

Because the class label of each training tuple *is provided*, this step is also known as **supervised learning** (i.e., the learning of the classifier is “supervised” in that it is told to which class each training tuple belongs). It contrasts with **unsupervised learning** (or **clustering**), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance. For example, if we did not have the *loan\_decision* data available for the training set, we could use clustering to try to determine “groups of like tuples,” which may correspond to risk groups within the loan application data. Clustering is the topic of Chapters 10 and 11.

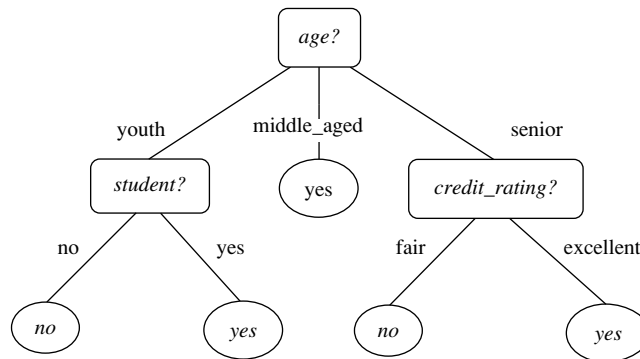
This first step of the classification process can also be viewed as the learning of a mapping or function,  $y = f(X)$ , that can predict the associated class label  $y$  of a given tuple  $X$ . In this view, we wish to learn a mapping or function that separates the data classes. Typically, this mapping is represented in the form of classification rules, decision trees, or mathematical formulae. In our example, the mapping is represented as classification rules that identify loan applications as being either safe or risky (Figure 8.1a). The rules can be used to categorize future data tuples, as well as provide deeper insight into the data contents. They also provide a compressed data representation.

“What about classification accuracy?” In the second step (Figure 8.1b), the model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the classifier’s accuracy, this estimate would likely be optimistic, because the classifier tends to **overfit** the data (i.e., during learning it may incorporate some particular anomalies of the training data that are not present in the general data set overall). Therefore, a **test set** is used, made up of **test tuples** and their associated class labels. They are independent of the training tuples, meaning that they were not used to construct the classifier.

The **accuracy** of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier’s class prediction for that tuple. Section 8.5 describes several methods for estimating classifier accuracy. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known. (Such data are also referred to in the machine learning literature as “unknown” or “previously unseen” data.) For example, the classification rules learned in Figure 8.1(a) from the analysis of data from previous loan applications can be used to approve or reject new or future loan applicants.

## 8.2 Decision Tree Induction

**Decision tree induction** is the learning of decision trees from class-labeled training tuples. A **decision tree** is a flowchart-like tree structure, where each **internal node** (non-leaf node) denotes a test on an attribute, each **branch** represents an outcome of the test, and each **leaf node** (or *terminal node*) holds a class label. The topmost node in a tree is the **root** node. A typical decision tree is shown in Figure 8.2. It represents the concept *buys\_computer*, that is, it predicts whether a customer at *Allelectronics* is



**Figure 8.2** A decision tree for the concept *buys\_computer*, indicating whether an *Allelectronics* customer is likely to purchase a computer. Each internal (nonleaf) node represents a test on an attribute. Each leaf node represents a class (either *buys\_computer* = *yes* or *buys\_computer* = *no*).

likely to purchase a computer. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decision tree algorithms produce only *binary* trees (where each internal node branches to exactly two other nodes), whereas others can produce nonbinary trees.

“How are decision trees used for classification?” Given a tuple,  $X$ , for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules.

“Why are decision tree classifiers so popular?” The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle multidimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule induction systems.

In Section 8.2.1, we describe a basic algorithm for learning decision trees. During tree construction, *attribute selection measures* are used to select the attribute that best partitions the tuples into distinct classes. Popular measures of attribute selection are given in Section 8.2.2. When decision trees are built, many of the branches may reflect noise or outliers in the training data. *Tree pruning* attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data. Tree pruning is described in Section 8.2.3. Scalability issues for the induction of decision trees

from large databases are discussed in Section 8.2.4. Section 8.2.5 presents a visual mining approach to decision tree induction.

## 8.2.1 Decision Tree Induction

During the late 1970s and early 1980s, J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as **ID3** (Iterative Dichotomiser). This work expanded on earlier work on *concept learning systems*, described by E. B. Hunt, J. Marin, and P. T. Stone. Quinlan later presented **C4.5** (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared. In 1984, a group of statisticians (L. Breiman, J. Friedman, R. Olshen, and C. Stone) published the book *Classification and Regression Trees* (**CART**), which described the generation of binary decision trees. ID3 and CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples. These two cornerstone algorithms spawned a flurry of work on decision tree induction.

ID3, C4.5, and CART adopt a greedy (i.e., nonbacktracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built. A basic decision tree algorithm is summarized in Figure 8.3. At first glance, the algorithm may appear long, but fear not! It is quite straightforward. The strategy is as follows.

- The algorithm is called with three parameters: *D*, *attribute\_list*, and *Attribute\_selection\_method*. We refer to *D* as a data partition. Initially, it is the complete set of training tuples and their associated class labels. The parameter *attribute\_list* is a list of attributes describing the tuples. *Attribute\_selection\_method* specifies a heuristic procedure for selecting the attribute that “best” discriminates the given tuples according to class. This procedure employs an attribute selection measure such as information gain or the Gini index. Whether the tree is strictly binary is generally driven by the attribute selection measure. Some attribute selection measures, such as the Gini index, enforce the resulting tree to be binary. Others, like information gain, do not, therein allowing multiway splits (i.e., two or more branches to be grown from a node).
- The tree starts as a single node, *N*, representing the training tuples in *D* (step 1).<sup>3</sup>

---

<sup>3</sup>The partition of class-labeled training tuples at node *N* is the set of tuples that follow a path from the root of the tree to node *N* when being processed by the tree. This set is sometimes referred to in the literature as the *family* of tuples at node *N*. We have referred to this set as the “tuples represented at node *N*,” “the tuples that reach node *N*,” or simply “the tuples at node *N*.” Rather than storing the actual tuples at a node, most implementations store pointers to these tuples.

**Algorithm: Generate\_decision\_tree.** Generate a decision tree from the training tuples of data partition,  $D$ .

**Input:**

- Data partition,  $D$ , which is a set of training tuples and their associated class labels;
- *attribute\_list*, the set of candidate attributes;
- *Attribute\_selection\_method*, a procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes. This criterion consists of a *splitting\_attribute* and, possibly, either a *split-point* or *splitting\_subset*.

**Output:** A decision tree.

**Method:**

- (1) create a node  $N$ ;
- (2) **if** tuples in  $D$  are all of the same class,  $C$ , **then**
- (3)     return  $N$  as a leaf node labeled with the class  $C$ ;
- (4) **if** *attribute\_list* is empty **then**
- (5)     return  $N$  as a leaf node labeled with the majority class in  $D$ ; // majority voting
- (6) apply **Attribute\_selection\_method**( $D$ , *attribute\_list*) to **find** the “best” *splitting\_criterion*;
- (7) label node  $N$  with *splitting\_criterion*;
- (8) **if** *splitting\_attribute* is discrete-valued **and**  
       multiway splits allowed **then** // not restricted to binary trees
- (9)     *attribute\_list*  $\leftarrow$  *attribute\_list* – *splitting\_attribute*; // remove *splitting\_attribute*
- (10) **for each** outcome  $j$  of *splitting\_criterion*  
       // partition the tuples and grow subtrees for each partition
- (11)     let  $D_j$  be the set of data tuples in  $D$  satisfying outcome  $j$ ; // a partition
- (12)     **if**  $D_j$  is empty **then**
- (13)         attach a leaf labeled with the majority class in  $D$  to node  $N$ ;
- (14)     **else** attach the node returned by **Generate\_decision\_tree**( $D_j$ , *attribute\_list*) to node  $N$ ;
- endfor**
- (15) return  $N$ ;

---

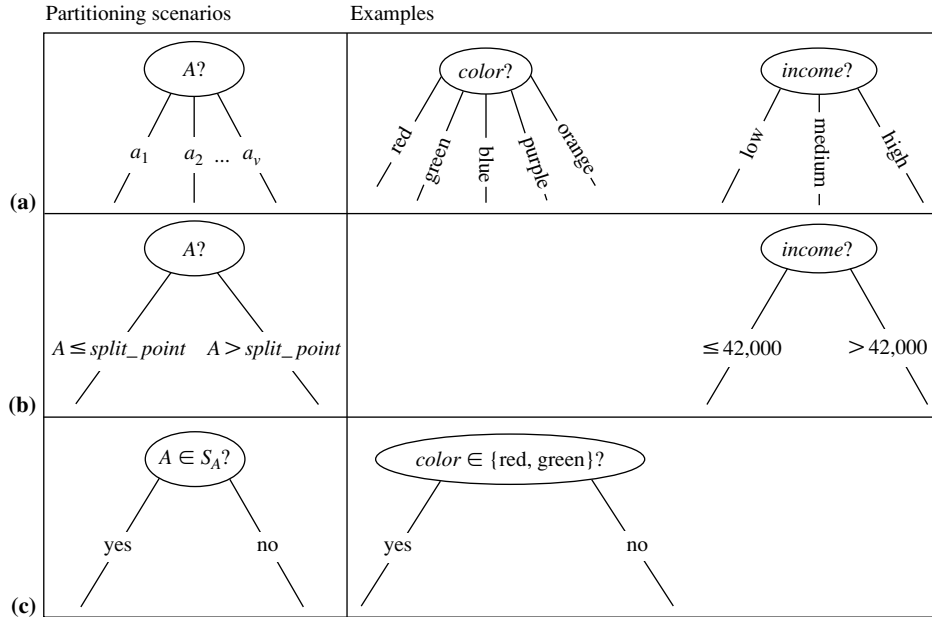
**Figure 8.3** Basic algorithm for inducing a decision tree from training tuples.

- If the tuples in  $D$  are all of the same class, then node  $N$  becomes a leaf and is labeled with that class (steps 2 and 3). Note that steps 4 and 5 are terminating conditions. All terminating conditions are explained at the end of the algorithm.
- Otherwise, the algorithm calls *Attribute\_selection\_method* to determine the **splitting criterion**. The splitting criterion tells us which attribute to test at node  $N$  by determining the “best” way to separate or partition the tuples in  $D$  into individual classes (step 6). The splitting criterion also tells us which branches to grow from node  $N$  with respect to the outcomes of the chosen test. More specifically, the splitting criterion indicates the **splitting attribute** and may also indicate either a **split-point** or a **splitting subset**. The splitting criterion is determined so that, ideally, the resulting

partitions at each branch are as “pure” as possible. A partition is **pure** if all the tuples in it belong to the same class. In other words, if we split up the tuples in  $D$  according to the mutually exclusive outcomes of the splitting criterion, we hope for the resulting partitions to be as pure as possible.

- The node  $N$  is labeled with the splitting criterion, which serves as a test at the node (step 7). A branch is grown from node  $N$  for each of the outcomes of the splitting criterion. The tuples in  $D$  are partitioned accordingly (steps 10 to 11). There are three possible scenarios, as illustrated in Figure 8.4. Let  $A$  be the splitting attribute.  $A$  has  $v$  distinct values,  $\{a_1, a_2, \dots, a_v\}$ , based on the training data.

- A is discrete-valued:* In this case, the outcomes of the test at node  $N$  correspond directly to the known values of  $A$ . A branch is created for each known value,  $a_j$ , of  $A$  and labeled with that value (Figure 8.4a). Partition  $D_j$  is the subset of class-labeled tuples in  $D$  having value  $a_j$  of  $A$ . Because all the tuples in a



**Figure 8.4** This figure shows three possibilities for partitioning tuples based on the splitting criterion, each with examples. Let  $A$  be the splitting attribute. (a) If  $A$  is discrete-valued, then one branch is grown for each known value of  $A$ . (b) If  $A$  is continuous-valued, then two branches are grown, corresponding to  $A \leq \text{split\_point}$  and  $A > \text{split\_point}$ . (c) If  $A$  is discrete-valued and a binary tree must be produced, then the test is of the form  $A \in S_A$ , where  $S_A$  is the splitting subset for  $A$ .

given partition have the same value for  $A$ ,  $A$  need not be considered in any future partitioning of the tuples. Therefore, it is removed from *attribute\_list* (steps 8 and 9).

2. *A is continuous-valued*: In this case, the test at node  $N$  has two possible outcomes, corresponding to the conditions  $A \leq \textit{split\_point}$  and  $A > \textit{split\_point}$ , respectively, where *split\_point* is the split-point returned by *Attribute\_selection\_method* as part of the splitting criterion. (In practice, the split-point,  $a$ , is often taken as the midpoint of two known adjacent values of  $A$  and therefore may not actually be a preexisting value of  $A$  from the training data.) Two branches are grown from  $N$  and labeled according to the previous outcomes (Figure 8.4b). The tuples are partitioned such that  $D_1$  holds the subset of class-labeled tuples in  $D$  for which  $A \leq \textit{split\_point}$ , while  $D_2$  holds the rest.
  3. *A is discrete-valued* and a *binary tree* must be produced (as dictated by the attribute selection measure or algorithm being used): The test at node  $N$  is of the form “ $A \in S_A?$ ,” where  $S_A$  is the splitting subset for  $A$ , returned by *Attribute\_selection\_method* as part of the splitting criterion. It is a subset of the known values of  $A$ . If a given tuple has value  $a_j$  of  $A$  and if  $a_j \in S_A$ , then the test at node  $N$  is satisfied. Two branches are grown from  $N$  (Figure 8.4c). By convention, the left branch out of  $N$  is labeled *yes* so that  $D_1$  corresponds to the subset of class-labeled tuples in  $D$  that satisfy the test. The right branch out of  $N$  is labeled *no* so that  $D_2$  corresponds to the subset of class-labeled tuples from  $D$  that do not satisfy the test.
- The algorithm uses the same process recursively to form a decision tree for the tuples at each resulting partition,  $D_j$ , of  $D$  (step 14).
  - The recursive partitioning stops only when any one of the following terminating conditions is true:
    1. All the tuples in partition  $D$  (represented at node  $N$ ) belong to the same class (steps 2 and 3).
    2. There are no remaining attributes on which the tuples may be further partitioned (step 4). In this case, **majority voting** is employed (step 5). This involves converting node  $N$  into a leaf and labeling it with the most common class in  $D$ . Alternatively, the class distribution of the node tuples may be stored.
    3. There are no tuples for a given branch, that is, a partition  $D_j$  is empty (step 12). In this case, a leaf is created with the majority class in  $D$  (step 13).
  - The resulting decision tree is returned (step 15).

The computational complexity of the algorithm given training set  $D$  is  $O(n \times |D| \times \log(|D|))$ , where  $n$  is the number of attributes describing the tuples in  $D$  and  $|D|$  is the number of training tuples in  $D$ . This means that the computational cost of growing a tree grows at most  $n \times |D| \times \log(|D|)$  with  $|D|$  tuples. The proof is left as an exercise for the reader.

**Incremental** versions of decision tree induction have also been proposed. When given new training data, these restructure the decision tree acquired from learning on previous training data, rather than relearning a new tree from scratch.

Differences in decision tree algorithms include how the attributes are selected in creating the tree (Section 8.2.2) and the mechanisms used for pruning (Section 8.2.3). The basic algorithm described earlier requires one pass over the training tuples in  $D$  for each level of the tree. This can lead to long training times and lack of available memory when dealing with large databases. Improvements regarding the scalability of decision tree induction are discussed in Section 8.2.4. Section 8.2.5 presents a visual interactive approach to decision tree construction. A discussion of strategies for extracting rules from decision trees is given in Section 8.4.2 regarding rule-based classification.

## 8.2.2 Attribute Selection Measures

An **attribute selection measure** is a heuristic for selecting the splitting criterion that “best” separates a given data partition,  $D$ , of class-labeled training tuples into individual classes. If we were to split  $D$  into smaller partitions according to the outcomes of the splitting criterion, ideally each partition would be pure (i.e., all the tuples that fall into a given partition would belong to the same class). Conceptually, the “best” splitting criterion is the one that most closely results in such a scenario. Attribute selection measures are also known as **splitting rules** because they determine how the tuples at a given node are to be split.

The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure<sup>4</sup> is chosen as the *splitting attribute* for the given tuples. If the splitting attribute is continuous-valued or if we are restricted to binary trees, then, respectively, either a *split point* or a *splitting subset* must also be determined as part of the splitting criterion. The tree node created for partition  $D$  is labeled with the splitting criterion, branches are grown for each outcome of the criterion, and the tuples are partitioned accordingly. This section describes three popular attribute selection measures—*information gain*, *gain ratio*, and *Gini index*.

The notation used herein is as follows. Let  $D$ , the data partition, be a training set of class-labeled tuples. Suppose the class label attribute has  $m$  distinct values defining  $m$  distinct classes,  $C_i$  (for  $i = 1, \dots, m$ ). Let  $C_{i,D}$  be the set of tuples of class  $C_i$  in  $D$ . Let  $|D|$  and  $|C_{i,D}|$  denote the number of tuples in  $D$  and  $C_{i,D}$ , respectively.

### Information Gain

ID3 uses **information gain** as its attribute selection measure. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages. Let node  $N$  represent or hold the tuples of partition  $D$ . The attribute with the highest information gain is chosen as the splitting attribute for node  $N$ . This attribute minimizes the information needed to classify the tuples in the

---

<sup>4</sup>Depending on the measure, either the highest or lowest score is chosen as the best (i.e., some measures strive to maximize while others strive to minimize).

resulting partitions and reflects the least randomness or “impurity” in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found.

The expected information needed to classify a tuple in  $D$  is given by

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i), \quad (8.1)$$

where  $p_i$  is the nonzero probability that an arbitrary tuple in  $D$  belongs to class  $C_i$  and is estimated by  $|C_{i,D}|/|D|$ . A log function to the base 2 is used, because the information is encoded in bits.  $Info(D)$  is just the average amount of information needed to identify the class label of a tuple in  $D$ . Note that, at this point, the information we have is based solely on the proportions of tuples of each class.  $Info(D)$  is also known as the **entropy** of  $D$ .

Now, suppose we were to partition the tuples in  $D$  on some attribute  $A$  having  $v$  distinct values,  $\{a_1, a_2, \dots, a_v\}$ , as observed from the training data. If  $A$  is discrete-valued, these values correspond directly to the  $v$  outcomes of a test on  $A$ . Attribute  $A$  can be used to split  $D$  into  $v$  partitions or subsets,  $\{D_1, D_2, \dots, D_v\}$ , where  $D_j$  contains those tuples in  $D$  that have outcome  $a_j$  of  $A$ . These partitions would correspond to the branches grown from node  $N$ . Ideally, we would like this partitioning to produce an exact classification of the tuples. That is, we would like for each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class).

How much more information would we still need (after the partitioning) to arrive at an exact classification? This amount is measured by

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j). \quad (8.2)$$

The term  $\frac{|D_j|}{|D|}$  acts as the weight of the  $j$ th partition.  $Info_A(D)$  is the expected information required to classify a tuple from  $D$  based on the partitioning by  $A$ . The smaller the expected information (still) required, the greater the purity of the partitions.

Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on  $A$ ). That is,

$$Gain(A) = Info(D) - Info_A(D). \quad (8.3)$$

In other words,  $Gain(A)$  tells us how much would be gained by branching on  $A$ . It is the expected reduction in the information requirement caused by knowing the value of  $A$ . The attribute  $A$  with the highest information gain,  $Gain(A)$ , is chosen as the splitting attribute at node  $N$ . This is equivalent to saying that we want to partition on the attribute  $A$  that would do the “best classification,” so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum  $Info_A(D)$ ).



**Table 8.1** Class-Labeled Training Tuples from the *AllElectronics* Customer Database

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

**Example 8.1 Induction of a decision tree using information gain.** Table 8.1 presents a training set,  $D$ , of class-labeled tuples randomly selected from the *AllElectronics* customer database. (The data are adapted from Quinlan [Qui86]. In this example, each attribute is discrete-valued. Continuous-valued attributes have been generalized.) The class label attribute, *buys\_computer*, has two distinct values (namely,  $\{yes, no\}$ ); therefore, there are two distinct classes (i.e.,  $m = 2$ ). Let class  $C_1$  correspond to *yes* and class  $C_2$  correspond to *no*. There are nine tuples of class *yes* and five tuples of class *no*. A (root) node  $N$  is created for the tuples in  $D$ . To find the splitting criterion for these tuples, we must compute the information gain of each attribute. We first use Eq. (8.1) to compute the expected information needed to classify a tuple in  $D$ :

$$Info(D) = -\frac{9}{14} \log_2 \left( \frac{9}{14} \right) - \frac{5}{14} \log_2 \left( \frac{5}{14} \right) = 0.940 \text{ bits.}$$

Next, we need to compute the expected information requirement for each attribute. Let's start with the attribute *age*. We need to look at the distribution of *yes* and *no* tuples for each category of *age*. For the *age* category "youth," there are two *yes* tuples and three *no* tuples. For the category "middle\_aged," there are four *yes* tuples and zero *no* tuples. For the category "senior," there are three *yes* tuples and two *no* tuples. Using Eq. (8.2), the expected information needed to classify a tuple in  $D$  if the tuples are partitioned according to *age* is

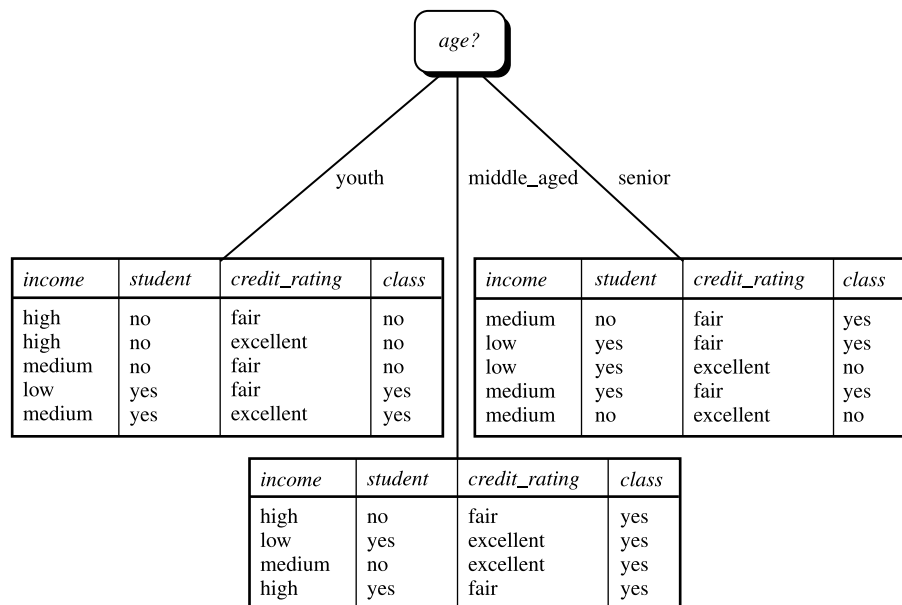
$$Info_{age}(D) = \frac{5}{14} \times \left( -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right)$$

$$\begin{aligned}
& + \frac{4}{14} \times \left( -\frac{4}{4} \log_2 \frac{4}{4} \right) \\
& + \frac{5}{14} \times \left( -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) \\
& = 0.694 \text{ bits.}
\end{aligned}$$

Hence, the gain in information from such a partitioning would be

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

Similarly, we can compute  $Gain(income) = 0.029$  bits,  $Gain(student) = 0.151$  bits, and  $Gain(credit\_rating) = 0.048$  bits. Because  $age$  has the highest information gain, it is selected as the splitting attribute. Node  $N$  is labeled with  $age$ , and branches are grown for each of the attribute's values. The tuples are then partitioned accordingly, as shown in Figure 8.5. Notice that the tuples falling into the partition for  $age = middle\_aged$  all belong to the same class. Because they all belong to class “yes,” a leaf should therefore be created at the end of this branch and labeled “yes.” The final decision tree returned by the algorithm was shown earlier in Figure 8.2. ■



**Figure 8.5** The attribute  $age$  has the highest information gain and therefore becomes the splitting attribute at the root node of the decision tree. Branches are grown for each outcome of  $age$ . The tuples are shown partitioned accordingly.

“But how can we compute the information gain of an attribute that is continuous-valued, unlike in the example?” Suppose, instead, that we have an attribute  $A$  that is continuous-valued, rather than discrete-valued. (For example, suppose that instead of the discretized version of *age* from the example, we have the raw values for this attribute.) For such a scenario, we must determine the “best” **split-point** for  $A$ , where the split-point is a threshold on  $A$ .

We first sort the values of  $A$  in increasing order. Typically, the midpoint between each pair of adjacent values is considered as a possible split-point. Therefore, given  $v$  values of  $A$ , then  $v - 1$  possible splits are evaluated. For example, the midpoint between the values  $a_i$  and  $a_{i+1}$  of  $A$  is

$$\frac{a_i + a_{i+1}}{2}. \quad (8.4)$$

If the values of  $A$  are sorted in advance, then determining the best split for  $A$  requires only one pass through the values. For each possible split-point for  $A$ , we evaluate  $\text{Info}_A(D)$ , where the number of partitions is two, that is,  $v = 2$  (or  $j = 1, 2$ ) in Eq. (8.2). The point with the minimum expected information requirement for  $A$  is selected as the *split-point* for  $A$ .  $D_1$  is the set of tuples in  $D$  satisfying  $A \leq \text{split-point}$ , and  $D_2$  is the set of tuples in  $D$  satisfying  $A > \text{split-point}$ .

## Gain Ratio

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. For example, consider an attribute that acts as a unique identifier such as *product.ID*. A split on *product.ID* would result in a large number of partitions (as many as there are values), each one containing just one tuple. Because each partition is pure, the information required to classify data set  $D$  based on this partitioning would be  $\text{Info}_{\text{product.ID}}(D) = 0$ . Therefore, the information gained by partitioning on this attribute is maximal. Clearly, such a partitioning is useless for classification.

C4.5, a successor of ID3, uses an extension to information gain known as *gain ratio*, which attempts to overcome this bias. It applies a kind of normalization to information gain using a “split information” value defined analogously with  $\text{Info}(D)$  as

$$\text{SplitInfo}_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right). \quad (8.5)$$

This value represents the potential information generated by splitting the training data set,  $D$ , into  $v$  partitions, corresponding to the  $v$  outcomes of a test on attribute  $A$ . Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in  $D$ . It differs from information gain, which measures the information with respect to classification that is acquired based on the

same partitioning. The gain ratio is defined as

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}_A(D)}. \quad (8.6)$$

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that as the split information approaches 0, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large—at least as great as the average gain over all tests examined.

**Example 8.2 Computation of gain ratio for the attribute *income*.** A test on *income* splits the data of Table 8.1 into three partitions, namely *low*, *medium*, and *high*, containing four, six, and four tuples, respectively. To compute the gain ratio of *income*, we first use Eq. (8.5) to obtain

$$\begin{aligned} \text{SplitInfo}_{\text{income}}(D) &= -\frac{4}{14} \times \log_2 \left( \frac{4}{14} \right) - \frac{6}{14} \times \log_2 \left( \frac{6}{14} \right) - \frac{4}{14} \times \log_2 \left( \frac{4}{14} \right) \\ &= 1.557. \end{aligned}$$

From Example 8.1, we have  $\text{Gain}(\text{income}) = 0.029$ . Therefore,  $\text{GainRatio}(\text{income}) = 0.029/1.557 = 0.019$ . ■

## Gini Index

The Gini index is used in CART. Using the notation previously described, the Gini index measures the impurity of  $D$ , a data partition or set of training tuples, as

$$\text{Gini}(D) = 1 - \sum_{i=1}^m p_i^2, \quad (8.7)$$

where  $p_i$  is the probability that a tuple in  $D$  belongs to class  $C_i$  and is estimated by  $|C_{i,D}|/|D|$ . The sum is computed over  $m$  classes.

The Gini index considers a binary split for each attribute. Let's first consider the case where  $A$  is a discrete-valued attribute having  $v$  distinct values,  $\{a_1, a_2, \dots, a_v\}$ , occurring in  $D$ . To determine the best binary split on  $A$ , we examine all the possible subsets that can be formed using known values of  $A$ . Each subset,  $S_A$ , can be considered as a binary test for attribute  $A$  of the form " $A \in S_A$ ?" Given a tuple, this test is satisfied if the value of  $A$  for the tuple is among the values listed in  $S_A$ . If  $A$  has  $v$  possible values, then there are  $2^v$  possible subsets. For example, if *income* has three possible values, namely *low*, *medium*, *high*, then the possible subsets are *{low, medium, high}*, *{low, medium}*, *{low, high}*, *{medium, high}*, *{low}*, *{medium}*, *{high}*, and *{}*. We exclude the power set, *{low, medium, high}*, and the empty set from consideration since, conceptually, they do not represent a split. Therefore, there are  $2^v - 2$  possible ways to form two partitions of the data,  $D$ , based on a binary split on  $A$ .

When considering a binary split, we compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on  $A$  partitions  $D$  into  $D_1$  and  $D_2$ , the Gini index of  $D$  given that partitioning is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2). \quad (8.8)$$

For each attribute, each of the possible binary splits is considered. For a discrete-valued attribute, the subset that gives the minimum Gini index for that attribute is selected as its splitting subset.

For continuous-valued attributes, each possible split-point must be considered. The strategy is similar to that described earlier for information gain, where the midpoint between each pair of (sorted) adjacent values is taken as a possible split-point. The point giving the minimum Gini index for a given (continuous-valued) attribute is taken as the split-point of that attribute. Recall that for a possible split-point of  $A$ ,  $D_1$  is the set of tuples in  $D$  satisfying  $A \leq \text{split\_point}$ , and  $D_2$  is the set of tuples in  $D$  satisfying  $A > \text{split\_point}$ .

The reduction in impurity that would be incurred by a binary split on a discrete- or continuous-valued attribute  $A$  is

$$\Delta Gini(A) = Gini(D) - Gini_A(D). \quad (8.9)$$

The attribute that maximizes the reduction in impurity (or, equivalently, has the minimum Gini index) is selected as the splitting attribute. This attribute and either its splitting subset (for a discrete-valued splitting attribute) or split-point (for a continuous-valued splitting attribute) together form the splitting criterion.

**Example 8.3 Induction of a decision tree using the Gini index.** Let  $D$  be the training data shown earlier in Table 8.1, where there are nine tuples belonging to the class *buys\_computer* = *yes* and the remaining five tuples belong to the class *buys\_computer* = *no*. A (root) node  $N$  is created for the tuples in  $D$ . We first use Eq. (8.7) for the Gini index to compute the impurity of  $D$ :

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459.$$

To find the splitting criterion for the tuples in  $D$ , we need to compute the Gini index for each attribute. Let's start with the attribute *income* and consider each of the possible splitting subsets. Consider the subset  $\{low, medium\}$ . This would result in 10 tuples in partition  $D_1$  satisfying the condition "*income*  $\in \{low, medium\}$ ." The remaining four tuples of  $D$  would be assigned to partition  $D_2$ . The Gini index value computed based on

this partitioning is

$$\begin{aligned}
 & \text{Gini}_{\text{income} \in \{\text{low}, \text{medium}\}}(D) \\
 &= \frac{10}{14} \text{Gini}(D_1) + \frac{4}{14} \text{Gini}(D_2) \\
 &= \frac{10}{14} \left( 1 - \left( \frac{7}{10} \right)^2 - \left( \frac{3}{10} \right)^2 \right) + \frac{4}{14} \left( 1 - \left( \frac{2}{4} \right)^2 - \left( \frac{2}{4} \right)^2 \right) \\
 &= 0.443 \\
 &= \text{Gini}_{\text{income} \in \{\text{high}\}}(D).
 \end{aligned}$$

Similarly, the Gini index values for splits on the remaining subsets are 0.458 (for the subsets  $\{\text{low}, \text{high}\}$  and  $\{\text{medium}\}$ ) and 0.450 (for the subsets  $\{\text{medium}, \text{high}\}$  and  $\{\text{low}\}$ ). Therefore, the best binary split for attribute *income* is on  $\{\text{low}, \text{medium}\}$  (or  $\{\text{high}\}$ ) because it minimizes the Gini index. Evaluating *age*, we obtain  $\{\text{youth}, \text{senior}\}$  (or  $\{\text{middle\_aged}\}$ ) as the best split for *age* with a Gini index of 0.375; the attributes *student* and *credit\_rating* are both binary, with Gini index values of 0.367 and 0.429, respectively.

The attribute *age* and splitting subset  $\{\text{youth}, \text{senior}\}$  therefore give the minimum Gini index overall, with a reduction in impurity of  $0.459 - 0.357 = 0.102$ . The binary split “*age*  $\in \{\text{youth}, \text{senior}\}?$ ” results in the maximum reduction in impurity of the tuples in *D* and is returned as the splitting criterion. Node *N* is labeled with the criterion, two branches are grown from it, and the tuples are partitioned accordingly. ■

## Other Attribute Selection Measures

This section on attribute selection measures was not intended to be exhaustive. We have shown three measures that are commonly used for building decision trees. These measures are not without their biases. Information gain, as we saw, is biased toward multivalued attributes. Although the gain ratio adjusts for this bias, it tends to prefer unbalanced splits in which one partition is much smaller than the others. The Gini index is biased toward multivalued attributes and has difficulty when the number of classes is large. It also tends to favor tests that result in equal-size partitions and purity in both partitions. Although biased, these measures give reasonably good results in practice.

Many other attribute selection measures have been proposed. CHAID, a decision tree algorithm that is popular in marketing, uses an attribute selection measure that is based on the statistical  $\chi^2$  test for independence. Other measures include C-SEP (which performs better than information gain and the Gini index in certain cases) and G-statistic (an information theoretic measure that is a close approximation to  $\chi^2$  distribution).

Attribute selection measures based on the **Minimum Description Length (MDL)** principle have the least bias toward multivalued attributes. MDL-based measures use encoding techniques to define the “best” decision tree as the one that requires the fewest number of bits to both (1) encode the tree and (2) encode the exceptions to the tree

(i.e., cases that are not correctly classified by the tree). Its main idea is that the simplest of solutions is preferred.

Other attribute selection measures consider **multivariate splits** (i.e., where the partitioning of tuples is based on a *combination* of attributes, rather than on a single attribute). The CART system, for example, can find multivariate splits based on a linear combination of attributes. Multivariate splits are a form of **attribute construction**, where new attributes are created based on the existing ones. (Attribute construction was also discussed in Chapter 3, as a form of data transformation.) These other measures mentioned here are beyond the scope of this book. Additional references are given in the bibliographic notes at the end of this chapter (Section 8.9).

*“Which attribute selection measure is the best?”* All measures have some bias. It has been shown that the time complexity of decision tree induction generally increases exponentially with tree height. Hence, measures that tend to produce shallower trees (e.g., with multiway rather than binary splits, and that favor more balanced splits) may be preferred. However, some studies have found that shallow trees tend to have a large number of leaves and higher error rates. Despite several comparative studies, no one attribute selection measure has been found to be significantly superior to others. Most measures give quite good results.

### 8.2.3 Tree Pruning

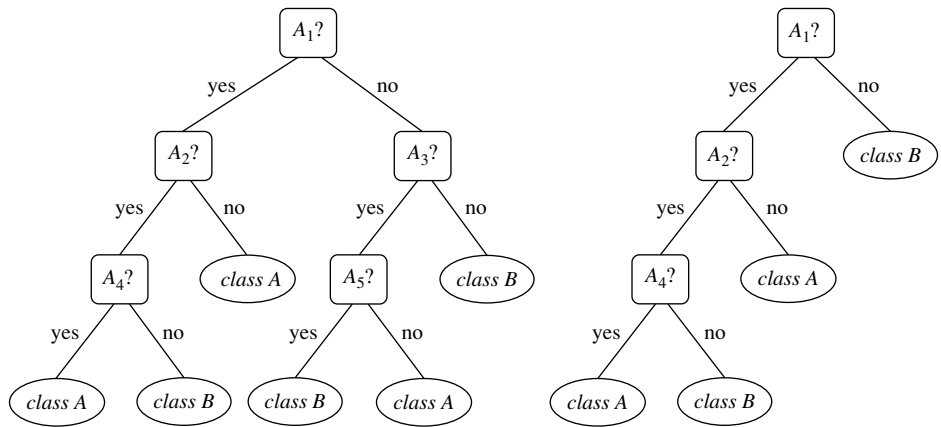
When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of *overfitting* the data. Such methods typically use statistical measures to remove the least-reliable branches. An unpruned tree and a pruned version of it are shown in Figure 8.6. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data (i.e., of previously unseen tuples) than unpruned trees.

*“How does tree pruning work?”* There are two common approaches to tree pruning: *prepruning* and *postpruning*.

In the **prepruning** approach, a tree is “pruned” by halting its construction early (e.g., by deciding not to further split or partition the subset of training tuples at a given node). Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples.

When constructing a tree, measures such as statistical significance, information gain, Gini index, and so on, can be used to assess the goodness of a split. If partitioning the tuples at a node would result in a split that falls below a prespecified threshold, then further partitioning of the given subset is halted. There are difficulties, however, in choosing an appropriate threshold. High thresholds could result in oversimplified trees, whereas low thresholds could result in very little simplification.

The second and more common approach is **postpruning**, which removes subtrees from a “fully grown” tree. A subtree at a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced. For example, notice the subtree at node “ $A_3$ ?” in the unpruned



**Figure 8.6** An unpruned decision tree and a pruned version of it.

tree of Figure 8.6. Suppose that the most common class within this subtree is “class B.” In the pruned version of the tree, the subtree in question is pruned by replacing it with the leaf “class B.”

The **cost complexity** pruning algorithm used in CART is an example of the postpruning approach. This approach considers the cost complexity of a tree to be a function of the number of leaves in the tree and the error rate of the tree (where the **error rate** is the percentage of tuples misclassified by the tree). It starts from the bottom of the tree. For each internal node,  $N$ , it computes the cost complexity of the subtree at  $N$ , and the cost complexity of the subtree at  $N$  if it were to be pruned (i.e., replaced by a leaf node). The two values are compared. If pruning the subtree at node  $N$  would result in a smaller cost complexity, then the subtree is pruned. Otherwise, it is kept.

A **pruning set** of class-labeled tuples is used to estimate cost complexity. This set is independent of the training set used to build the unpruned tree and of any test set used for accuracy estimation. The algorithm generates a set of progressively pruned trees. In general, the smallest decision tree that minimizes the cost complexity is preferred.

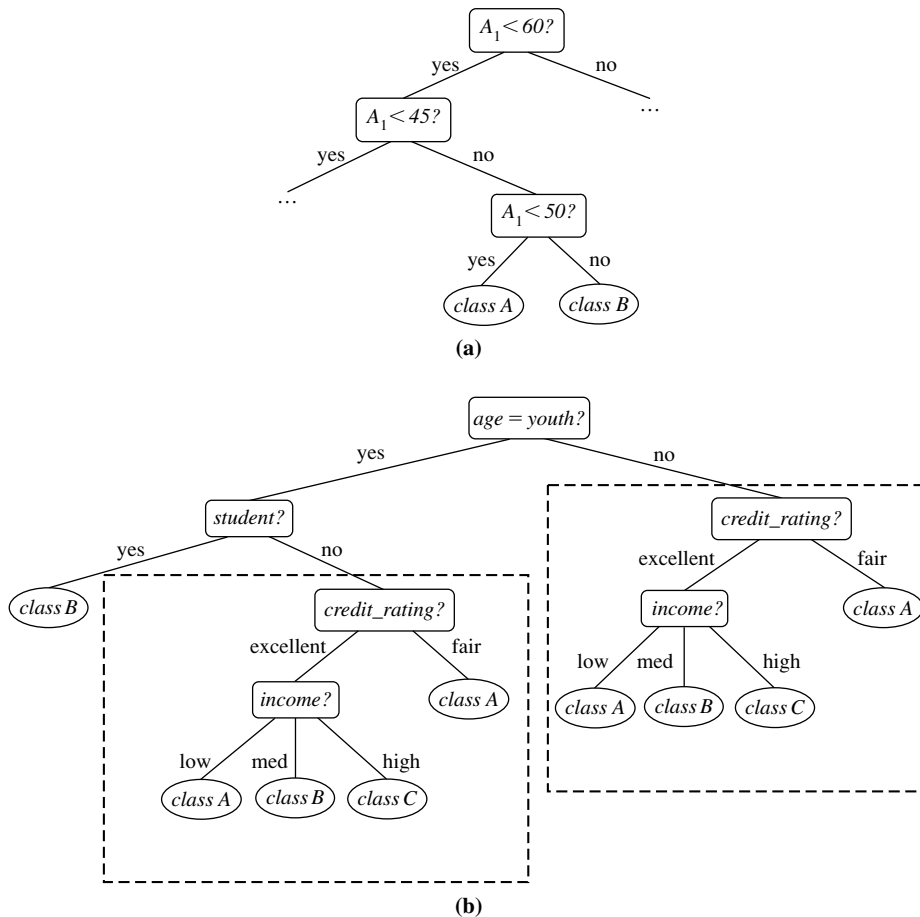
C4.5 uses a method called **pessimistic pruning**, which is similar to the cost complexity method in that it also uses error rate estimates to make decisions regarding subtree pruning. Pessimistic pruning, however, does not require the use of a prune set. Instead, it uses the training set to estimate error rates. Recall that an estimate of accuracy or error based on the training set is overly optimistic and, therefore, strongly biased. The pessimistic pruning method therefore adjusts the error rates obtained from the training set by adding a penalty, so as to counter the bias incurred.

Rather than pruning trees based on estimated error rates, we can prune trees based on the number of bits required to encode them. The “best” pruned tree is the one that minimizes the number of encoding bits. This method adopts the MDL principle, which was briefly introduced in Section 8.2.2. The basic idea is that the simplest solution is preferred. Unlike cost complexity pruning, it does not require an independent set of tuples.



Alternatively, prepruning and postpruning may be interleaved for a combined approach. Postpruning requires more computation than prepruning, yet generally leads to a more reliable tree. No single pruning method has been found to be superior over all others. Although some pruning methods do depend on the availability of additional data for pruning, this is usually not a concern when dealing with large databases.

Although pruned trees tend to be more compact than their unpruned counterparts, they may still be rather large and complex. Decision trees can suffer from *repetition* and *replication* (Figure 8.7), making them overwhelming to interpret. **Repetition** occurs when an attribute is repeatedly tested along a given branch of the tree (e.g., “age < 60?”



**Figure 8.7** An example of: (a) subtree **repetition**, where an attribute is repeatedly tested along a given branch of the tree (e.g., age) and (b) subtree **replication**, where duplicate subtrees exist within a tree (e.g., the subtree headed by the node “credit\_rating?”).

followed by “ $age < 45?$ ,” and so on). In **replication**, duplicate subtrees exist within the tree. These situations can impede the accuracy and comprehensibility of a decision tree. The use of multivariate splits (splits based on a combination of attributes) can prevent these problems. Another approach is to use a different form of knowledge representation, such as rules, instead of decision trees. This is described in Section 8.4.2, which shows how a *rule-based classifier* can be constructed by extracting IF-THEN rules from a decision tree.

### 8.2.4 Scalability and Decision Tree Induction

*“What if  $D$ , the disk-resident training set of class-labeled tuples, does not fit in memory? In other words, how scalable is decision tree induction?”* The efficiency of existing decision tree algorithms, such as ID3, C4.5, and CART, has been well established for relatively small data sets. Efficiency becomes an issue of concern when these algorithms are applied to the mining of very large real-world databases. The pioneering decision tree algorithms that we have discussed so far have the restriction that the training tuples should reside *in memory*.

In data mining applications, very large training sets of millions of tuples are common. Most often, the training data will not fit in memory! Therefore, decision tree construction becomes inefficient due to swapping of the training tuples in and out of main and cache memories. More scalable approaches, capable of handling training data that are too large to fit in memory, are required. Earlier strategies to “save space” included discretizing continuous-valued attributes and sampling data at each node. These techniques, however, still assume that the training set can fit in memory.

Several scalable decision tree induction methods have been introduced in recent studies. RainForest, for example, adapts to the amount of main memory available and applies to any decision tree induction algorithm. The method maintains an **AVC-set** (where “AVC” stands for “*Attribute-Value, Classlabel*”) for each attribute, at each tree node, describing the training tuples at the node. The AVC-set of an attribute  $A$  at node  $N$  gives the class label counts for each value of  $A$  for the tuples at  $N$ . Figure 8.8 shows AVC-sets for the tuple data of Table 8.1. The set of all AVC-sets at a node  $N$  is the **AVC-group** of  $N$ . The size of an AVC-set for attribute  $A$  at node  $N$  depends only on the number of distinct values of  $A$  and the number of classes in the set of tuples at  $N$ . Typically, this size should fit in memory, even for real-world data. RainForest also has techniques, however, for handling the case where the AVC-group does not fit in memory. Therefore, the method has high scalability for decision tree induction in very large data sets.

BOAT (Bootstrapped Optimistic Algorithm for Tree construction) is a decision tree algorithm that takes a completely different approach to scalability—it is not based on the use of any special data structures. Instead, it uses a statistical technique known as “bootstrapping” (Section 8.5.4) to create several smaller samples (or subsets) of the given training data, each of which fits in memory. Each subset is used to construct a tree, resulting in several trees. The trees are examined and used to construct a new tree,  $T'$ , that turns out to be “very close” to the tree that would have been generated if all the original training data had fit in memory.

<i>age</i>	<i>buys_computer</i>	
	yes	no
youth	2	3
middle_aged	4	0
senior	3	2

<i>income</i>	<i>buys_computer</i>	
	yes	no
low	3	1
medium	4	2
high	2	2

<i>student</i>	<i>buys_computer</i>	
	yes	no
yes	6	1
no	3	4

<i>credit_rating</i>	<i>buys_computer</i>	
	yes	no
fair	6	2
excellent	3	3

**Figure 8.8** The use of data structures to hold aggregate information regarding the training data (e.g., these AVC-sets describing Table 8.1’s data) are one approach to improving the scalability of decision tree induction.

BOAT can use any attribute selection measure that selects binary splits and that is based on the notion of purity of partitions such as the Gini index. BOAT uses a lower bound on the attribute selection measure to detect if this “very good” tree,  $T'$ , is different from the “real” tree,  $T$ , that would have been generated using all of the data. It refines  $T'$  to arrive at  $T$ .

BOAT usually requires only two scans of  $D$ . This is quite an improvement, even in comparison to traditional decision tree algorithms (e.g., the basic algorithm in Figure 8.3), which require one scan per tree level! BOAT was found to be two to three times faster than RainForest, while constructing exactly the same tree. An additional advantage of BOAT is that it can be used for incremental updates. That is, BOAT can take new insertions and deletions for the training data and update the decision tree to reflect these changes, without having to reconstruct the tree from scratch.

### 8.2.5 Visual Mining for Decision Tree Induction

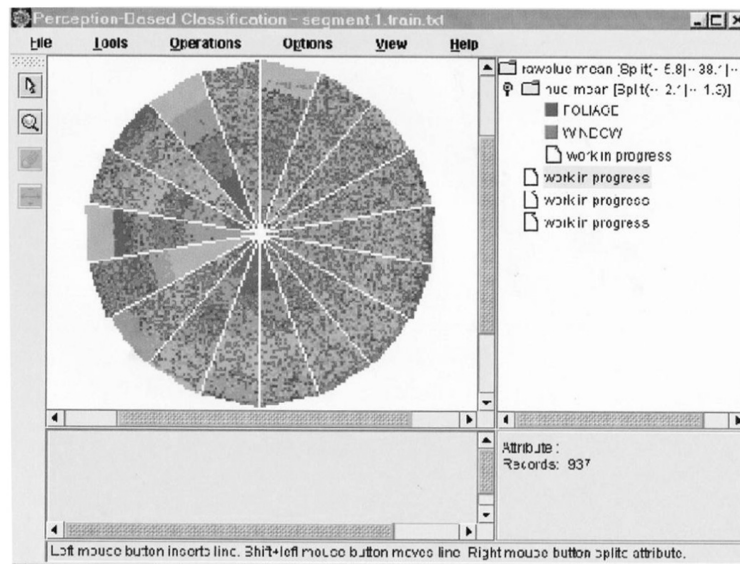
“Are there any interactive approaches to decision tree induction that allow us to visualize the data and the tree as it is being constructed? Can we use any knowledge of our data to help in building the tree?” In this section, you will learn about an approach to decision tree induction that supports these options. **Perception-based classification (PBC)** is an interactive approach based on multidimensional visualization techniques and allows the user to incorporate background knowledge about the data when building a decision tree. By visually interacting with the data, the user is also likely to develop a deeper understanding of the data. The resulting trees tend to be smaller than those built using traditional decision tree induction methods and so are easier to interpret, while achieving about the same accuracy.

“How can the data be visualized to support interactive decision tree construction?” PBC uses a pixel-oriented approach to view multidimensional data with its class label

information. The circle segments approach is adapted, which maps  $d$ -dimensional data objects to a circle that is partitioned into  $d$  segments, each representing one attribute (Section 2.3.1). Here, an attribute value of a data object is mapped to one colored pixel, reflecting the object's class label. This mapping is done for each attribute–value pair of each data object. Sorting is done for each attribute to determine the arrangement order within a segment. For example, attribute values within a given segment may be organized so as to display homogeneous (with respect to class label) regions within the same attribute value. The amount of training data that can be visualized at one time is approximately determined by the product of the number of attributes and the number of data objects.

The PBC system displays a split screen, consisting of a Data Interaction window and a Knowledge Interaction window (Figure 8.9). The Data Interaction window displays the circle segments of the data under examination, while the Knowledge Interaction window displays the decision tree constructed so far. Initially, the complete training set is visualized in the Data Interaction window, while the Knowledge Interaction window displays an empty decision tree.

Traditional decision tree algorithms allow only binary splits for numeric attributes. PBC, however, allows the user to specify multiple split-points, resulting in multiple branches to be grown from a single tree node.



**Figure 8.9** A screenshot of PBC, a system for interactive decision tree construction. Multidimensional training data are viewed as circle segments in the Data Interaction window (*left*). The Knowledge Interaction window (*right*) displays the current decision tree. *Source:* From Ankerst, Elsen, Ester, and Kriegel [AEEK99].

A tree is interactively constructed as follows. The user visualizes the multidimensional data in the Data Interaction window and selects a splitting attribute and one or more split-points. The current decision tree in the Knowledge Interaction window is expanded. The user selects a node of the decision tree. The user may either assign a class label to the node (which makes the node a leaf) or request the visualization of the training data corresponding to the node. This leads to a new visualization of every attribute except the ones used for splitting criteria on the same path from the root. The interactive process continues until a class has been assigned to each leaf of the decision tree.

The trees constructed with PBC were compared with trees generated by the CART, C4.5, and SPRINT algorithms from various data sets. The trees created with PBC were of comparable accuracy with the tree from the algorithmic approaches, yet were significantly smaller and, thus, easier to understand. Users can use their domain knowledge in building a decision tree, but also gain a deeper understanding of their data during the construction process.

## 8.3 Bayes Classification Methods

“What are Bayesian classifiers?” Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Bayesian classification is based on Bayes’ theorem, described next. Studies comparing classification algorithms have found a simple Bayesian classifier known as the *naïve Bayesian classifier* to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called *class-conditional independence*. It is made to simplify the computations involved and, in this sense, is considered “naïve.”

Section 8.3.1 reviews basic probability notation and Bayes’ theorem. In Section 8.3.2 you will learn how to do naïve Bayesian classification.

### 8.3.1 Bayes’ Theorem

Bayes’ theorem is named after Thomas Bayes, a nonconformist English clergyman who did early work in probability and decision theory during the 18th century. Let  $\mathbf{X}$  be a data tuple. In Bayesian terms,  $\mathbf{X}$  is considered “evidence.” As usual, it is described by measurements made on a set of  $n$  attributes. Let  $H$  be some hypothesis such as that the data tuple  $\mathbf{X}$  belongs to a specified class  $C$ . For classification problems, we want to determine  $P(H|\mathbf{X})$ , the probability that the hypothesis  $H$  holds given the “evidence” or observed data tuple  $\mathbf{X}$ . In other words, we are looking for the probability that tuple  $\mathbf{X}$  belongs to class  $C$ , given that we know the attribute description of  $\mathbf{X}$ .

$P(H|X)$  is the **posterior probability**, or a *posteriori probability*, of  $H$  conditioned on  $X$ . For example, suppose our world of data tuples is confined to customers described by the attributes *age* and *income*, respectively, and that  $X$  is a 35-year-old customer with an income of \$40,000. Suppose that  $H$  is the hypothesis that our customer will buy a computer. Then  $P(H|X)$  reflects the probability that customer  $X$  will buy a computer given that we know the customer's age and income.

In contrast,  $P(H)$  is the **prior probability**, or a *priori probability*, of  $H$ . For our example, this is the probability that any given customer will buy a computer, regardless of age, income, or any other information, for that matter. The posterior probability,  $P(H|X)$ , is based on more information (e.g., customer information) than the prior probability,  $P(H)$ , which is independent of  $X$ .

Similarly,  $P(X|H)$  is the posterior probability of  $X$  conditioned on  $H$ . That is, it is the probability that a customer,  $X$ , is 35 years old and earns \$40,000, given that we know the customer will buy a computer.

$P(X)$  is the prior probability of  $X$ . Using our example, it is the probability that a person from our set of customers is 35 years old and earns \$40,000.

"How are these probabilities estimated?"  $P(H)$ ,  $P(X|H)$ , and  $P(X)$  may be estimated from the given data, as we shall see next. **Bayes' theorem** is useful in that it provides a way of calculating the posterior probability,  $P(H|X)$ , from  $P(H)$ ,  $P(X|H)$ , and  $P(X)$ . Bayes' theorem is

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}. \quad (8.10)$$

Now that we have that out of the way, in the next section, we will look at how Bayes' theorem is used in the naïve Bayesian classifier.

### 8.3.2 Naïve Bayesian Classification

The **naïve Bayesian** classifier, or **simple Bayesian** classifier, works as follows:

1. Let  $D$  be a training set of tuples and their associated class labels. As usual, each tuple is represented by an  $n$ -dimensional attribute vector,  $X = (x_1, x_2, \dots, x_n)$ , depicting  $n$  measurements made on the tuple from  $n$  attributes, respectively,  $A_1, A_2, \dots, A_n$ .
2. Suppose that there are  $m$  classes,  $C_1, C_2, \dots, C_m$ . Given a tuple,  $X$ , the classifier will predict that  $X$  belongs to the class having the highest posterior probability, conditioned on  $X$ . That is, the naïve Bayesian classifier predicts that tuple  $X$  belongs to the class  $C_i$  if and only if

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \leq j \leq m, j \neq i.$$

Thus, we maximize  $P(C_i|X)$ . The class  $C_i$  for which  $P(C_i|X)$  is maximized is called the *maximum posteriori hypothesis*. By Bayes' theorem (Eq. 8.10),

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}. \quad (8.11)$$

3. As  $P(\mathbf{X})$  is constant for all classes, only  $P(\mathbf{X}|C_i)P(C_i)$  needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is,  $P(C_1) = P(C_2) = \dots = P(C_m)$ , and we would therefore maximize  $P(\mathbf{X}|C_i)$ . Otherwise, we maximize  $P(\mathbf{X}|C_i)P(C_i)$ . Note that the class prior probabilities may be estimated by  $P(C_i) = |C_{i,D}|/|D|$ , where  $|C_{i,D}|$  is the number of training tuples of class  $C_i$  in  $D$ .
4. Given data sets with many attributes, it would be extremely computationally expensive to compute  $P(\mathbf{X}|C_i)$ . To reduce computation in evaluating  $P(\mathbf{X}|C_i)$ , the naïve assumption of **class-conditional independence** is made. This presumes that the attributes' values are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$\begin{aligned}
 P(\mathbf{X}|C_i) &= \prod_{k=1}^n P(x_k|C_i) \\
 &= P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i).
 \end{aligned} \tag{8.12}$$

We can easily estimate the probabilities  $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$  from the training tuples. Recall that here  $x_k$  refers to the value of attribute  $A_k$  for tuple  $\mathbf{X}$ . For each attribute, we look at whether the attribute is categorical or continuous-valued. For instance, to compute  $P(\mathbf{X}|C_i)$ , we consider the following:

- (a) If  $A_k$  is categorical, then  $P(x_k|C_i)$  is the number of tuples of class  $C_i$  in  $D$  having the value  $x_k$  for  $A_k$ , divided by  $|C_{i,D}|$ , the number of tuples of class  $C_i$  in  $D$ .
- (b) If  $A_k$  is continuous-valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean  $\mu$  and standard deviation  $\sigma$ , defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{8.13}$$

so that

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}). \tag{8.14}$$

These equations may appear daunting, but hold on! We need to compute  $\mu_{C_i}$  and  $\sigma_{C_i}$ , which are the mean (i.e., average) and standard deviation, respectively, of the values of attribute  $A_k$  for training tuples of class  $C_i$ . We then plug these two quantities into Eq. (8.13), together with  $x_k$ , to estimate  $P(x_k|C_i)$ .

For example, let  $\mathbf{X} = (35, \$40,000)$ , where  $A_1$  and  $A_2$  are the attributes *age* and *income*, respectively. Let the class label attribute be *buys\_computer*. The associated class label for  $\mathbf{X}$  is *yes* (i.e., *buys\_computer* = *yes*). Let's suppose that *age* has not been discretized and therefore exists as a continuous-valued attribute. Suppose that from the training set, we find that customers in  $D$  who buy a computer are

$38 \pm 12$  years of age. In other words, for attribute *age* and this class, we have  $\mu = 38$  years and  $\sigma = 12$ . We can plug these quantities, along with  $x_1 = 35$  for our tuple  $\mathbf{X}$ , into Eq. (8.13) to estimate  $P(\text{age} = 35 | \text{buys\_computer} = \text{yes})$ . For a quick review of mean and standard deviation calculations, please see Section 2.2.

5. To predict the class label of  $\mathbf{X}$ ,  $P(\mathbf{X}|C_i)P(C_i)$  is evaluated for each class  $C_i$ . The classifier predicts that the class label of tuple  $\mathbf{X}$  is the class  $C_i$  if and only if

$$P(\mathbf{X}|C_i)P(C_i) > P(\mathbf{X}|C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i. \quad (8.15)$$

In other words, the predicted class label is the class  $C_i$  for which  $P(\mathbf{X}|C_i)P(C_i)$  is the maximum.

“How effective are Bayesian classifiers?” Various empirical studies of this classifier in comparison to decision tree and neural network classifiers have found it to be comparable in some domains. In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case, owing to inaccuracies in the assumptions made for its use, such as class-conditional independence, and the lack of available probability data.

Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers that do not explicitly use Bayes’ theorem. For example, under certain assumptions, it can be shown that many neural network and curve-fitting algorithms output the *maximum posteriori* hypothesis, as does the naïve Bayesian classifier.

**Example 8.4 Predicting a class label using naïve Bayesian classification.** We wish to predict the class label of a tuple using naïve Bayesian classification, given the same training data as in Example 8.3 for decision tree induction. The training data were shown earlier in Table 8.1. The data tuples are described by the attributes *age*, *income*, *student*, and *credit\_rating*. The class label attribute, *buys\_computer*, has two distinct values (namely, {yes, no}). Let  $C_1$  correspond to the class *buys\_computer* = yes and  $C_2$  correspond to *buys\_computer* = no. The tuple we wish to classify is

$$\mathbf{X} = (\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})$$

We need to maximize  $P(\mathbf{X}|C_i)P(C_i)$ , for  $i = 1, 2$ .  $P(C_i)$ , the prior probability of each class, can be computed based on the training tuples:

$$P(\text{buys\_computer} = \text{yes}) = 9/14 = 0.643$$

$$P(\text{buys\_computer} = \text{no}) = 5/14 = 0.357$$

To compute  $P(\mathbf{X}|C_i)$ , for  $i = 1, 2$ , we compute the following conditional probabilities:

$$P(\text{age} = \text{youth} | \text{buys\_computer} = \text{yes}) = 2/9 = 0.222$$

$$P(\text{age} = \text{youth} | \text{buys\_computer} = \text{no}) = 3/5 = 0.600$$

$$P(\text{income} = \text{medium} | \text{buys\_computer} = \text{yes}) = 4/9 = 0.444$$

$$P(\text{income} = \text{medium} | \text{buys\_computer} = \text{no}) = 2/5 = 0.400$$

$$P(\text{student} = \text{yes} | \text{buys\_computer} = \text{yes}) = 6/9 = 0.667$$



$$P(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{no}) = 1/5 = 0.200$$

$$P(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{yes}) = 6/9 = 0.667$$

$$P(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{no}) = 2/5 = 0.400$$

Using these probabilities, we obtain

$$\begin{aligned} P(\mathbf{X} \mid \text{buys\_computer} = \text{yes}) &= P(\text{age} = \text{youth} \mid \text{buys\_computer} = \text{yes}) \\ &\quad \times P(\text{income} = \text{medium} \mid \text{buys\_computer} = \text{yes}) \\ &\quad \times P(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{yes}) \\ &\quad \times P(\text{credit\_rating} = \text{fair} \mid \text{buys\_computer} = \text{yes}) \\ &= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044. \end{aligned}$$

Similarly,

$$P(\mathbf{X} \mid \text{buys\_computer} = \text{no}) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

To find the class,  $C_i$ , that maximizes  $P(\mathbf{X} \mid C_i)P(C_i)$ , we compute

$$P(\mathbf{X} \mid \text{buys\_computer} = \text{yes})P(\text{buys\_computer} = \text{yes}) = 0.044 \times 0.643 = 0.028$$

$$P(\mathbf{X} \mid \text{buys\_computer} = \text{no})P(\text{buys\_computer} = \text{no}) = 0.019 \times 0.357 = 0.007$$

Therefore, the naïve Bayesian classifier predicts *buys\_computer = yes* for tuple  $\mathbf{X}$ . ■

“What if I encounter probability values of zero?” Recall that in Eq. (8.12), we estimate  $P(\mathbf{X} \mid C_i)$  as the product of the probabilities  $P(x_1 \mid C_i)$ ,  $P(x_2 \mid C_i)$ , ...,  $P(x_n \mid C_i)$ , based on the assumption of class-conditional independence. These probabilities can be estimated from the training tuples (step 4). We need to compute  $P(\mathbf{X} \mid C_i)$  for *each* class ( $i = 1, 2, \dots, m$ ) to find the class  $C_i$  for which  $P(\mathbf{X} \mid C_i)P(C_i)$  is the maximum (step 5). Let’s consider this calculation. For each attribute–value pair (i.e.,  $A_k = x_k$ , for  $k = 1, 2, \dots, n$ ) in tuple  $\mathbf{X}$ , we need to count the number of tuples having that attribute–value pair, per class (i.e., per  $C_i$ , for  $i = 1, \dots, m$ ). In Example 8.4, we have two classes ( $m = 2$ ), namely *buys\_computer = yes* and *buys\_computer = no*. Therefore, for the attribute–value pair *student = yes* of  $\mathbf{X}$ , say, we need two counts—the number of customers who are students and for which *buys\_computer = yes* (which contributes to  $P(\mathbf{X} \mid \text{buys\_computer} = \text{yes})$ ) and the number of customers who are students and for which *buys\_computer = no* (which contributes to  $P(\mathbf{X} \mid \text{buys\_computer} = \text{no})$ ).

But what if, say, there are no training tuples representing students for the class *buys\_computer = no*, resulting in  $P(\text{student} = \text{yes} \mid \text{buys\_computer} = \text{no}) = 0$ ? In other words, what happens if we should end up with a probability value of zero for some  $P(x_k \mid C_i)$ ? Plugging this zero value into Eq. (8.12) would return a zero probability for  $P(\mathbf{X} \mid C_i)$ , even though, without the zero probability, we may have ended up with a high probability, suggesting that  $\mathbf{X}$  belonged to class  $C_i$ ! A zero probability cancels the effects of all the other (posteriori) probabilities (on  $C_i$ ) involved in the product.

There is a simple trick to avoid this problem. We can assume that our training database,  $D$ , is so large that adding one to each count that we need would only make a negligible difference in the estimated probability value, yet would conveniently avoid the

case of probability values of zero. This technique for probability estimation is known as the **Laplacian correction** or **Laplace estimator**, named after Pierre Laplace, a French mathematician who lived from 1749 to 1827. If we have, say,  $q$  counts to which we each add one, then we must remember to add  $q$  to the corresponding denominator used in the probability calculation. We illustrate this technique in Example 8.5.

**Example 8.5 Using the Laplacian correction to avoid computing probability values of zero.** Suppose that for the class *buys\_computer = yes* in some training database,  $D$ , containing 1000 tuples, we have 0 tuples with *income = low*, 990 tuples with *income = medium*, and 10 tuples with *income = high*. The probabilities of these events, without the Laplacian correction, are 0, 0.990 (from 990/1000), and 0.010 (from 10/1000), respectively. Using the Laplacian correction for the three quantities, we pretend that we have 1 more tuple for each income-value pair. In this way, we instead obtain the following probabilities (rounded up to three decimal places):

$$\frac{1}{1003} = 0.001, \frac{991}{1003} = 0.988, \text{ and } \frac{11}{1003} = 0.011,$$

respectively. The “corrected” probability estimates are close to their “uncorrected” counterparts, yet the zero probability value is avoided. ■

## 8.4 Rule-Based Classification

In this section, we look at rule-based classifiers, where the learned model is represented as a set of IF-THEN rules. We first examine how such rules are used for classification (Section 8.4.1). We then study ways in which they can be generated, either from a decision tree (Section 8.4.2) or directly from the training data using a *sequential covering algorithm* (Section 8.4.3).

### 8.4.1 Using IF-THEN Rules for Classification

Rules are a good way of representing information or bits of knowledge. A **rule-based classifier** uses a set of IF-THEN rules for classification. An **IF-THEN** rule is an expression of the form

IF *condition* THEN *conclusion*.

An example is rule  $R_1$ ,

$R_1$ : IF *age = youth* AND *student = yes* THEN *buys\_computer = yes*.

The “IF” part (or left side) of a rule is known as the **rule antecedent** or **precondition**. The “THEN” part (or right side) is the **rule consequent**. In the rule antecedent, the condition consists of one or more *attribute tests* (e.g., *age = youth* and *student = yes*)