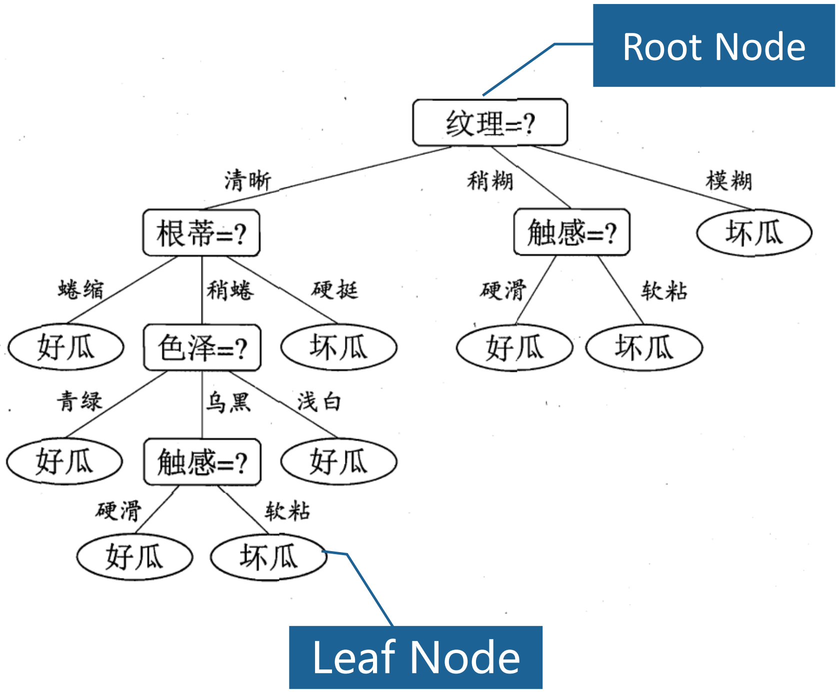
Decision Tree

Outline

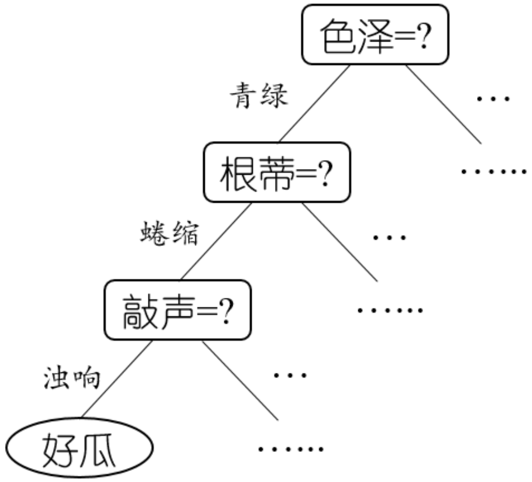
- What's a decision tree
- The algorithm of decision tree
  - Information Gain
  - Gain ratio
  - Pruning tree
  - Continuous attributes
  - Missing values
  - Interpretability
- Summary

Decision Tree



- Every non-leaf node represents a partition of an attribute
- The result of each partition either leads to a further decision problem or leads to the final conclusion
- Decision trees classify instances or examples by starting at the root of the tree and moving through branches until a leaf node
- The final conclusion of decision process corresponds to a target value

How to Construct a Decision Tree



(1) Which attribute to start? (root)

(2) Which attribute to proceed?

(3) When to stop and obtain the target value?

Decision Tree Algorithms

- The basic idea of decision tree algorithm:
  - Choose the *best attribute(s)* to split the remaining instances and make this attribute be a node
  - Repeat this process recursively for successor nodes
  - Stop when:
    - For the current node, all instances have same target value
    - Or there are no more attributes or the instances have the same values in all remaining attributes
    - Or there are no more instances

Choosing Attributes

- One key problem of decision tree algorithm: attribute selection
- Different decision tree algorithms : different methods for attribute selection
- We will focus on the *ID3 (Interactive Dichotomize 3)* algorithm [Ross Quinlan/1975]

Information gain

- ID3 selects attributes according to their information gain
- Information gain is calculated from entropy
- Entropy is the measure of purity of a set

Eg.

- Set1: 10 good watermelons
- Set2: 8 good watermelons and 2 bad watermelons
- Set3: 5 good watermelons and 5 bad watermelons

Purity： Set1 > Set2 > Set3

Entropy

- In general, when p_i is the fraction of instances labeled i,

  $$\text{Entropy}(\{p_1,\ldots,p_k\}) = -\text{sum}(p_i \log(p_i))$$

- Entropy of a set of instances relative to a binary classification is

  $$\text{Entropy} = -p_1\log(p_1) - (1-p_1)\log(1-p_1)$$

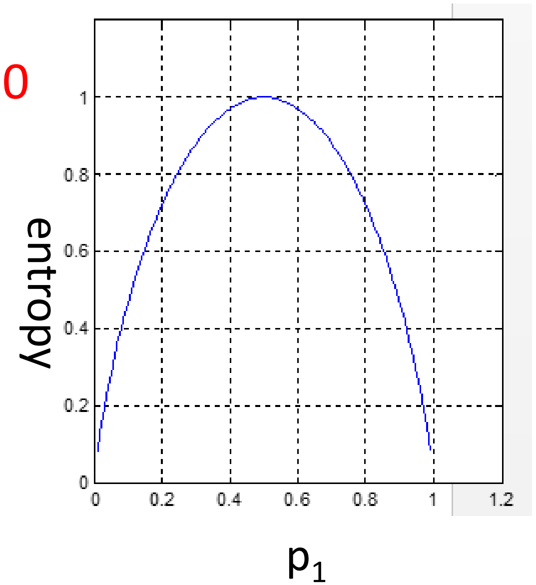
- If all the instances belong to the same class, entropy is 0
- Eg. Set1: 10 good watermelons

  $p_1=1$ or $p_1=0$

- If the instances are equally mixed, entropy is 1

  $p_1=0.5$

- Eg. Set2: 5 good watermelons , 5 bad watermelons



Entropy

- Entropy is minimum when all the instances belong to the same class (highest purity)
- Entropy is maximum when the instances are equally mixed (lowest purity)
- The higher the purity, the smaller the entropy is; the lower the purity, the larger the entropy is.

Information gain

- The information gain of an attribute is the expected reduction in entropy caused by partitioning on this attribute.
- D^i is the subset of D, a is an attribute:

  $$\text{Gain}(D, a) = \text{Entropy}(D) - \sum_{(i = 1 \text{ to } k)} |D^i| / |D| \ \text{Entropy}(D^i)$$

- Partitions: low entropy $\longrightarrow$ high gain
- Eg.

D: 5 good watermelons,5 bad watermelons

D^1: 2 good watermelons, 1 bad watermelon

D^2 : 3 good watermelons,4 bad watermelons

$$\text{Gain}(D, a) = \text{Entropy}(D) - \left(\frac{3}{10}\,\text{Entropy}(D^1) + \frac{7}{10}\,\text{Entropy}(D^2)\right)$$

The example

$\text{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -(\frac{8}{17}\log_2\frac{8}{17} + \frac{9}{17}\log_2\frac{9}{17}) = 0.998$

Training set

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
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
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
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
10	青绿	硬挺	清脆	清晰	平坦	软粘	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
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(色泽=青绿)={1+,4+,6+,10-,13-,17-}

D^2(色泽=乌黑)={2+,3+,7+,8+,9-,15-}

D^3(色泽=浅白)={5+,11-,12-,14-,16-}

$$\text{Ent}(D1) = -(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}) = 1.000$$

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

$$\text{Ent}(D^3) = -(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}) = 0.722$$

$$\sum_{v=1}^{3}\frac{|D^v|}{|D|}\text{Ent}(D^v) = \frac{6}{17}\times 1.000 + \frac{6}{17}\times 0.918 + \frac{5}{17}\times 0.722 = 0.889$$

$$\text{Gain}(D, \text{色泽}) = \text{Ent}(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}\text{Ent}(D^v)$$

$= 0.998 - (\frac{6}{17}\times 1.000 + \frac{6}{17}\times 0.918 + \frac{5}{17}\times 0.722)$

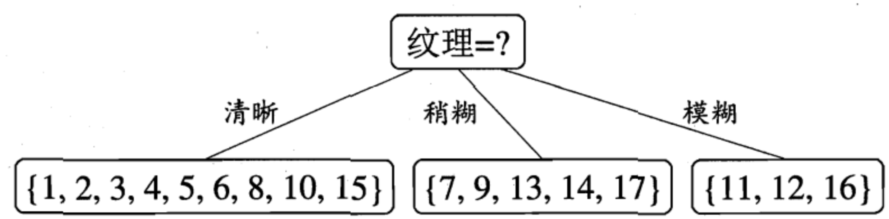
$= 0.109$

The example

$$\mathrm{Gain}(D,\text{色泽}) = \mathrm{Ent}(D) - \sum_{v=1}^{3} \frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$

$$=0.998-(\frac{6}{17}\times 1.000+\frac{6}{17}\times 0.918+\frac{5}{17}\times 0.722)$$

$=0.109$



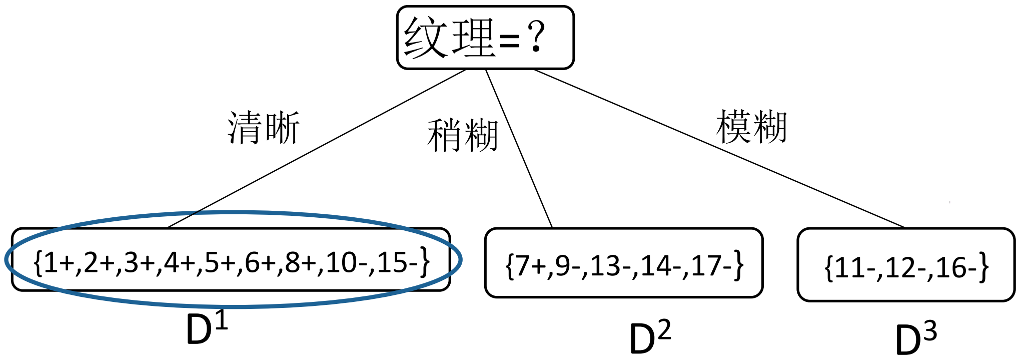
Similarly:

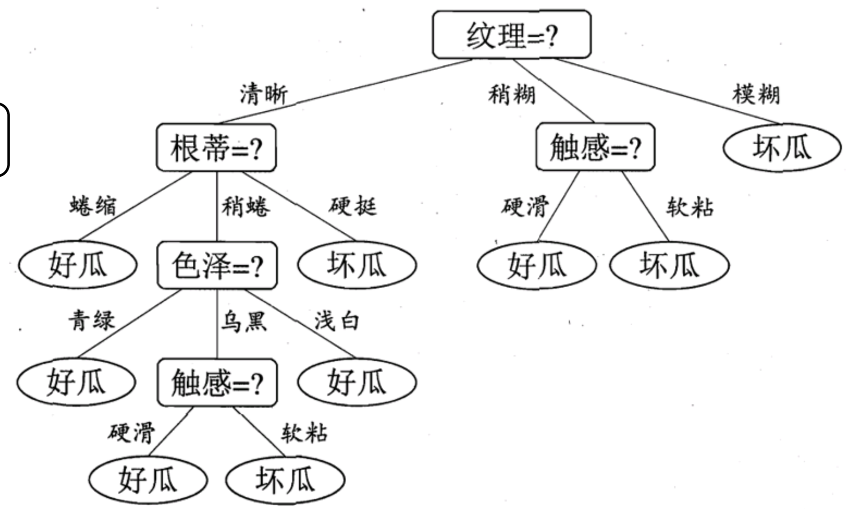
Gain(D,根蒂)=0.143; Gain(D,敲声)=0.141;

Gain(D,纹理)=0.381; Gain(D,脐部)=0.289;

Gain(D,触感)= 0.006

The example





D^1={1+,2+,3+,4+,5+,6+,8+,10-,15-}

Gain(D^1,色泽)=0.043; Gain(D^1,根蒂)=0.458;

Gain(D^1,敲声)=0.331; Gain(D^1,脐部)=0.458;

Gain(D^1,触感)= 0.458

One limitation of ID3

- ID3 tends to select the attribute with more values as the best attribute

如果我们把“编号”视为西瓜的一个属性，它将会被选择为最优属性。

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
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
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
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
10	青绿	硬挺	清脆	清晰	平坦	软粘	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
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

Gain ratio

Gain ratio:

The set of samples

The term is used to measure the number of the values of an attribute

$$\text{Gain\_ratio}(D,a)=\frac{\text{Gain}(D,a)}{\boxed{\mathrm{IV}(a)}}$$

An attribute

The number of the values of the attribute

$$\mathrm{IV}(a) = -\sum_{v=1}^{V}\frac{\boxed{|D^v|}}{|D|}\log_2\frac{|D^v|}{|D|}$$

- Gin ratio tends to the attribute with less values.
- **C4.5** firstly selects these attributes whose information gain is higher than the average information gain, then chooses the attribute with highest gain ratio among these attributes.

Pruning Trees

- Too many branches may cause overfitting.
- There is a technique for reducing the number of branches used in a tree – *pruning*
- Two types of pruning:
  - Pre-pruning (forward pruning)
  - Post-pruning (backward pruning)

Pruning

- Generalization ability is estimated by the accuracy on validation set

- Prepruning: we stop adding attributes during the process of building the decision tree
- Postpruning: we prune the attributes after the full decision tree has been built
- Prepruning & Postpruning: according to generalization ability

Example of Prepruning

Training set

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

Validation set

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

脐部=?

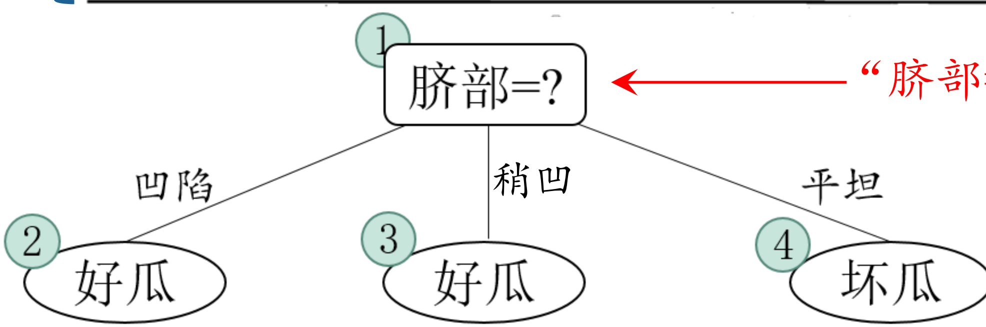
If stop adding this attribute and the label of the node is good:

Accuracy on validation set : 3/7=42.9%

Example of Prepruning

Validation set

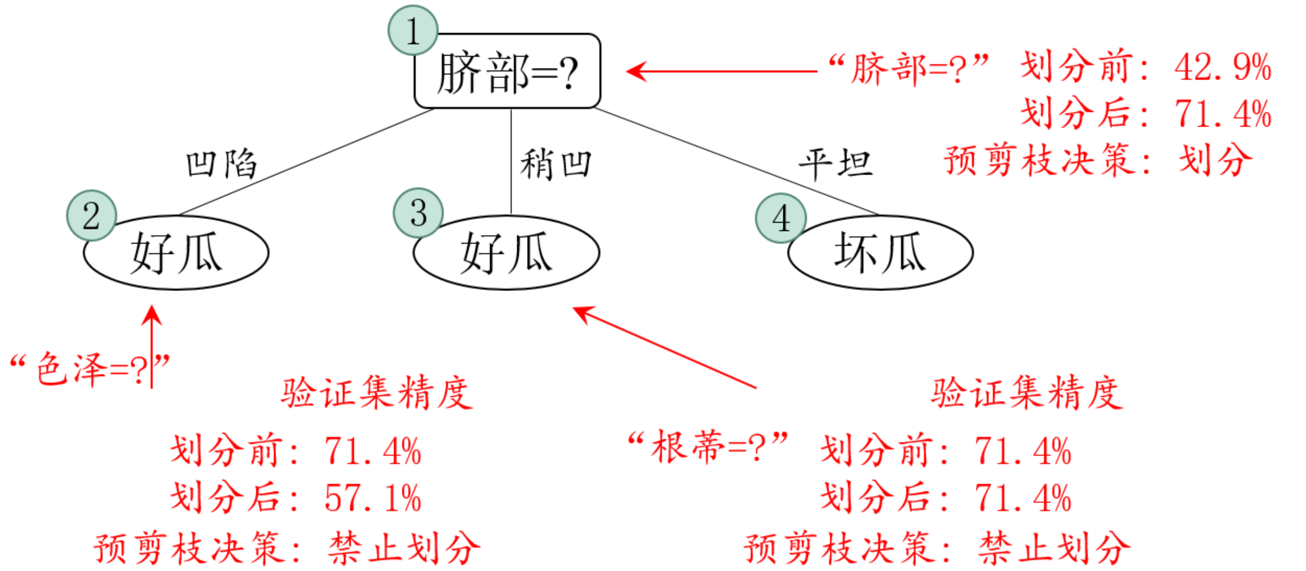
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否



“脐部=?” 划分前：42.9%
划分后：71.4%
预剪枝决策：划分

If don’t stop adding this attribute:
Accuracy on validation set :
(1+1+1+1+1)/7=71.4% > 42.9%

Example of Prepruning

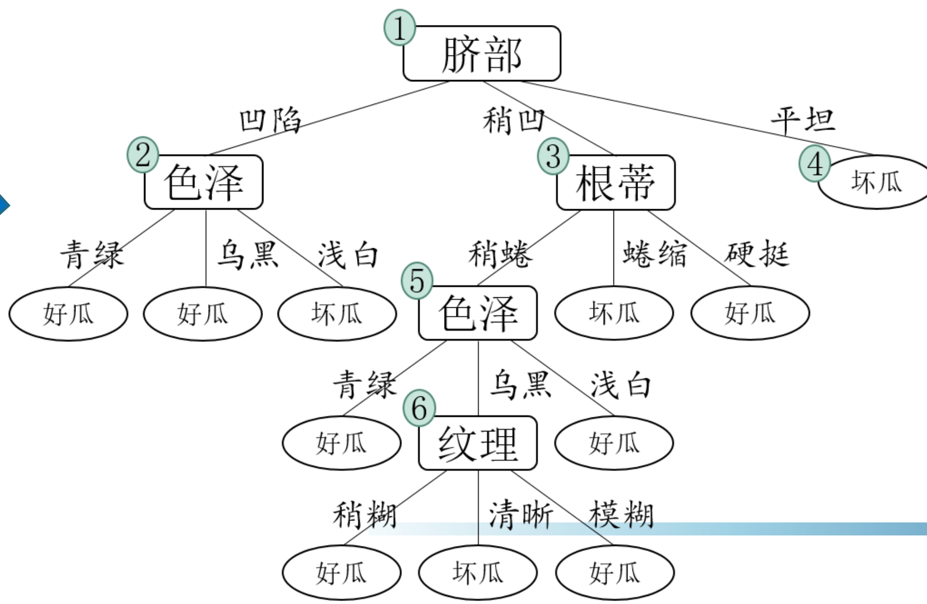


- Prepruning can reduce the risk of overfitting , but it may lead to underfitting.
- Sometimes attributes individually may cause the reduction of generalization ability, but combined, they may improve the generalization ability.

Example of Postpruning

Training set

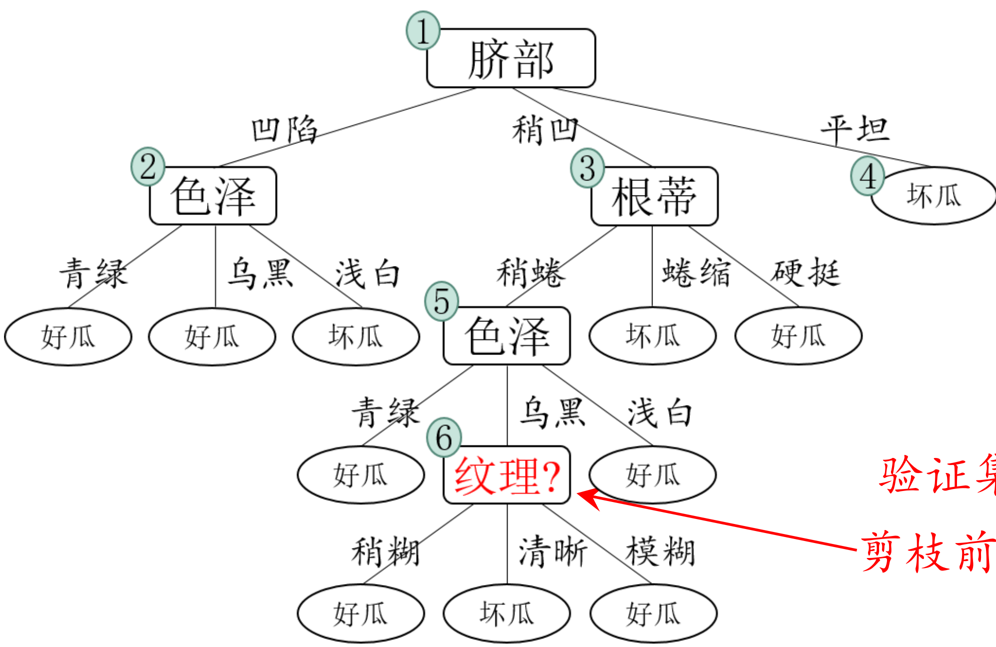
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否



Example of Postpruning

Validation set

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



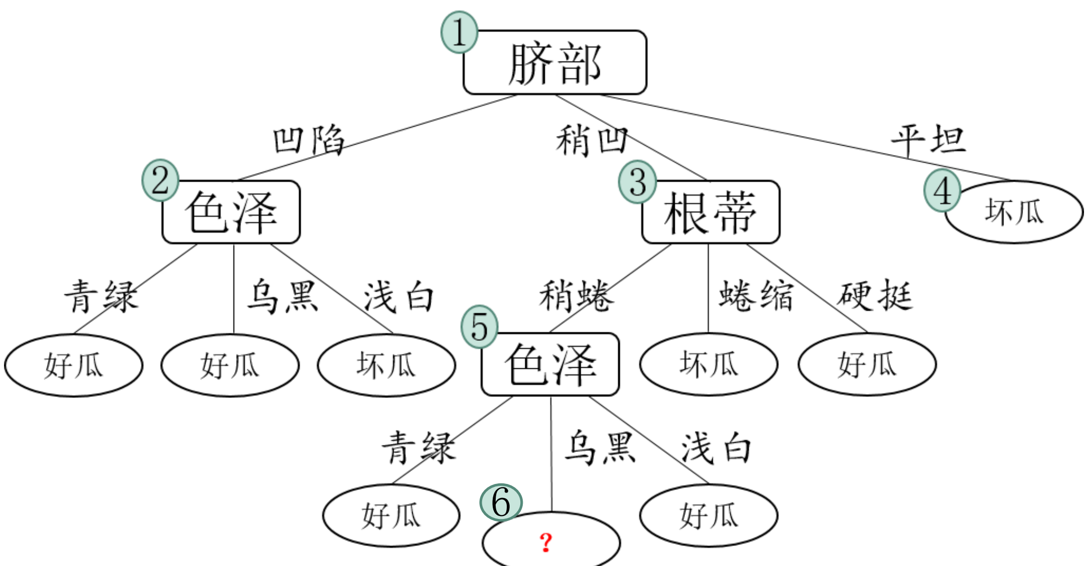
对于节点6，剪枝前
验证集精度：3/7=42.9%

验证集精度
剪枝前：42.9%

Example of Postpruning

Training set

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

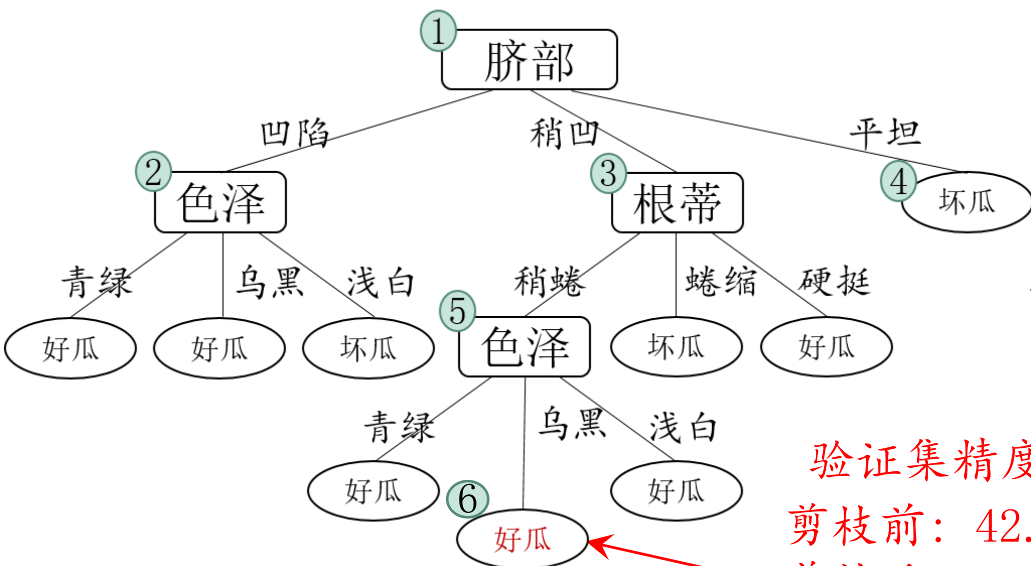


将对节点6进行剪枝，即将节点6替换为叶子节点，当前包含的训练样本为{7+, 15-}，标记为“好瓜”。

Example of Postpruning

Validation set

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

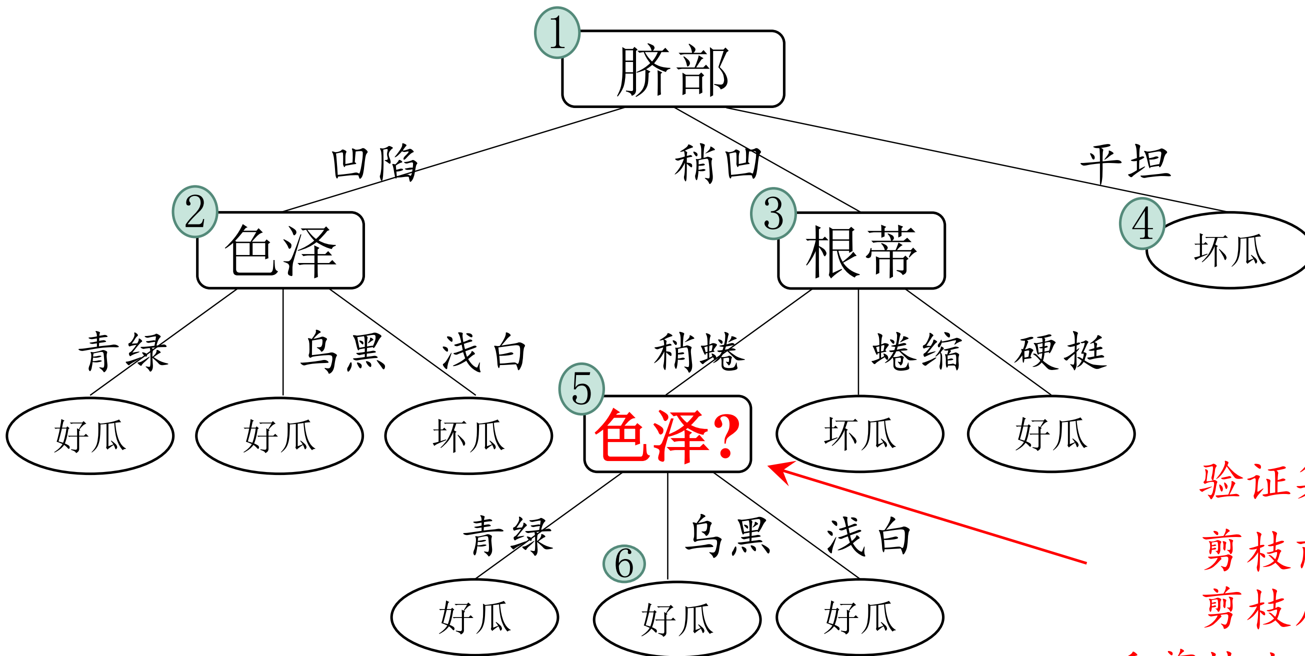


此时验证集精度: 57.1% > 42.9%
因此剪枝。

验证集精度
剪枝前：42.9%
剪枝后：57.1%
后剪枝决策：剪枝

Example of Postpruning

对于节点5:



验证集精度

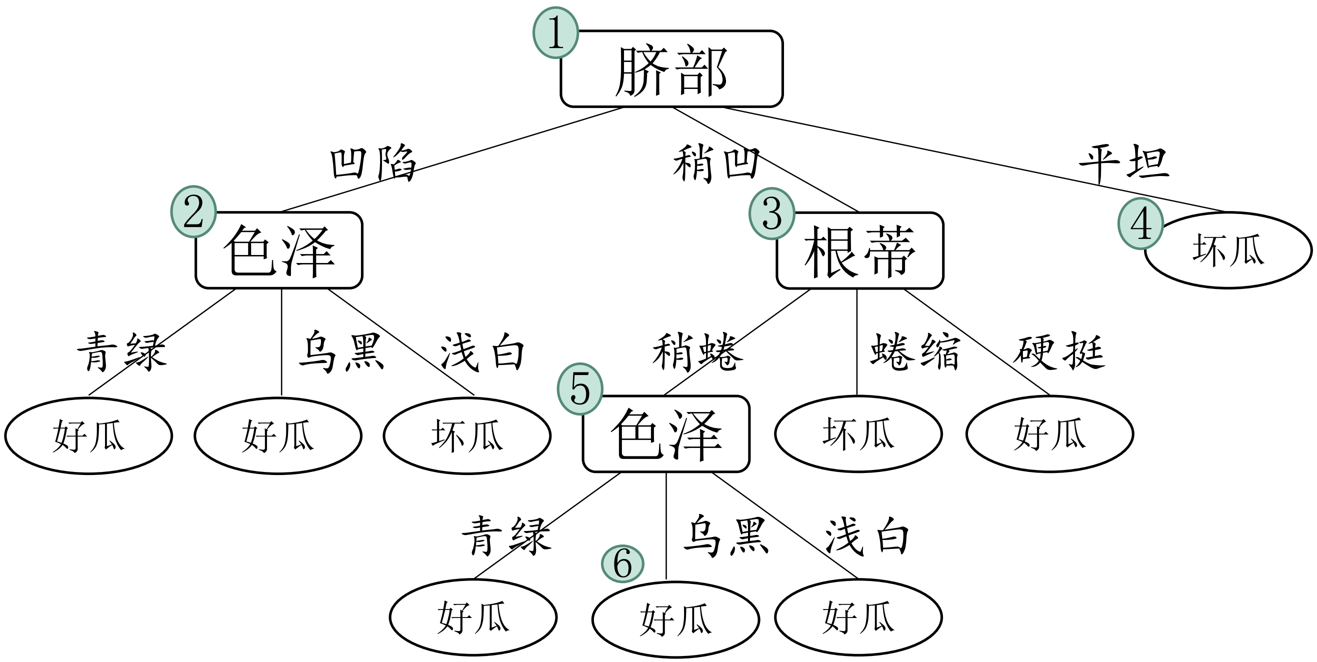
剪枝前：57.1 %

剪枝后：57.1%

后剪枝决策：不剪枝

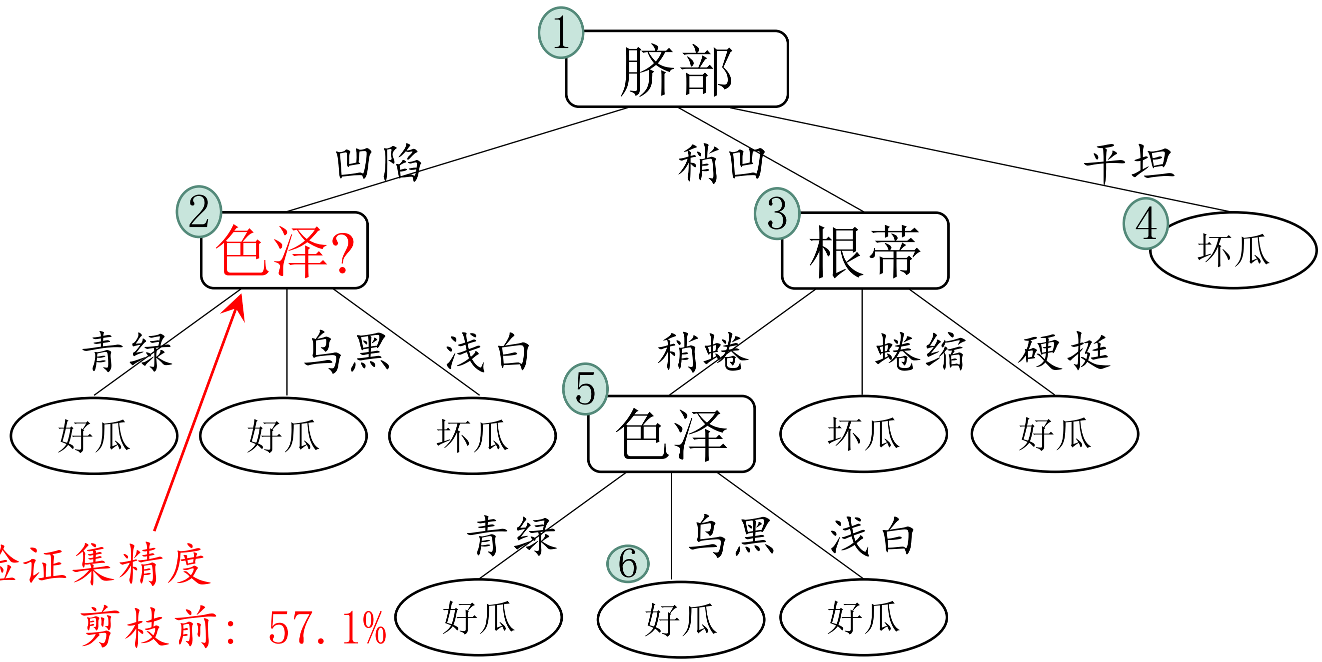
Example of Postpruning

对于节点5:



Example of Postpruning

对于节点2:



验证集精度

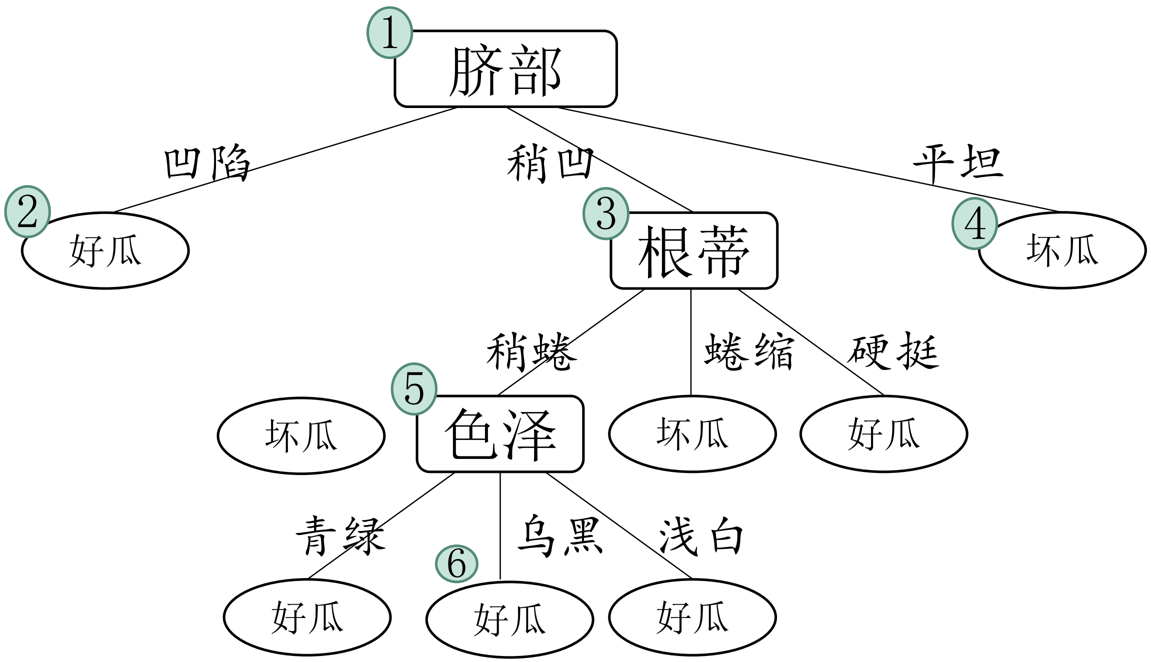
剪枝前：57.1%

剪枝后：71.4%

后剪枝决策：剪枝

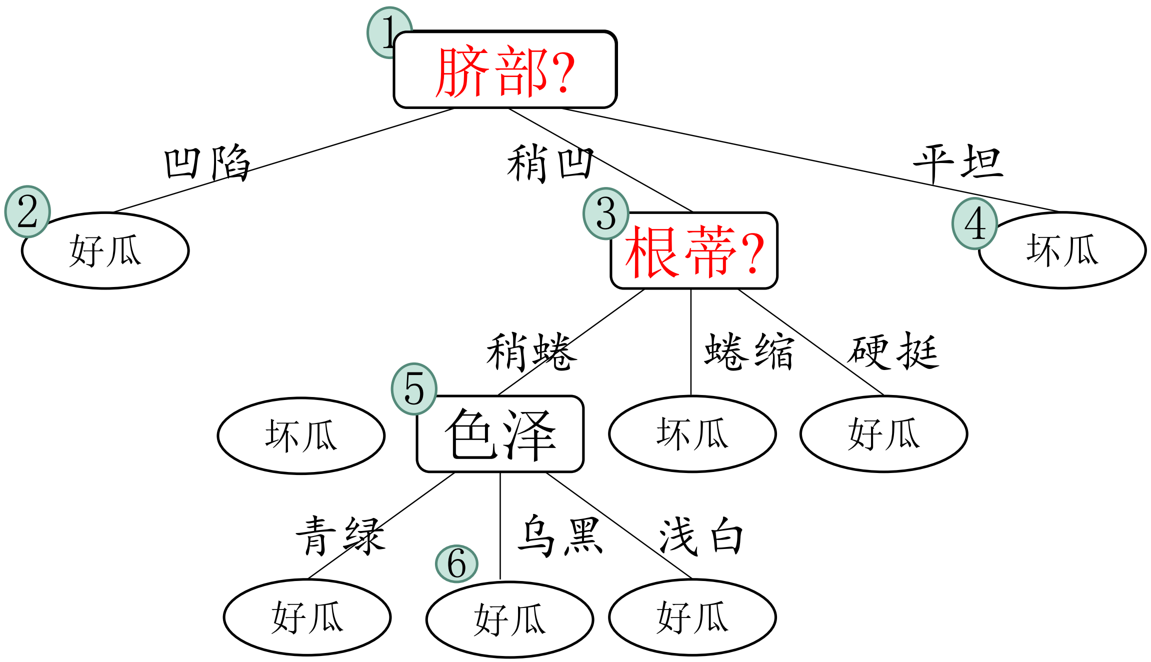
Example of Postpruning

对于节点2:



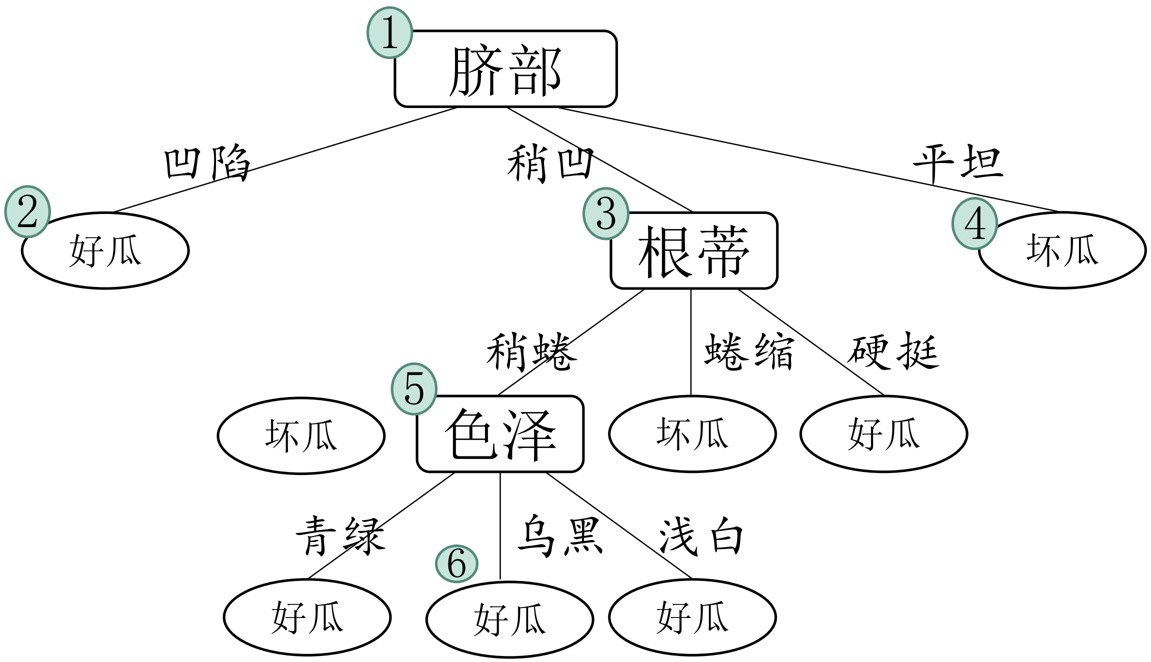
Example of Postpruning

同理，先后把节点3和节点1替换为叶子节点，
验证集精度均未提升，保留分支。



Example of Pruning

最终得到的后剪枝树：



Postpruning

- Advantages:
  - Compared to prepruning, the under-fitting risk of postpruning is low.
  - The generalization ability of postpruning is typically better than that of prepruning.

- Disadvantages:
  - The computational time is expensive.

Continuous Attribute

- Each non-leaf node represents the partition of the attribute (easy for discrete attributes).

- ***C4.5*** use Bi-partition to process continuous attributes:
  - Find a threshold $\boldsymbol{T_a}$ to change continuous attribute A_c to discrete attribute A_d which has two values

$$A_d = \begin{cases} true, & if\ Ad < Ta \\ false, & \texttt{otherwise} \end{cases}$$

How to choose the threshold T_a?

Continuous Attribute

Training set $\{a^1, a^2, \ldots, a^n\}$

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \middle| 1 \le i \le n-1 \right\}$$ (Possible partitions)

$$\mathrm{Gain}(D, a) = \max_{t \in Ta} \mathrm{Gain}(D, a, t)$$
$$= \max_{t \in Ta} \mathrm{Ent}(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^{\lambda}|}{|D|} \mathrm{Ent}(D_t^{\lambda})$$

We choose the threshold corresponding to the partition with highest information gain.

Missing value

Training set

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

Q1: How to select the attribute when some values are missed?

Q2: Given the partitioning attribute, how to partition these examples which miss values on the attribute?

Missing value

- $\widetilde{D}$ which is the subset of D contains the samples which have values on the attribute a
- $\widetilde{D^v}$ which is the subset of $\widetilde{D}$ contains the samples which have value a^v on the attribute a
- $\widetilde{D_k}$ which is the subset of $\widetilde{D}$ contains the samples labeled K

We assign a weight ω_x for each sample $\boldsymbol{x}$.

- The weight ratio of the samples which have values on the attribute a :

$$\rho = \frac{\sum_{x \in \widetilde{D}} \omega_x}{\sum_{x \in D} \omega_x}$$

- The weight ratio of the samples labeled K in $\widetilde{D}$:

$$\widetilde{p_k} = \frac{\sum_{x \in \widetilde{D_k}} \omega_x}{\sum_{x \in \widetilde{D}} \omega_x} \quad (1 \le k \le |y|)$$

- The weight ratio of the samples which have value a^v on the attribute a in $\widetilde{D}$:

$$\widetilde{r_v} = \frac{\sum_{x \in \widetilde{D^v}} \omega_x}{\sum_{x \in \widetilde{D}} \omega_x} \quad (1 \le v \le V)$$

Q1：How to select the attribute when some values are missed?

Missing value

Then, $\text{Gain}(D, a) = \rho \times \text{Gain}(\tilde{D}, a)$

$= \rho \times (\text{Ent}(\tilde{D}) - \sum_{v=1}^{V} \tilde{r}_v \text{Ent}(\widetilde{D^v}))$

$$\text{Ent}(\tilde{D}) = -\sum_{k=1}^{|y|} \widetilde{p_k} \log_2 \widetilde{p_k}$$

As for Q2:

1. For the sample $\boldsymbol{x}$ which has value on the attribute a, we put $\boldsymbol{x}$ in its corresponding child node, and its weight does not change (ω_x).

2. For the sample $\boldsymbol{x}$ which misses value on the attribute a, we put it in all child nodes, and it weight changes to $\tilde{r}_v * \omega_x$

Missing-value example

Training set

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

- 学习开始时，根结点包含样本集中全部17个样本，各样本的权值均初始化为1
- 以“色泽”属性为例，在色泽属性有取值的样本为14个：

$\tilde{D}$={2+,3+,4+,6+,7+,8+,9−,10−,11−,12−,14−,15−,16−,17−}

$$\mathrm{Ent}(\tilde{D}) = -\sum_{k=1}^{2} \tilde{p_k} \log_2 \tilde{p_k}$$

$$=-\left(\frac{6}{14}\log_2\frac{6}{14}+\frac{8}{14}\log_2\frac{8}{14}\right)=0.985$$

Missing-value example

Training set

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

色泽: $\tilde{D}$={2+,3+,4+,6+,7+,8+,9−,10−,11−,12−,14−,15−,16−,17−}

青绿： $\tilde{D}^1$={4+,6+,10-,17- }

乌黑： $\tilde{D}^2$={2+,3+,7+,8+,9-,15-}

浅白： $\tilde{D}^3$={11-,12-14-,16-}

$$\mathrm{Ent}\left(\tilde{D}^1\right)=-\left(\frac{2}{4}\log_2\frac{2}{4}+\frac{2}{4}\log_2\frac{2}{4}\right)=1.000$$

$$\mathrm{Ent}\left(\tilde{D}^2\right)=-\left(\frac{4}{6}\log_2\frac{4}{6}+\frac{2}{6}\log_2\frac{2}{6}\right)=0.918$$

$$\mathrm{Ent}\left(\tilde{D}^3\right)=-\left(\frac{0}{4}\log_2\frac{0}{4}+\frac{4}{4}\log_2\frac{4}{4}\right)=0.000$$

$$\sum_{v=1}^{3}\tilde{r}_v\mathrm{Ent}(\tilde{D}^v)=\frac{4}{14}\times 1.000+\frac{6}{14}\times 0.918+\frac{4}{14}\times 0.000=0.679$$

Missing-value example

色泽: $\tilde{D}$={2+,3+,4+,6+,7+,8+,9−,10−,11−,12−,14−,15−,16−,17−}

Information gain：

$$\text{Gain}(\tilde{D}, \text{色泽})=\text{Ent}(\tilde{D})-\sum_{v=1}^{3}\tilde{r}_v\text{Ent}(\widetilde{D^v})$$

$$=0.985-(\frac{4}{14}\times 1.000+\frac{6}{14}\times 0.918+\frac{4}{14}\times 0.000)$$

$$=0.306$$

我们这里把含有色泽属性样本集的权重所占的比例考虑进去（每个样本的初始权重为1）：

$\tilde{D}$含有14个样本，每个样本的权重为1，所以 $\tilde{D}$总权重为14；
训练集D共包含17个样本，每个样本的权重为1，所以训练集D的总权重为17；
$\tilde{D}$所占权重比例为$\frac{14}{17}$：

$$\text{Gain}(D, \text{色泽})=\rho\times \text{Gain}(\tilde{D}, \text{色泽})=\frac{14}{17}\times 0.306=0.252$$

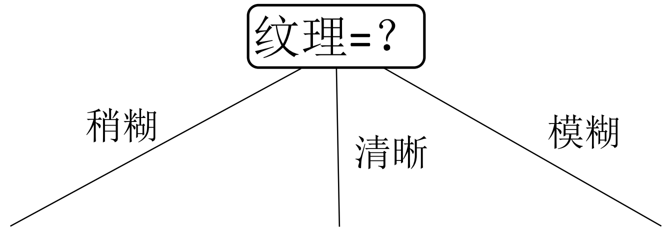
Missing- value example

Similarly,

Gain(D,色泽)=0.252; Gain(D,根蒂)=0.171;

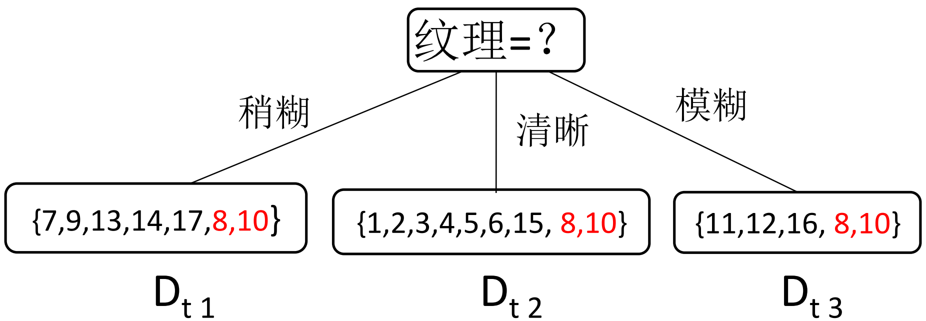
Gain(D,敲声)=0.145; Gain(D,纹理)=0.424;

Gain(D,脐部)=0.289; Gain(D,触感)= 0.006.



Missing-value example

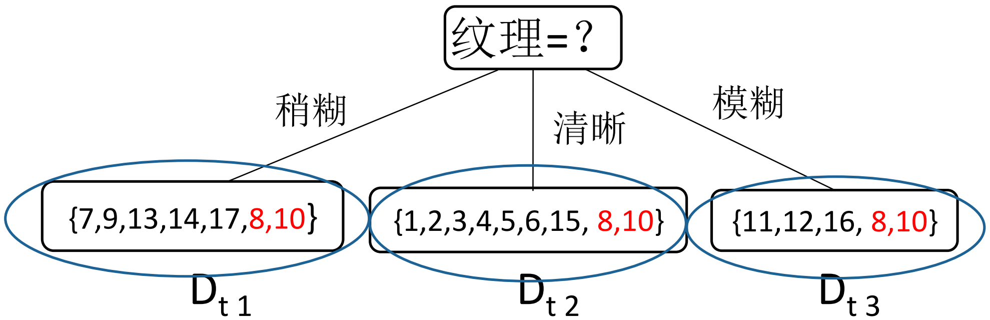
- ✓ 纹理(15个样本) :{1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15, 16, 17}
  其中：稍糊(5个样本): {7,9,13,14,17}
  清晰(7个样本): {1,2,3,4,5,6,15}
  模糊(3个样本): {11,12,16}
- ✓ 缺失纹理属性取值的样本：{8,10}



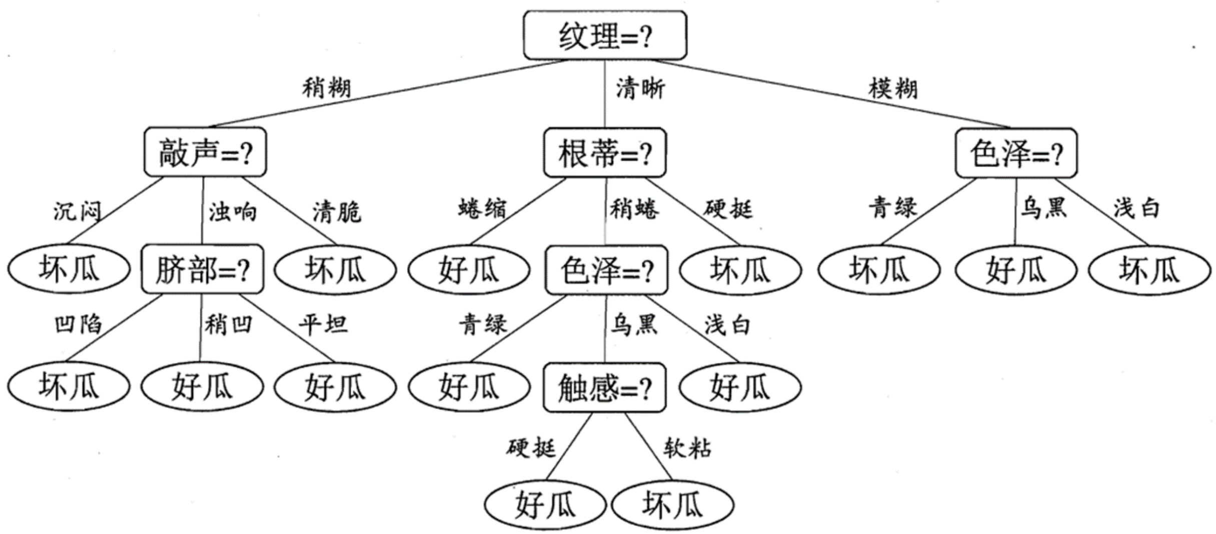
选择纹理属性后，我们把在纹理属性上有取值的样本划分到三个分支，权重不变；同时把在纹理属性上没有取值的样本8,10同时放进三个分支，在三个子节点的权重调整为$\tilde{r_v} * \omega_x$，即 $\frac{5}{15}, \frac{7}{15}, \frac{3}{15}$。则：

1. $D_{t\,1}$各个样本权重为：样本7,9,13,14,17的权重为1, 样本8，10的权重为$\frac{5}{15}$
2. $D_{t\,2}$各个样本权重为：样本1,2,3,4,5,6,15的权重为1, 样本8，10的权重为$\frac{7}{15}$
3. $D_{t\,3}$各个样本权重为：样本1,2,3,4,5,6,15的权重为1, 样本8，10的权重为$\frac{3}{15}$

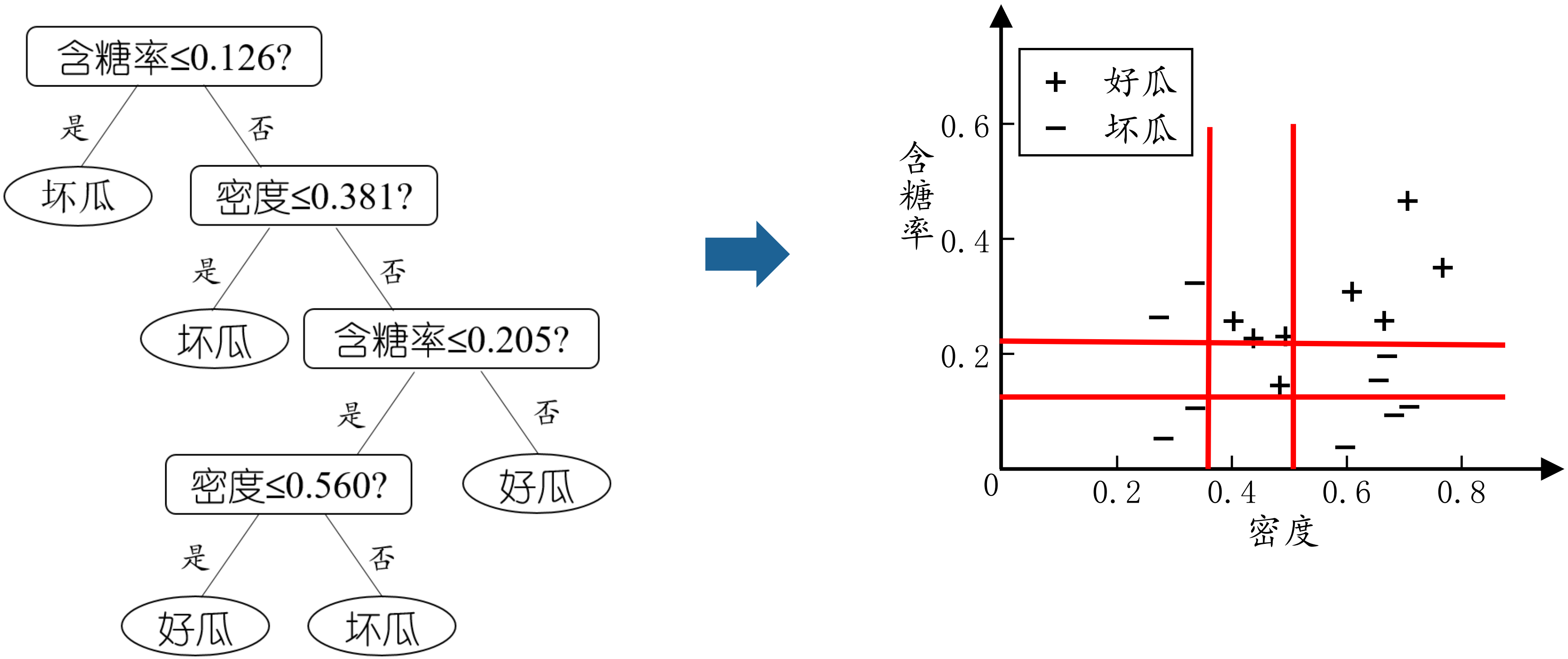
Missing-value example



对于后续节点同理：



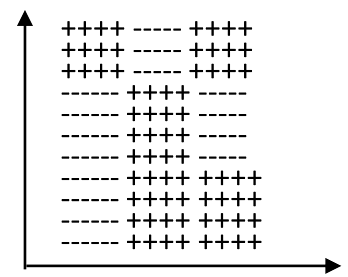
Interpretability



- The boundaries of classification are axis-parallel
- But for too complex problems, they may have too many small segments.

Summary

- Strengths
  - can generate understandable rules
  - perform classification without much computation
  - provide a clear indication of which attributes are most important for prediction or classification
  - Treat well rectangular regions



- Weaknesses
  - The trees may suffer from error propagation
  - Do not treat well non-rectangular regions

RESOURCES

- **C4.5 package:** http://www.rulequest.com/Personal/c4.5r8.tar.gz
- **Wikipedia page for decision tree:** http://en.wikipedia.org/wiki/Decision_tree_learning
- **Random Forests** (Leo Breiman and Adele Cutler): http://www.stat.berkeley.edu/~breiman/RandomForests/
- **ICCV 2013 tutorial:**

  Decision Forests and Fields for Computer Vision: http://research.microsoft.com/enus/um/cambridge/projects/iccv2013tutorial/

References

[Rastogi, et al., 1998] R. Rastogi and K. Shim. PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning. In Proceedings of the 24th International Conference on Very Large Data Bases (VLDB'98), pp. 404–415, Aug. 1998.

[Shafer, et al., 1996] J. C. Shafer, R. Agrawal, and M. Mehta. SPRINT:AScalable Parallel Classifier for Data Mining. In *Proceedings of the 22th International Conference on Very Large Data Bases (VLDB'96), pp. 544–555, Sep. 1996.*

[Gehrke, et al., 1999] J. Gehrke, V. Ganti, R. Ramakrishnan, and W.-H. Loh. BOAT: Optimistic Decision Tree Construction. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'99), pp. 169–180, 1999.

[Thomas, 2000] Dietterich, Thomas (2000). "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization". Machine Learning: 139–157.

[Breiman, 2001] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32.doi:10.1023/A:1010933404324.

Thanks !