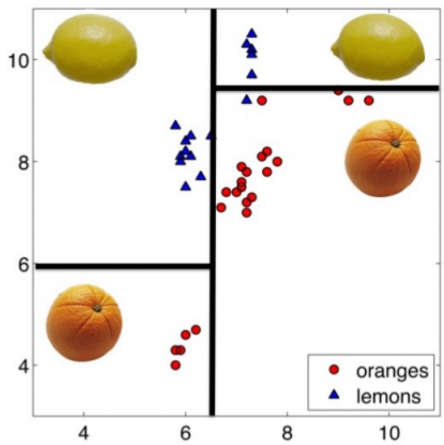
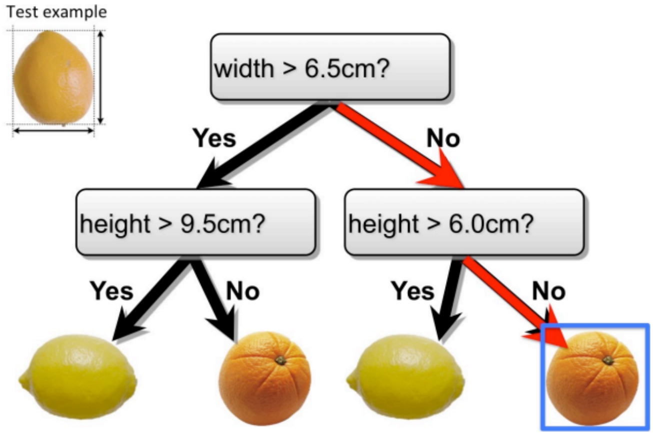
ch3: Decision Trees

Decision Tree: the Key Idea(P9)

- Classify a data sample through a sequence of if-then questions.
  - rule-based
  - splitting data attributes.

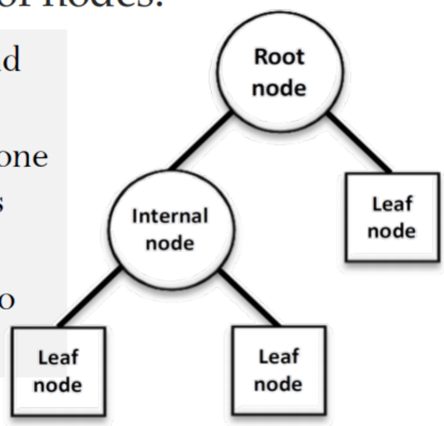




Model Structure(P10)

- A decision tree consists of three types of nodes:

1. A **root node** that has no incoming edges and zero or more outgoing edges
2. **Internal nodes**, each of which has exactly one incoming edge and multiple outgoing edges
3. **Leaf nodes** (or terminal nodes), each of which has exactly one incoming edge and no outgoing edges



▹ each leaf node is assigned a class label

▹ non-terminal nodes contain attribute test conditions to separate records that have different characteristics.

Training – (Build a Decision Tree)(P11)

A top-down divide-and-conquer learning procedure

1.Construct a root node which contains the whole data set.
2.Selecting an attribute that benefit the task most according to some criterion.
3.Split the examples of the current node into subsets based on values of the selected attributes.
4.Create a child node for each subset and passes the examples in the subset to the node.
5.Recursively repeat step 2~4 until some stopping criterion is met.

Training Algorithms

1. ID3(P14-17)

Split attributes that have the maximum **Information Gain.**

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log \frac{|C_k|}{|D|}$$

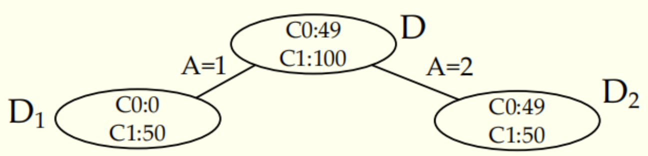
($|C_k|$: # samples of class C_k in dataset D)

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \left(\sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log \frac{|D_{ik}|}{|D_i|}\right)$$

($|D_i|$: # samples whose attribute A is set to the i-th value in D; $|D_{ik}|$: #samples of class C_k in D_i)

Gain $(D,A) = H(D) - H(D|A)$

What is the information gain of this split?



- root entropy: $H(D) = -{}^{49}\!/_{149} \log({}^{49}\!/_{149}) - {}^{100}\!/_{149}\log({}^{100}\!/_{149}) \approx 0.91$
- leaves entropy: $H(D|A=1) = 0, H(D|A=2) \approx 1$
- IG(D|A) $\approx 0.91 - (\frac{1}{3} \cdot 0 + \frac{2}{3} \cdot 1) \approx 0.24 > 0$

- **ID3 – Example(P15-17)**

2. CART(P18-19)

- Find the best split using **Gini Index.**

$$\textbf{Gini}\ (D) = \sum_{k=1}^{K} \frac{|C_k|}{|D|}\left(1 - \frac{|C_k|}{|D|}\right) = 1 - \sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2$$

$$\textbf{Gini}\ (D \mid A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|}\,\mathrm{Gini}(D_i)$$

Gini represents the probability that two randomly selected samples come from different classes.

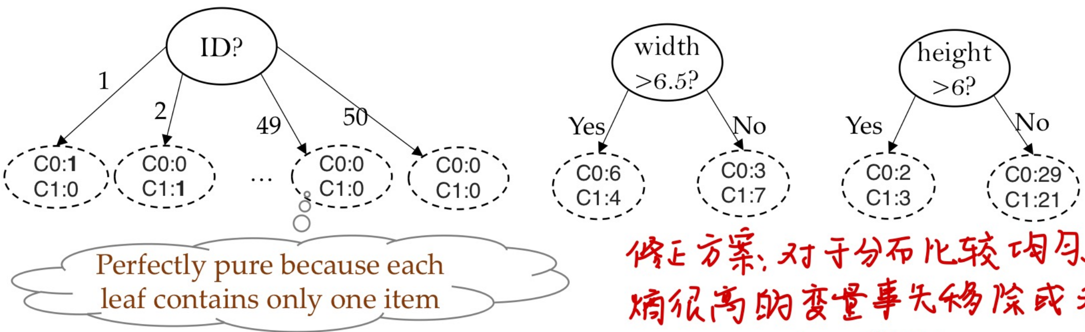
Gini is **cheaper in computation** than Entropy which needs to compute *log* functions.

- **CART – Example(P19)**

Shortcoming of ID3 and CART

- Entropy and Gini favor attributes with large number of distinct values.
喜欢纯净的数据

Possible nodes to split on:



修正方案：对于分布比较均匀熵很高的变量事先移除或者做出惩罚

- ID will result in perfectly pure children.
- Will have the greatest information gain.
- Should have been removed as a predictor variable.

C4.5

To avoid bias to attributes that have many distinct values, C4.5 uses gain ratio instead of information gain.

- ▻ takes into account the number of outcomes produced by attribute split condition. 考虑到由属性分割条件所产生的结果的数量。
- ▻ Adjusts information gain by the entropy of the partitioning.

$$\textbf{Gain Ratio}\ (D, A) = \frac{\text{Gain}(D,A)}{H_A(D)}$$

$$H_A(D) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log \frac{|D_i|}{|D|}$$

- **C4.5 – Example(P22)**

Stop Criteria(P23)

- Purity

  The leaves contain the training examples from the same class

- Minimum number of points

  The number of training examples contained in the leaves are less than a threshold 在叶子中包含的训练示例的数量小于一个阈值

- No more attribute to used for split (ID3)

Pruning(剪枝）(P24)

- Why pruning?
  - To reduce the chance of overfitting

- Pruning strategy
  - Pre-pruning

    Stop growing tree early if the goodness measure is less than a threshold 如果goodness测量值小于一个阈值，停止growing tree
  - Post-pruning

    Remove branches after a tree has been fully grown.

Post-pruning usually performs better than pre-pruning, but its computational cost is heavier.

Regression Trees（不考！😀)