凹	软粘	是
10	青绿	硬挺	[illegible]	[illegible]	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

测试集

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	[illegible]	[illegible]	稍凹	硬滑	是
9	乌黑	稍蜷	[illegible]	[illegible]	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

$Ent(D)=1$ **色泽：** $Ent(D_{青绿})=1$； $Ent(D_{乌黑})=0.8113$； $Ent(D_{浅白})=0$； $Gain(D,色泽)=0.2775$

根蒂： $Ent(D_{卷曲})=0.918$； $Ent(D_{稍卷})=1$； $Ent(D_{硬挺})=0$； $Gain(D,根蒂)=0.115$

敲声： $Ent(D_{浊响})=0.918$； $Ent(D_{沉闷})=0.918$； $Ent(D_{清脆})=0$； $Gain(D,敲声)=0.1738$

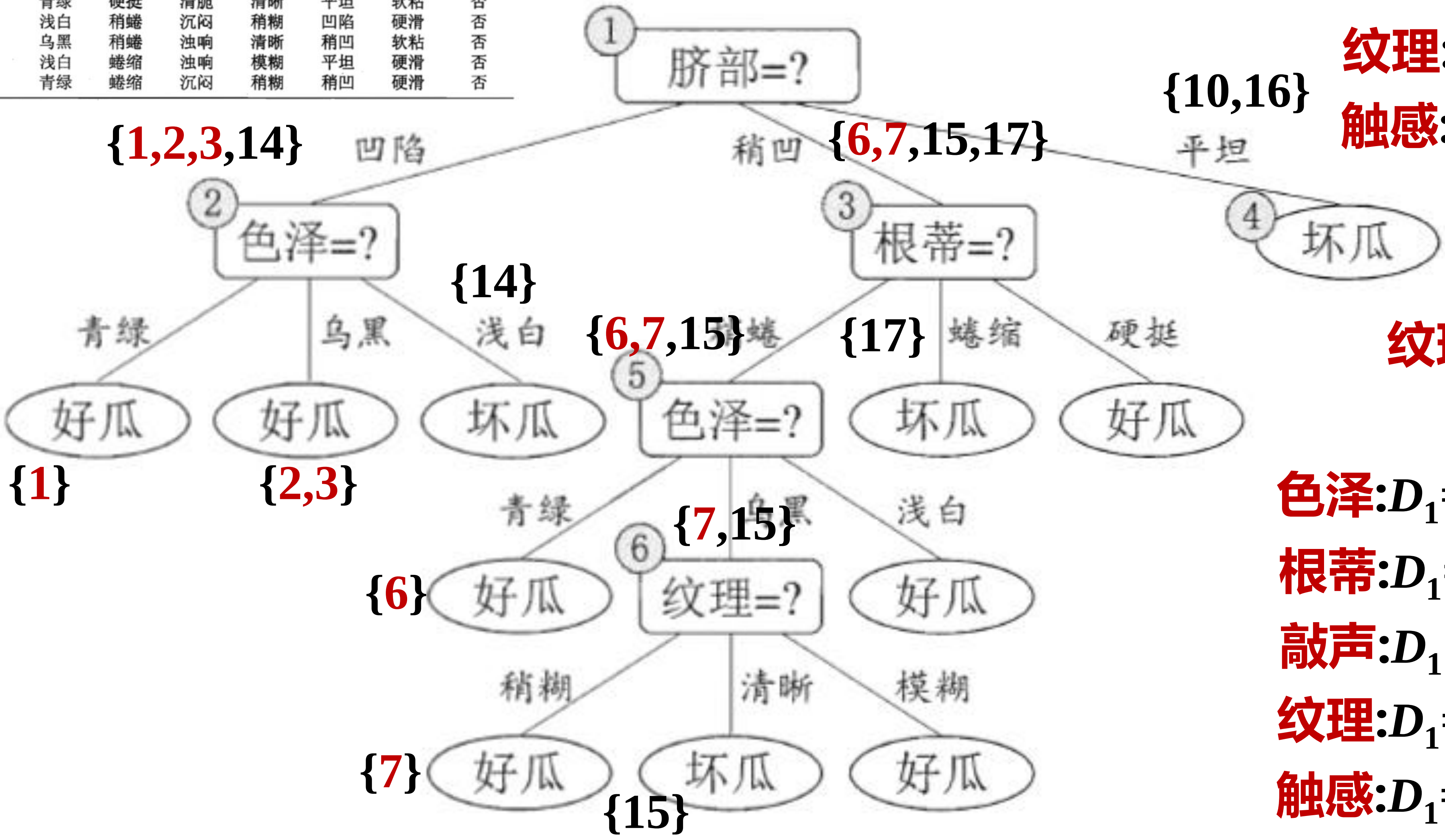
纹理： $Ent(D_{清晰})=0.918$； $Ent(D_{稍糊})=0.918$； $Ent(D_{模糊})=0$； $Gain(D,纹理)=0.1738$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

脐部： $Ent(D_{凹陷})=0.8113$；
$Ent(D_{稍凹})=1$；
$Ent(D_{平坦})=0$；
$Gain(D,脐部)=0.2775$

触感： $Ent(D_{硬滑})=1$；
$Ent(D_{软粘})=1$；
$Gain(D,触感)=0$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否



色泽:$D_1=\{6,\},D_2=\{7,15\}$

敲声:$D_1=\{6,7,15\}$

纹理:$D_1=\{6,15\},D_2=\{7\}$

触感:$D_1=\{6,7,15\}$

敲声:$D_1=\{7,15\}$

纹理:$D_1=\{15\},D_2=\{7\}$

触感:$D_1=\{7,15\}$

色泽:$D_1=\{6,17\},D_2=\{7,15\}$

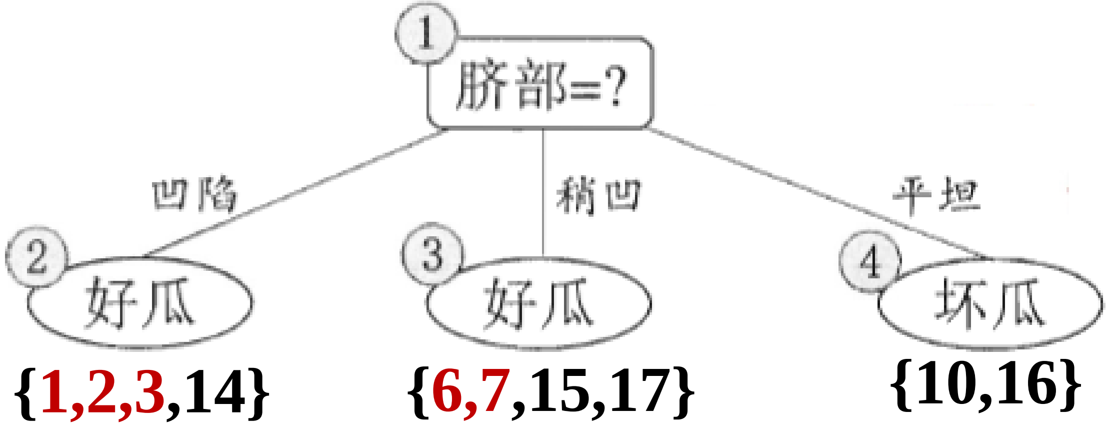
根蒂:$D_1=\{6,7,15\},D_2=\{17\}$

敲声:$D_1=\{6,7,15\},D_2=\{17\}$

纹理:$D_1=\{6,15\},D_2=\{7,17\}$

触感:$D_1=\{6,7,15\},D_2=\{17\}$

二、预剪枝(prepruning)：



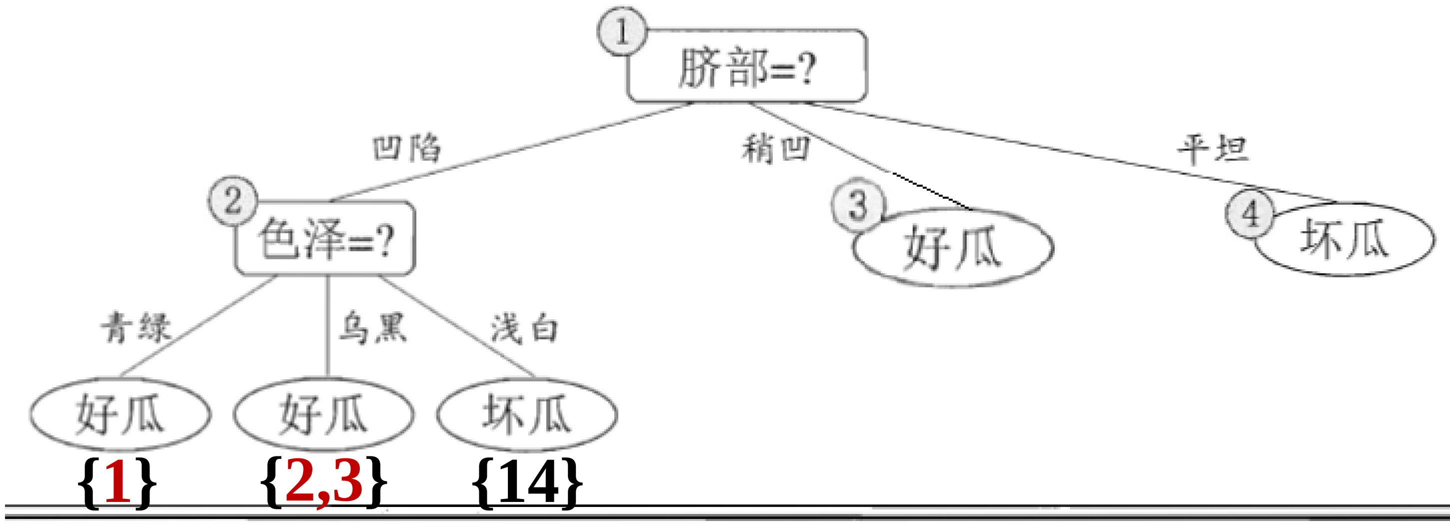
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

对节点①不做划分-叶节点-坏瓜-正确率4/7

对节点做划分-子节点作为叶节点
②-好瓜；③-好瓜；④-坏瓜-正确率

样例：
4：脐部=凹陷-好瓜-正确
5：脐部=凹陷-好瓜-正确
8：脐部=稍凹-好瓜-正确
9：脐部=稍凹-好瓜-错误
11：脐部=平坦-坏瓜-正确
12：脐部=平坦-坏瓜-正确
13：脐部=凹陷-好瓜-错误

划分后正确率：5/7-划分



编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

划分前正确率：5/7-不划分

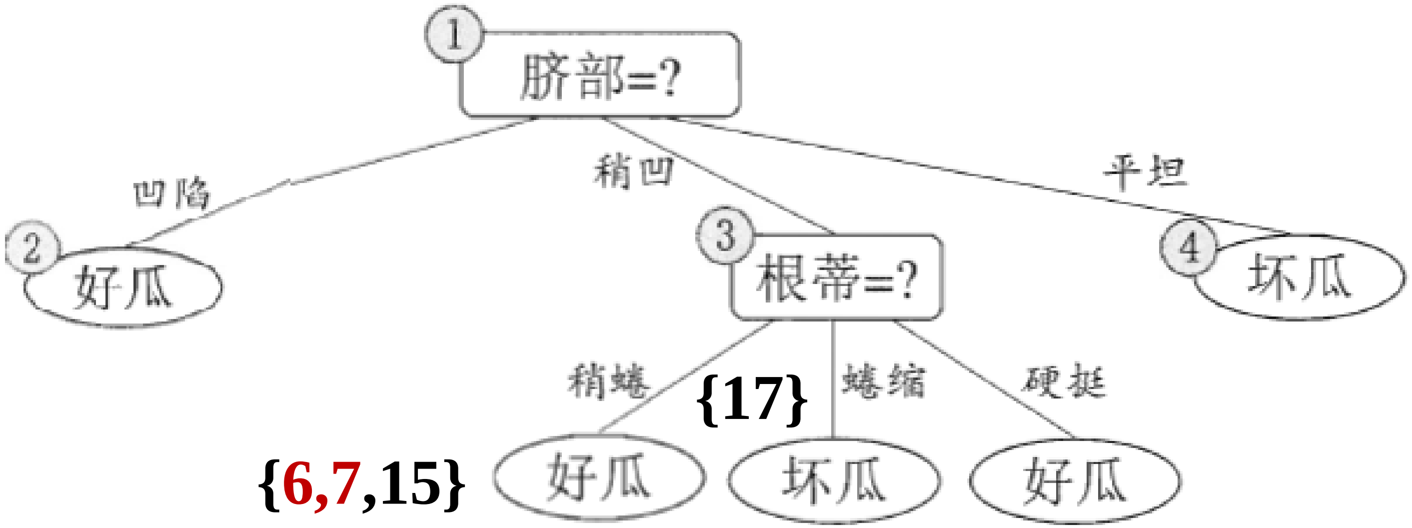
样例：

4：脐部=凹陷-好瓜-正确
and色泽=青绿
好瓜-正确

5：脐部=凹陷-好瓜-正确
and色泽=浅白
坏瓜-错误

13：脐部=凹陷-好瓜-错误
and色泽=青绿
好瓜-错误

划分后正确率：4/7-不划分



编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

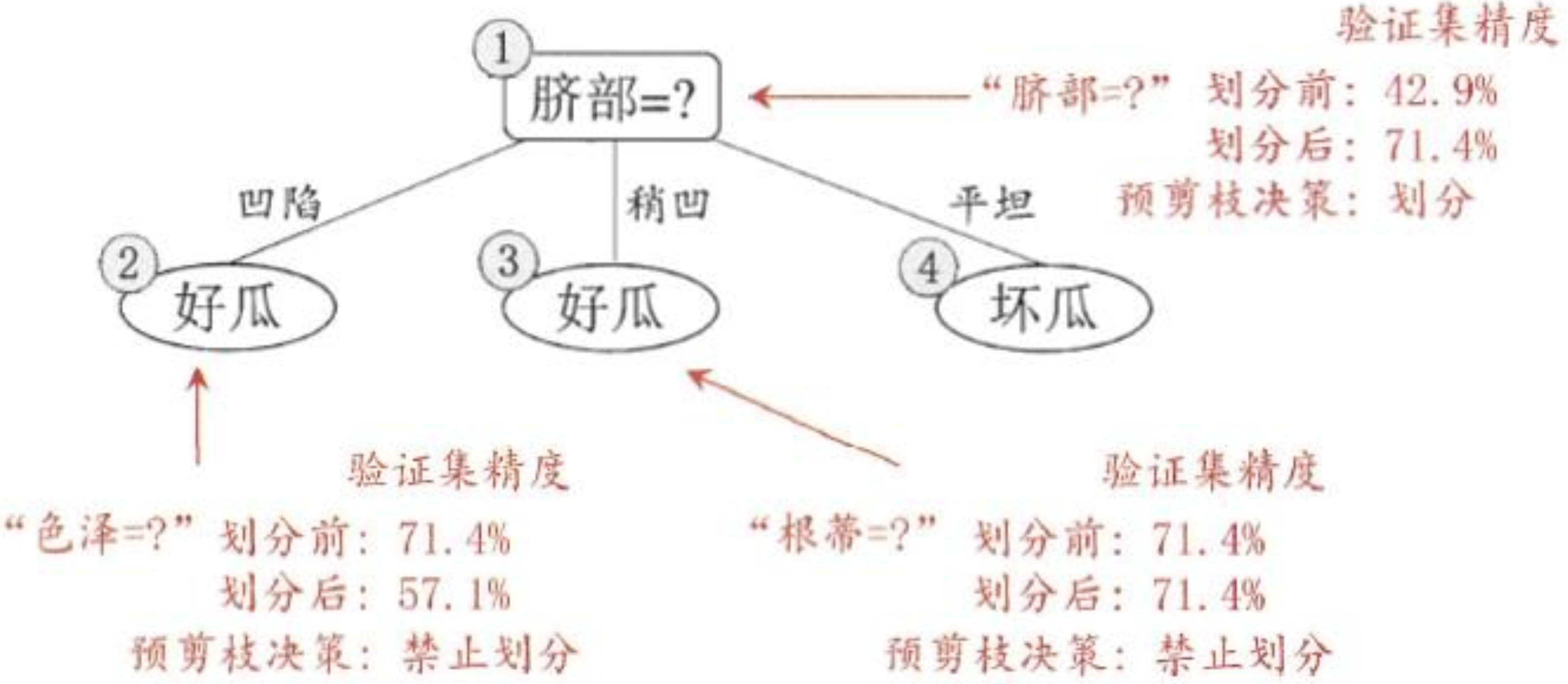
样例集：

8：脐部=稍陷-好瓜-正确
and根蒂=稍卷
好瓜-正确

9：脐部=稍凹-好瓜-错误
and根蒂=稍卷
好-错误

划分后正确率：5/7-不划分

保留分支数	训练时间	泛化性能	欠拟合风险
少	小	强	大
很多分支未展开	学到的特点不足以对样本进行正确分类		

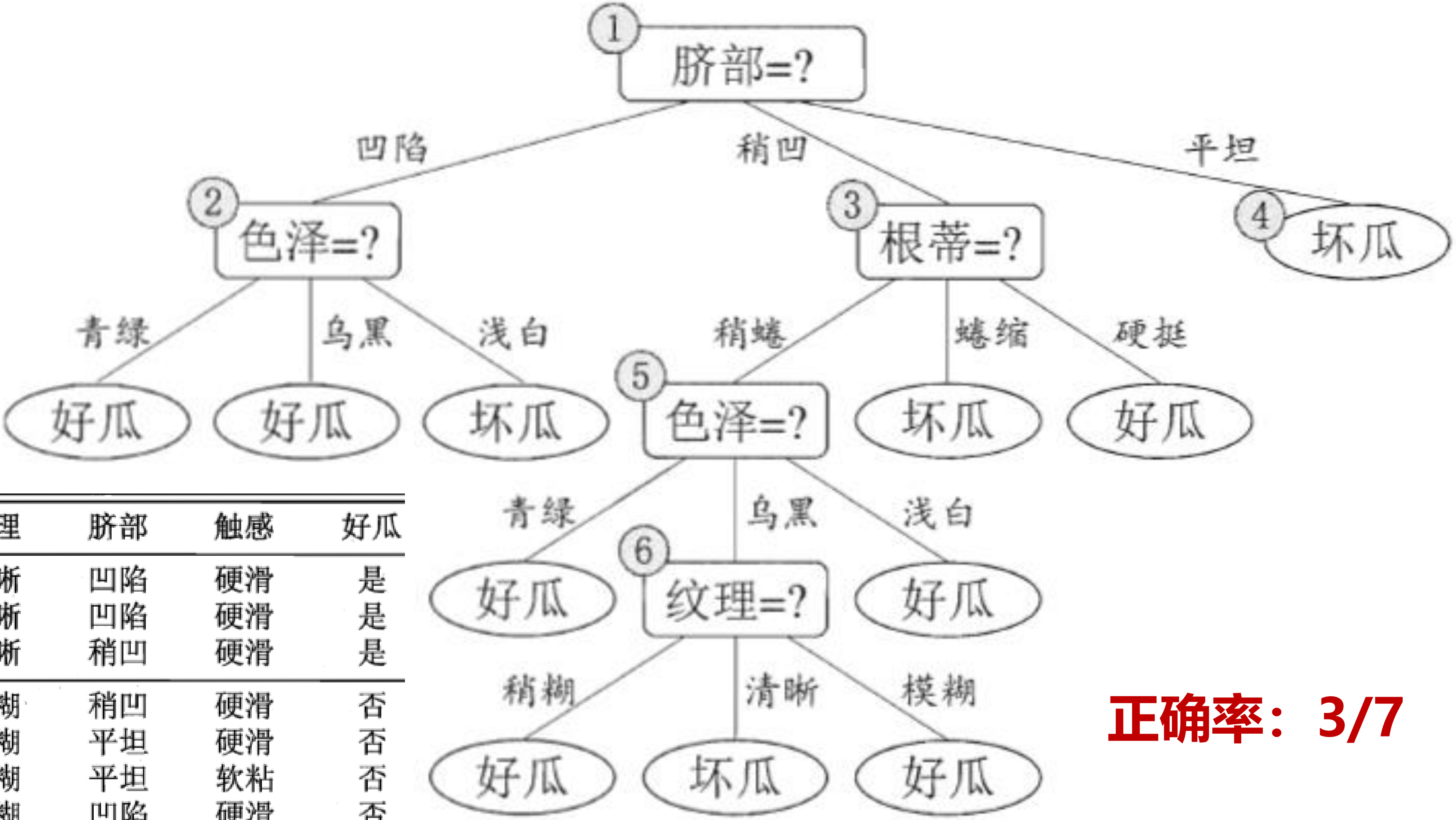

1
脐部=?
凹陷
稍凹
平坦
2
好瓜
3
好瓜
4
坏瓜
验证集精度
“脐部=?” 划分前: 42.9%
划分后: 71.4%
预剪枝决策: 划分
验证集精度
“色泽=?” 划分前: 71.4%
划分后: 57.1%
预剪枝决策: 禁止划分
验证集精度
“根蒂=?” 划分前: 71.4%
划分后: 71.4%
预剪枝决策: 禁止划分

三、后剪枝：

样例：

4：脐部=凹陷-好瓜-正确
5：脐部=凹陷-坏瓜-错误
8：脐部=稍凹-坏瓜-错误
9：脐部=稍凹-好瓜-错误
11：脐部=平坦-坏瓜-正确
12：脐部=平坦-坏瓜-正确
13：脐部=凹陷-好瓜-错误

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否



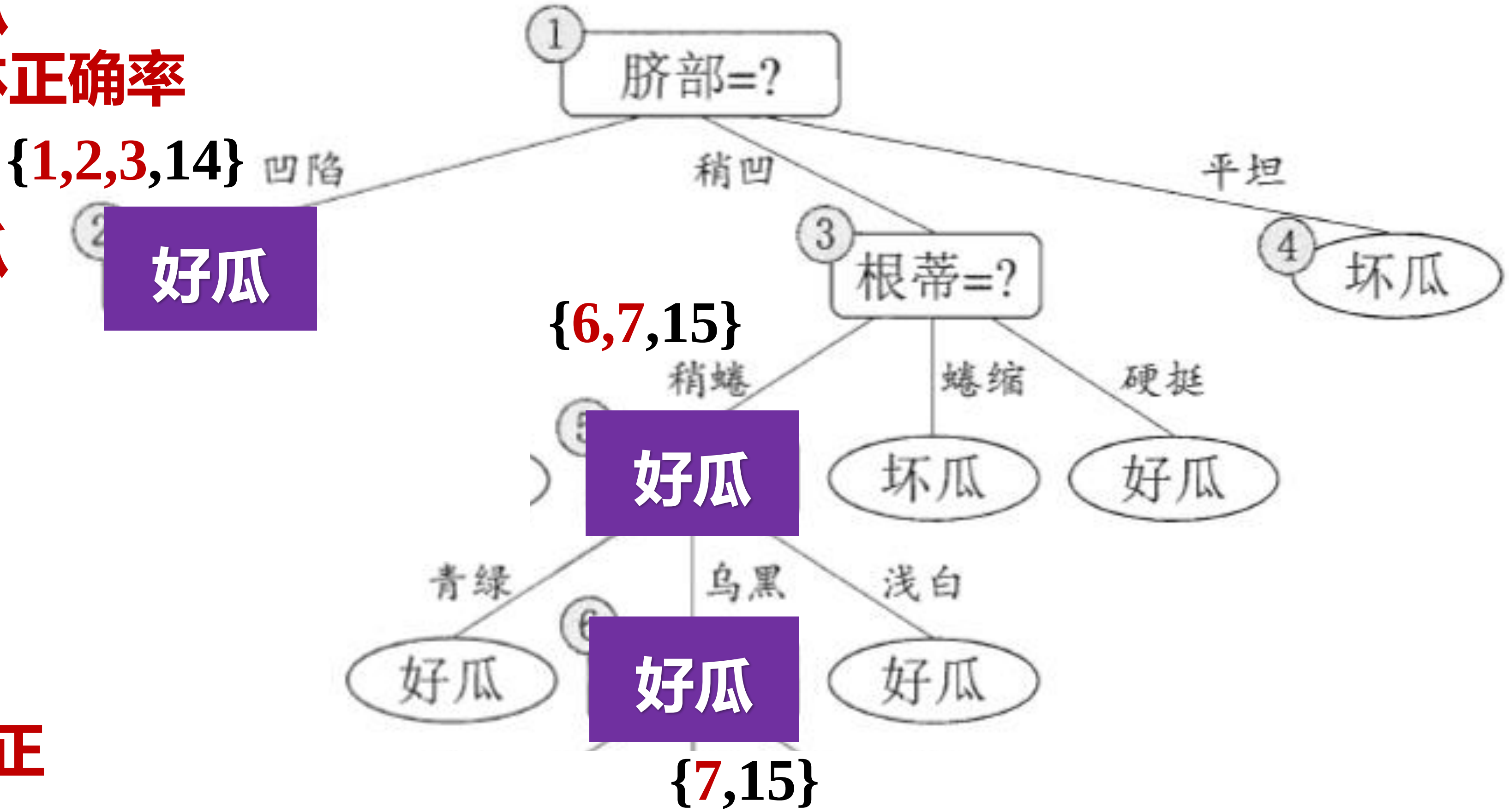
正确率：3/7

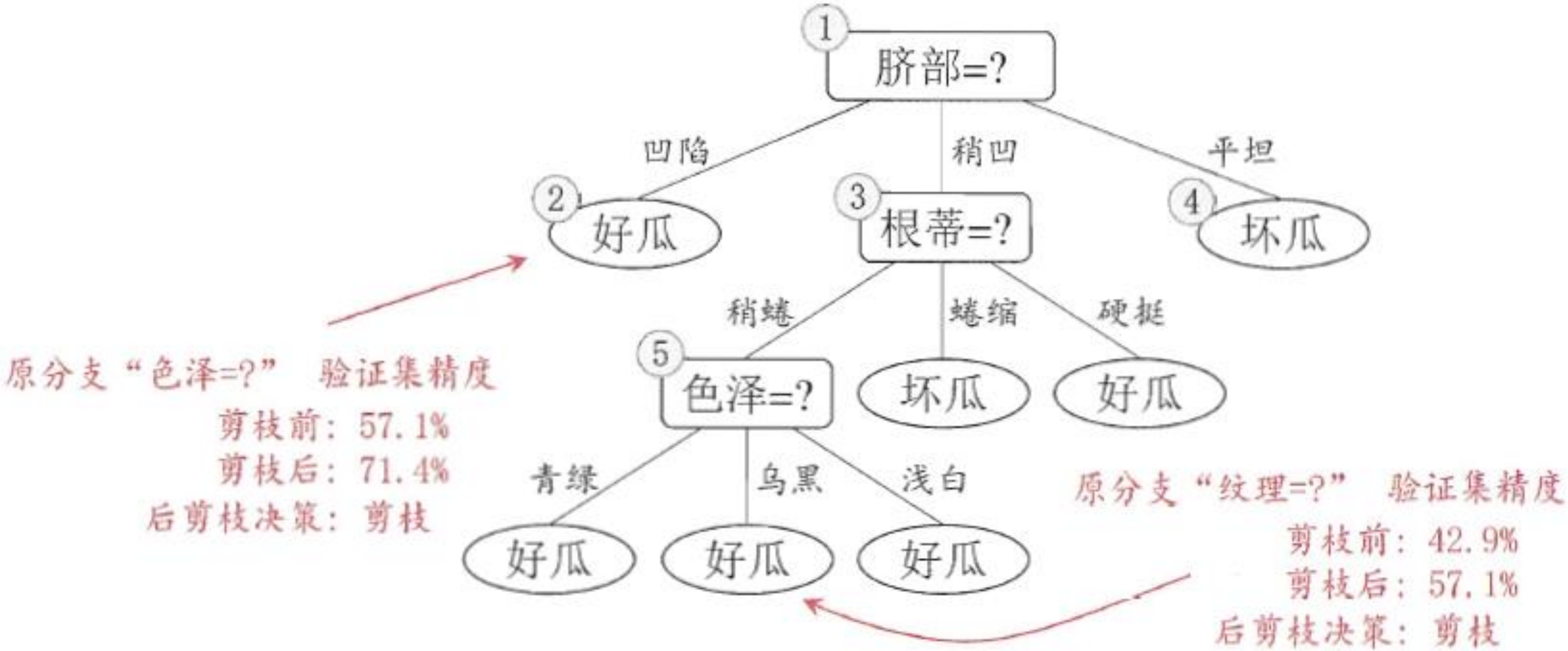
1、考虑节点⑥-叶节点-好瓜
样例8由原来错误-正确-整体正确率提高-剪枝

2、考虑节点⑤-叶节点-好瓜
整体正确率不变-不减枝

3、考虑节点②
-叶节点-好瓜
整体正确率提高
-减枝

验证集中{4,5,8,11,12}分类正确，决策树验证集精度为71.4%





1
脐部=?
凹陷
稍凹
平坦
2
好瓜
3
根蒂=?
4
坏瓜
稍蜷
蜷缩
硬挺
5
色泽=?
坏瓜
好瓜
青绿
乌黑
浅白
好瓜
好瓜
好瓜
原分支“色泽=?” 验证集精度
剪枝前：57.1%
剪枝后：71.4%
后剪枝决策：剪枝
原分支“纹理=?” 验证集精度
剪枝前：42.9%
剪枝后：57.1%
后剪枝决策：剪枝

表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

Entropy Bias

- The entropy measure guides the entire tree building process.
- There is a natural bias that favours attributes with many values.
- Consider the attribute “Birth Date”
    - Separate the training data into very small subsets.
    - Very high information gain
    - A very poor predicator of the target function over unseen instances.
- Such attributes need to be penalized!

$$SplitInformation(S,A) = -\sum_{i=1}^{C} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}$$

GINI

- Gini Index for a given node t :

$$Gini(D) = \sum_{k=1}^{N} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{N} p_k^2$$

(NOTE: p is the relative frequency of class j at node t).

- **Maximum** (1 - 1/n_c) when records are equally distributed among all classes, implying least interesting information
- **Minimum** (0.0) when all records belong to one class, implying most interesting information

GINI

$$Gini(D) = \sum_{k=1}^{N} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{N} p_k^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – $P(C1)^2$ – $P(C2)^2$ = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – $(1/6)^2$ – $(5/6)^2$ = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – $(2/6)^2$ – $(4/6)^2$ = 0.444

GINI

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,

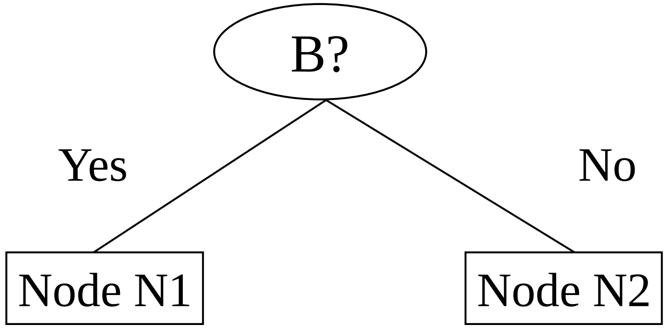
$$\boxed{GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)}$$

where, n_i = number of records at child i,

n = number of records at node p.

GINI

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.



	Parent
C1	**6**
C2	**6**
Gini = 0.500	

Gini (N1)
$= 1 - (5/7)^2 - (2/7)^2 = 0.408$

Gini (N2)
$= 1 - (1/5)^2 - (4/5)^2 = 0.32$

Gini(Children)
$= 7/12 * 0.408 + 5/12 * 0.32$
$= 0.371$

	N1	N2
C1	**5**	**1**
C2	**2**	**4**
Gini=0.371		

Error

- Classification error at a node t :

$$\boxed{Error(t) = 1 - \max_i P(i \mid t)}$$

- Measures misclassification error made by a node.
  - Maximum (1 - 1/n_c) when records are equally distributed among all classes, implying least interesting information
  - Minimum (0.0) when all records belong to one class, implying most interesting information

Error

$$\boxed{Error(t) = 1 - \max_i P(i \mid t)}$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0   P(C2) = 6/6 = 1

Error = 1 – max (0, 1) = 1 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6   P(C2) = 5/6

Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

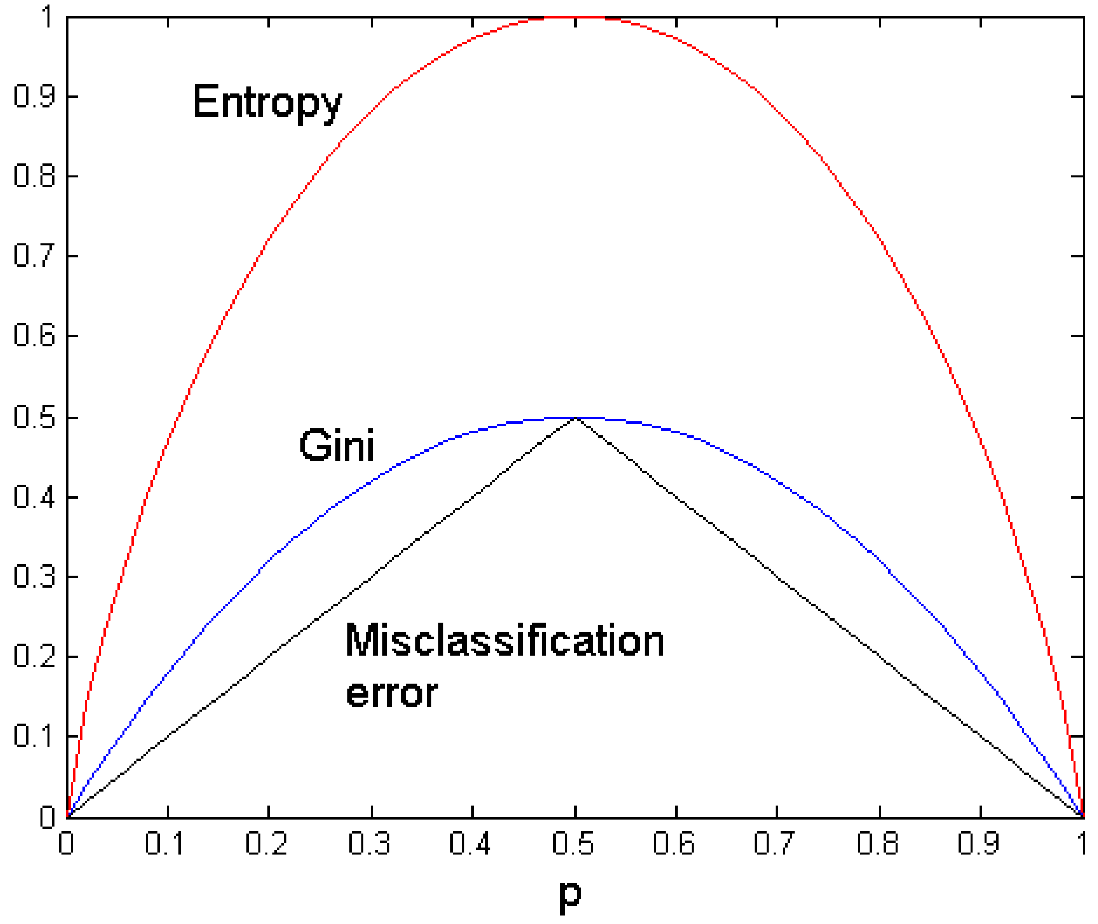
C1	**2**
C2	**4**

P(C1) = 2/6   P(C2) = 4/6

Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

Comparison

For a 2-class problem:



Continuous Attributes

Samples are sorted based on *Temperature*.

Temperature	40	48	60	72	80	90
Play Tennis	No	No	Yes	Yes	Yes	No

Threshold **A**

Threshold **B**

$$Gain(S,A) = Entropy(S) - \frac{1}{3}\cdot 0 - \frac{2}{3}\cdot(-\frac{3}{4}\cdot\log_2\frac{3}{4} - \frac{1}{4}\cdot\log_2\frac{1}{4}) = 1 - 0.541 = 0.459$$

$$Gain(S,B) = Entropy(S) - \frac{1}{6}\cdot 0 - \frac{5}{6}\cdot(-\frac{3}{5}\cdot\log_2\frac{3}{5} - \frac{2}{5}\cdot\log_2\frac{2}{5}) = 1 - 0.809 = 0.191$$

缺失值处理

要解决的问题-决策树

1、如何选择属性?

2、选择划分属性后，如何划分样本集合?

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	–	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	–	是
3	乌黑	蜷缩	–	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	–	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	–	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	–	稍凹	硬滑	是
9	乌黑	–	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	–	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	–	否
12	浅白	蜷缩	–	模糊	平坦	软粘	否
13	–	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	–	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	–	沉闷	稍糊	稍凹	硬滑	否

缺失值处理

D有17个样例，设ω_x=1，a=色泽，

D_a={2,3,4,6,7,8,9,10,11,12,14,15,16,17}

$$Ent(D_a) = -\frac{6}{14}\log_2\frac{6}{14} - \frac{8}{14}\log_2\frac{8}{14} = 0.985$$

{a^1, a^2, a^3}={青绿，乌黑，浅白}

$$Ent(D_a^1) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$Ent(D_a^2) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.918$$

$$Ent(D_a^3) = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4} = 0$$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	–	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	–	是
3	乌黑	蜷缩	–	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	–	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	–	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	–	稍凹	硬滑	是
9	乌黑	–	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	–	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	–	否
12	浅白	蜷缩	–	模糊	平坦	软粘	否
13	–	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	–	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	–	沉闷	稍糊	稍凹	硬滑	否

$$Gain(D_a, 色泽) = Ent(D_a) - \sum_{v=1}^{3} r_a^v Ent(D_a^v)$$

$$= 0.985 - \frac{4}{14}\times 1 - \frac{6}{14}\times 0.918 - \frac{4}{14}\times 0 = 0.306$$

$$Gain(D, 色泽) = \rho \times Gain(D_a, 色泽)$$

$$= \frac{14}{17}\times 0.306 = 0.252$$

缺失值处理

$$\mathrm{Gain}(D,色泽)=0.252;\quad \mathrm{Gain}(D,根蒂)=0.171;$$

$$\mathrm{Gain}(D,敲声)=0.145;\quad \mathrm{Gain}(D,纹理)=0.424;$$

$$\mathrm{Gain}(D,脐部)=0.289;\quad \mathrm{Gain}(D,触感)=0.006.$$

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	–	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	–	是
3	乌黑	蜷缩	–	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	–	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	–	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	–	稍凹	硬滑	是
9	乌黑	–	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	–	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	–	否
12	浅白	蜷缩	–	模糊	平坦	软粘	否
13	–	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	–	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	–	沉闷	稍糊	稍凹	硬滑	否

$$D_{纹理}^{清晰}=\{1,2,3,4,5,6,15,8,10\}$$

$$D_{纹理}^{稍糊}=\{7,9,13,14,17,8,10\}$$

$$D_{纹理}^{模糊}=\{11,12,16,8,10\}$$

样例{8，10}在三个节点的权重分别为：{7/15，5/15，3/15}

缺失值处理

解决问题1：如何选择属性

训练数据集-D；属性-a
D中属性a上没有缺失的样例子集-D_a

设属性a的取值集合为 -$\{a^1, a^2, \cdots, a^V\}$；

$$D_a^v = \{(x_i, y_i) \in D_a \mid x_{ia} = a^v\} \Rightarrow D_a = \bigcup_{v=1}^{V} D_a^v$$

$$D_{ak} = \{(x_i, y_i) \in D_a \mid y_i = C_k\} \Rightarrow D_a = \bigcup_{k=1}^{|Y|} D_{ak}$$

设样例点的权重为ω_x，定义如下量

$$\rho_a = \frac{\sum_{x \in D_a} \omega_x}{\sum_{x \in D} \omega_x}$$

$$p_{ak} = \frac{\sum_{x \in D_{ak}} \omega_x}{\sum_{x \in D_a} \omega_x} \cdots (1 \le k \le |Y|)$$

$$r_a^v = \frac{\sum_{x \in D_a^v} \omega_x}{\sum_{x \in D_a} \omega_x} \cdots (1 \le v \le V)$$

信息增益推广为：

$$Ent(D_a) = -\sum_{k=1}^{|Y|} p_{ak} \log p_{ak}$$

$$Gain(D, a) = \rho_a Gain(D_a, a) = \rho_a \left(Ent(D_a) - \sum_{v=1}^{V} r_a^v Ent(D_a^v) \right)$$

缺失值处理

纹理=清晰

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	权重
1	-	蜷缩	浊响	清晰	凹陷	硬滑	是	1
2	乌黑	蜷缩	沉闷	清晰	凹陷	-	是	1
3	乌黑	蜷缩	-	清晰	凹陷	硬滑	是	1
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是	1
5	-	蜷缩	浊响	清晰	凹陷	硬滑	是	1
6	青绿	稍蜷	浊响	清晰	-	软粘	是	1
8	乌黑	稍蜷	浊响	-	稍凹	硬滑	是	7/15
10	青绿	硬挺	清脆	-	平坦	软粘	否	7/15
15	乌黑	稍蜷	浊响	清晰	-	软粘	否	1

$$Ent(\tilde{D}) = -\sum_{k=1}^{|Y|} \tilde{p}_k log_2 \tilde{p}_k = -0.753 \times log_2 0.753 - 0.247 \times log_2 0.247 = 0.806$$

$$Ent(\tilde{D}^1) = -(\frac{2.467}{3.467} log_2 \frac{2.467}{3.467} + \frac{1}{3.467} log_2 \frac{1}{3.467}) = 0.867$$

(“色泽=乌黑”)

$$Ent(\tilde{D}^2) = -(\frac{2}{2.467} log_2 \frac{2}{2.467} + \frac{0.467}{2.467} log_2 \frac{0.467}{2.467}) = 0.700$$

(“色泽=青绿”)

$$Gain(D,a) = \rho \times \left(Ent(\tilde{D}) - \sum_{v=1}^{V} \tilde{r}_v Ent(\tilde{D}^v) \right)$$

$$= 0.748 \times (0.806 - 0.584 \times 0.867 - 0.416 \times 0.700) = 0.006$$

色泽：

$$\rho = \frac{\sum_{x \in \tilde{D}} w_x}{\sum_{x \in D} w_x}$$

$$\tilde{p}_1 = \frac{\sum_{x \in \tilde{D}_1} w_x}{\sum_{x \in \tilde{D}} w_x}$$

(无缺失值样本中，好瓜的比例)

$$\tilde{p}_2 = \frac{\sum_{x \in \tilde{D}_2} w_x}{\sum_{x \in \tilde{D}} w_x}$$

(无缺失值样本中，坏瓜的比例)

$$\tilde{r}_1 = \frac{\sum_{x \in \tilde{D}^1} w_x}{\sum_{x \in \tilde{D}} w_x}$$

(无缺失值样本中，“色泽=乌黑”的样本的比例)

$$\tilde{r}_2 = \frac{\sum_{x \in \tilde{D}^2} w_x}{\sum_{x \in \tilde{D}} w_x}$$

(无缺失值样本中，“色泽=青绿”的样本的比例)

缺失值处理

解决问题2：如何划分样本集合

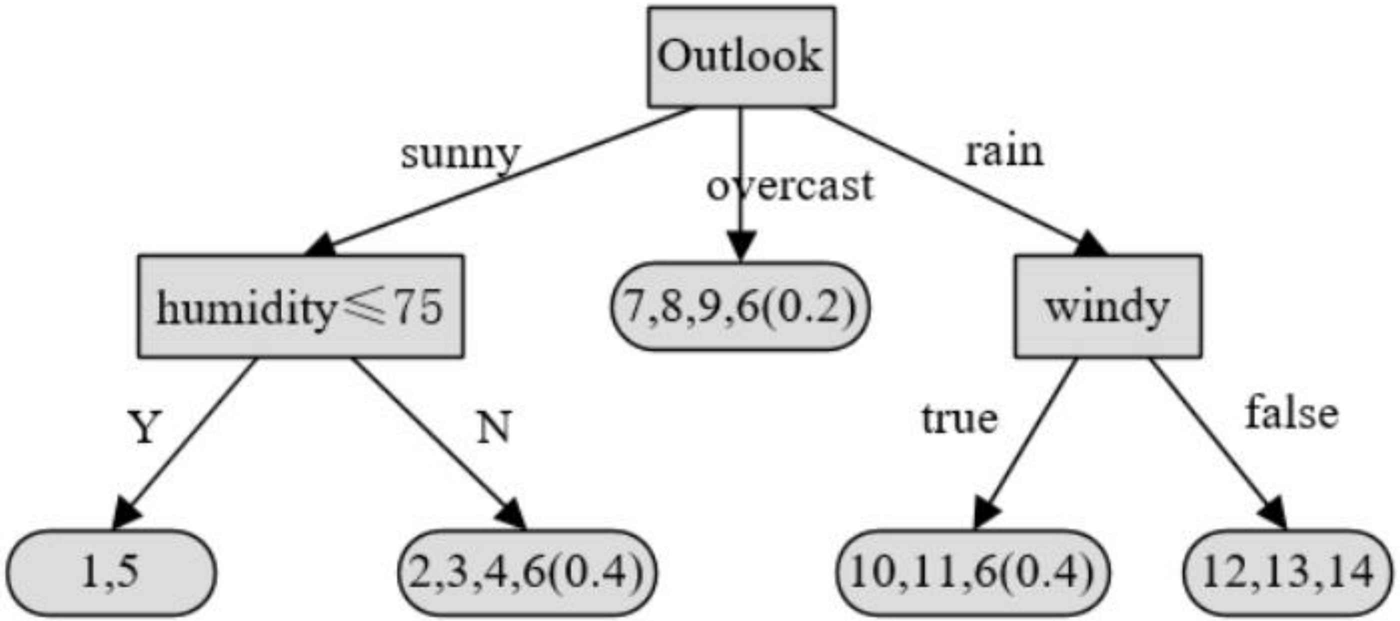
$x \in D_a$ **正常划分**

$x \notin D_a$ **将** x **划分到所有子节点中，但样例的权重值调整为：** $r_a^v \omega_x$

缺失值处理

一个新的问题：含有缺失属性的实际样本如何分类

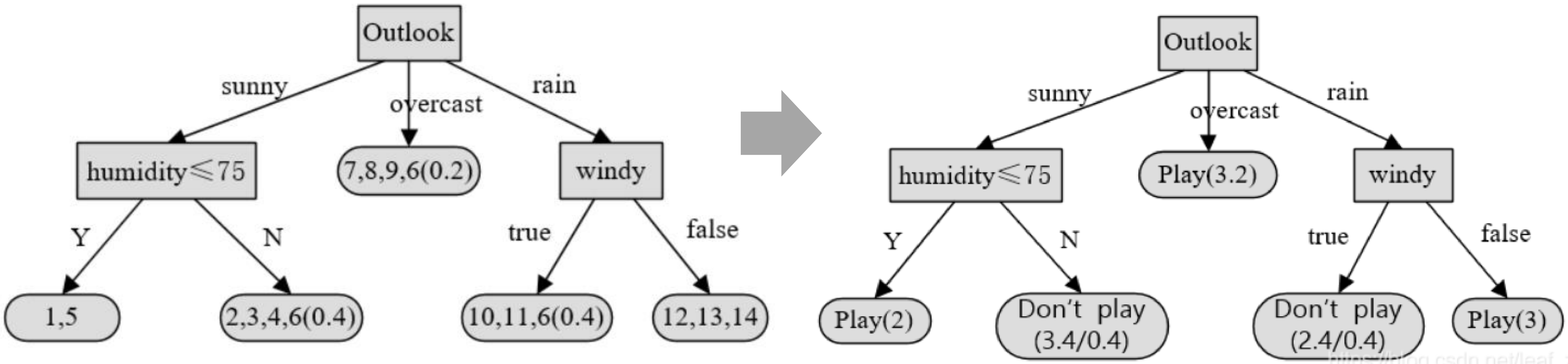
编号	Outlook	Temp(℉)	Humidity(
1	sunny	75	70
2	sunny	80	90
3	sunny	85	85
4	sunny	72	95
5	sunny	69	70
6	-	72	90
7	overcast	83	78
8	overcast	64	65
9	overcast	81	75
10	rain	71	80
11	rain	65	70
12	rain	75	80
13	rain	68	80
14	rain	70	96



outlook=sunny, temperature=70, humidity=?, windy=false.

缺失值处理

一个新的问题：含有缺失属性的实际样本如何分类

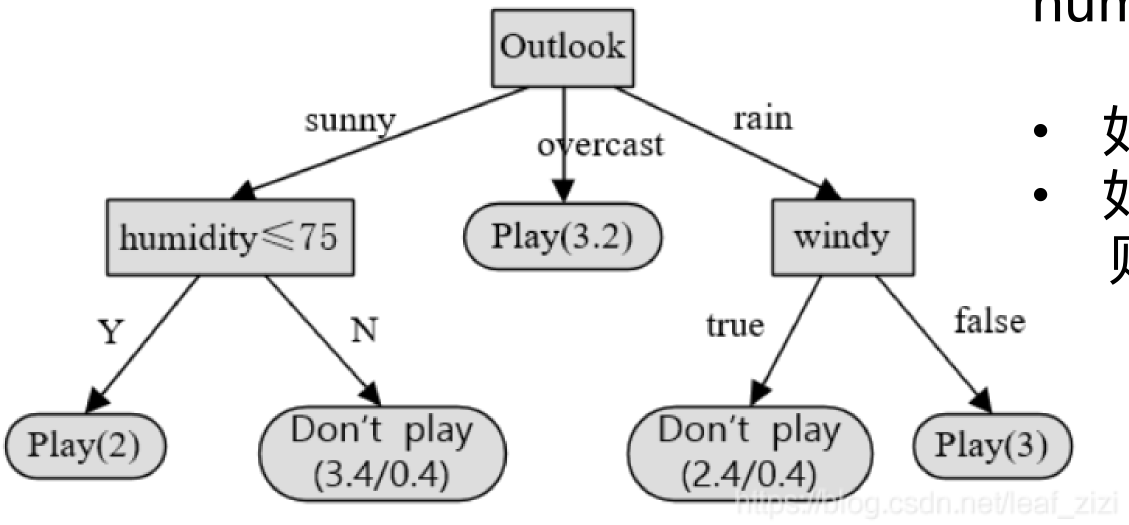


outlook=sunny, temperature=70, humidity=?, windy=false.

缺失值处理

一个新的问题：含有缺失属性的实际样本如何分类

outlook=sunny, temperature=70, humidity=?, windy=false.



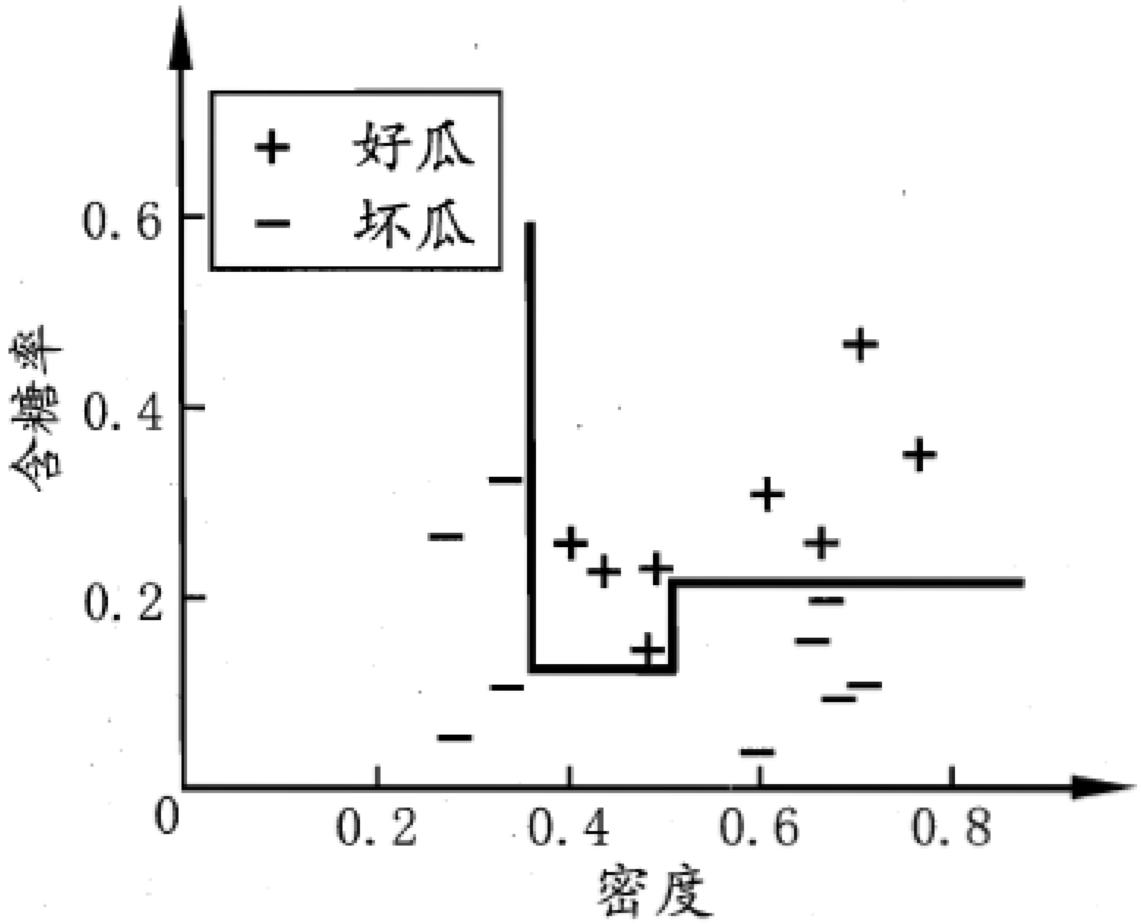
humidity属性值未知，两个分支的可能性都需考虑：

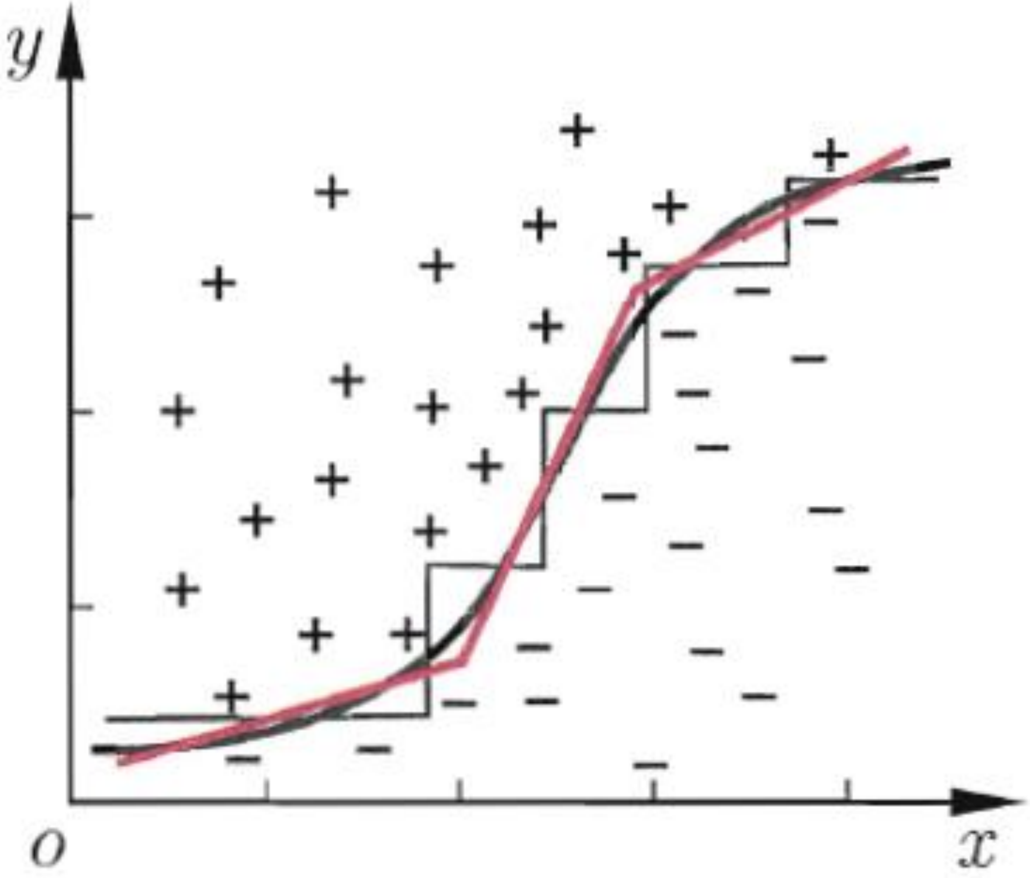
- 如果humidity<=75，类别是play。
- 如果humidity>75，
  则：Don’t play的概率为3/3.4（88%），
  Play的概率是0.4/3.4（12%）。

$$play: 2/5.4 \times 100\% + 3.4/5.4 \times 12\% = 44\%$$

$$Don't\ play: 3.4/5.4 \times 88\% = 56\%$$

多变量决策树的基本思想





Reading Materials

- Online Tutorial
  - http://www.decisiontrees.net/node/21 (with interactive demos)
  - http://www.autonlab.org/tutorials/dtree18.pdf
  - http://people.revoledu.com/kardi/tutorial/DecisionTree/index.html
  - http://www.public.asu.edu/~kirkwood/DAStuff/decisiontrees/index.html
- Tom Mitchell, *Machine Learning,* Chapters 3&6, McGraw-Hill.
- Additional reading about Naïve Bayes Classifier
  - http://www-2.cs.cmu.edu/~tom/NewChapters.html
- Software for text classification using Naïve Bayes Classifier
  - http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html


BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
SHARED
EXAMPLES
EVERY
DEFINITION
CURRENT
MAY
MOVING
MANAGE
WITHIN
ARCHIVES
PRACTITIONERS
CAPTURE
BUSINESS
TECHNOLOGIES
USING
INTERNET
DISTRIBUTED
SETS
GENOMICS
NEEDED
TIME
ONE
LARGE
MANAGEMENT
TERABYTES
CASE
ABILITY
SOFTWARE
INFORMATION
PROCESS
PETABYTES
INCLUDE
PERFORMANCE
STORE
ELAPSED
SYSTEMS
TOLERABLE
SIZE
RESEARCH
STORAGE
MPP
LOGS
COST
PARALLEL
SAN
MASSIVELY
GROW
SET
RELATIONAL
USE
LARGER
CONTINUES
SOCIAL
COMPLEX
CITATION
TENS
ANALYTICS
BIOLOGICAL
PROCESSING
CAPACITY
HUNDREDS
NETWORKS
DATABASES
RECORDS
GARTNER
WORKING
SEARCH
DIFFICULTY