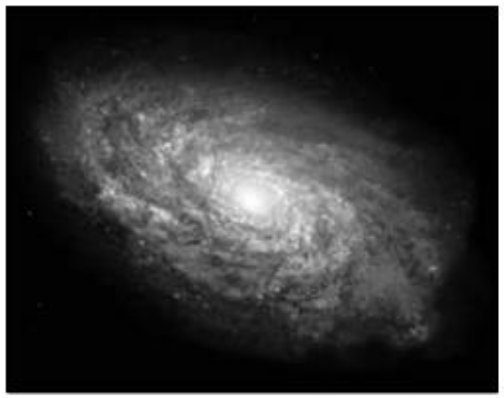
Data Mining – 7

Classification & Decision Trees

Dr. Tevfik Uyar



(a) A spiral galaxy.

(b) An elliptical galaxy.

Classification

Task of assigning objects to one of predefined categories.

Classification

- It may be possible to find a classification model.

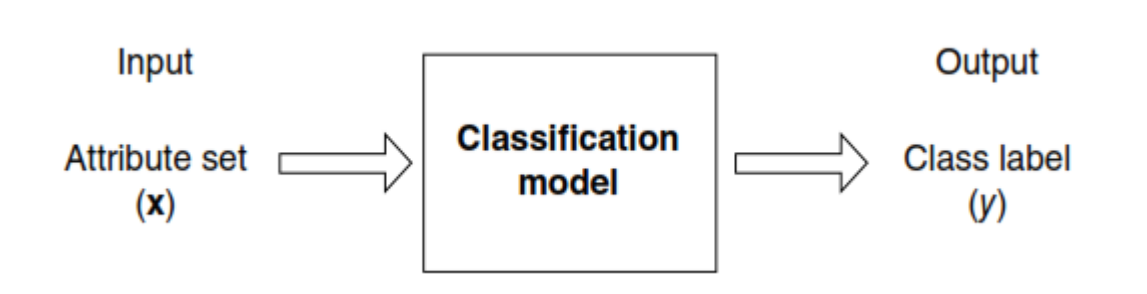


Table 4.1. The vertebrate data set.

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
gila monster	cold-blooded	scales	no	no	no	yes	yes	?

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

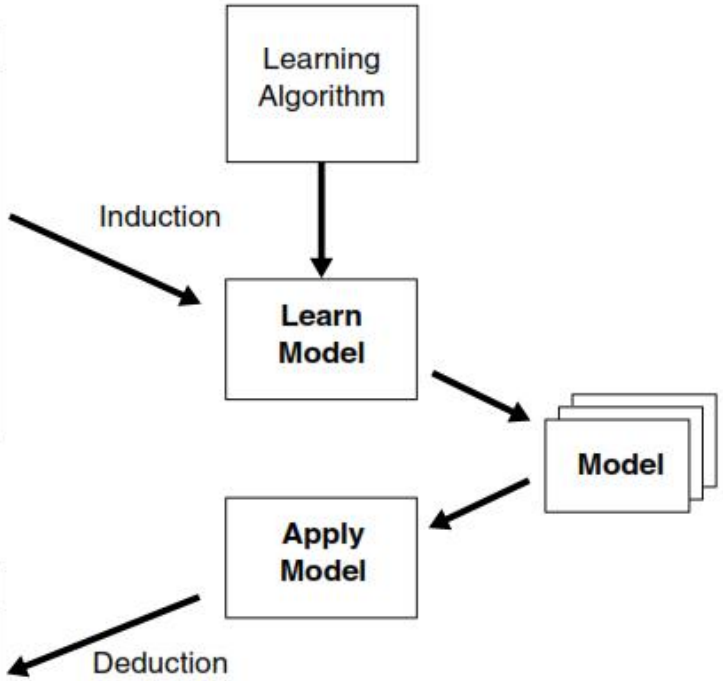


Figure 4.3. General approach for building a classification model.

- decision tree classifiers,
- rule-based classifiers,
- neural networks,
- support vector machines,
- and naive Bayes classifiers.
- Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data.

Learning Algorithm

Confusion Matrix

Table 4.2. Confusion matrix for a 2-class problem.

		Predicted Class	
		$Class = 1$	$Class = 0$
Actual Class	$Class = 1$	f_{11}	f_{10}
	$Class = 0$	f_{01}	f_{00}

For binary classifiers (e.g: Cancer or Not, COVID-19 (+) or (-):

False Positive

False Negative

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}. \quad (4.1)$$

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}. \quad (4.2)$$

Decision Tree

Decision Tree Classifier

- **Root Node:**
  - No incoming edges and zero or more outgoing edges
- **Internal Nodes:**
  - One incoming edge and two or more outgoing edge
- **Leaf Nodes:**
  - One incoming edge and no outgoing edges

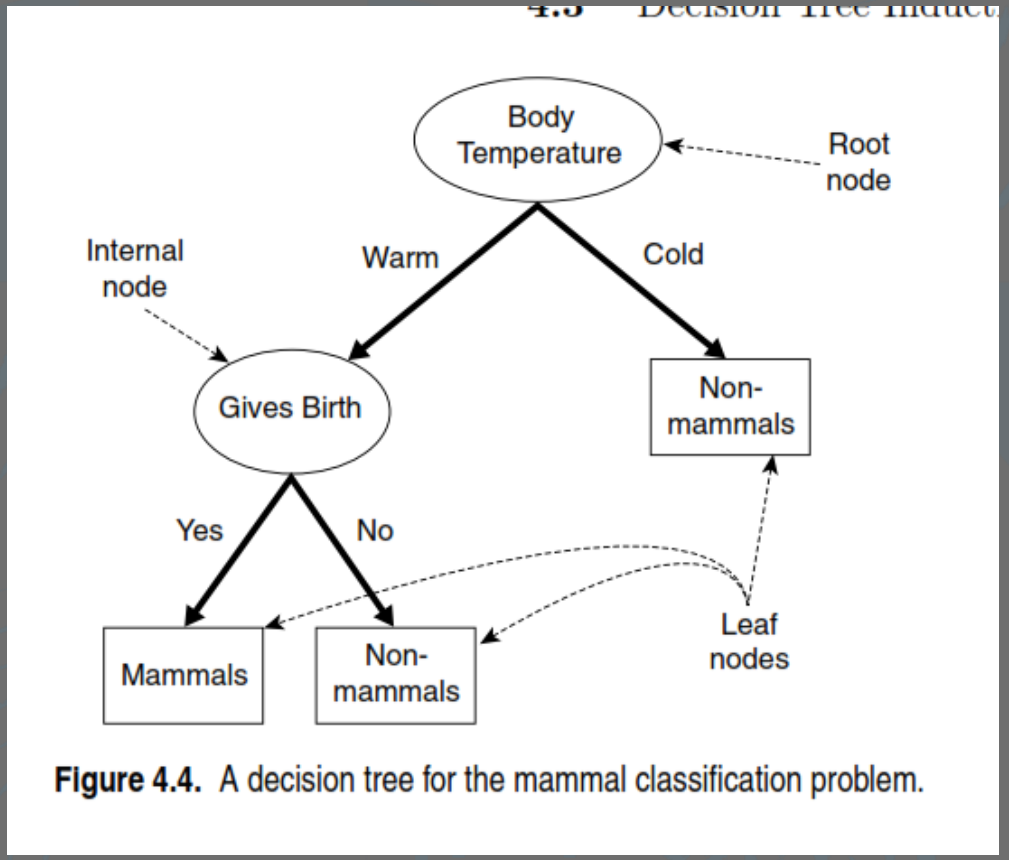


Figure 4.4. A decision tree for the mammal classification problem.

Example

- Let's convert it binary:
  - Mammals / Non-Mammals

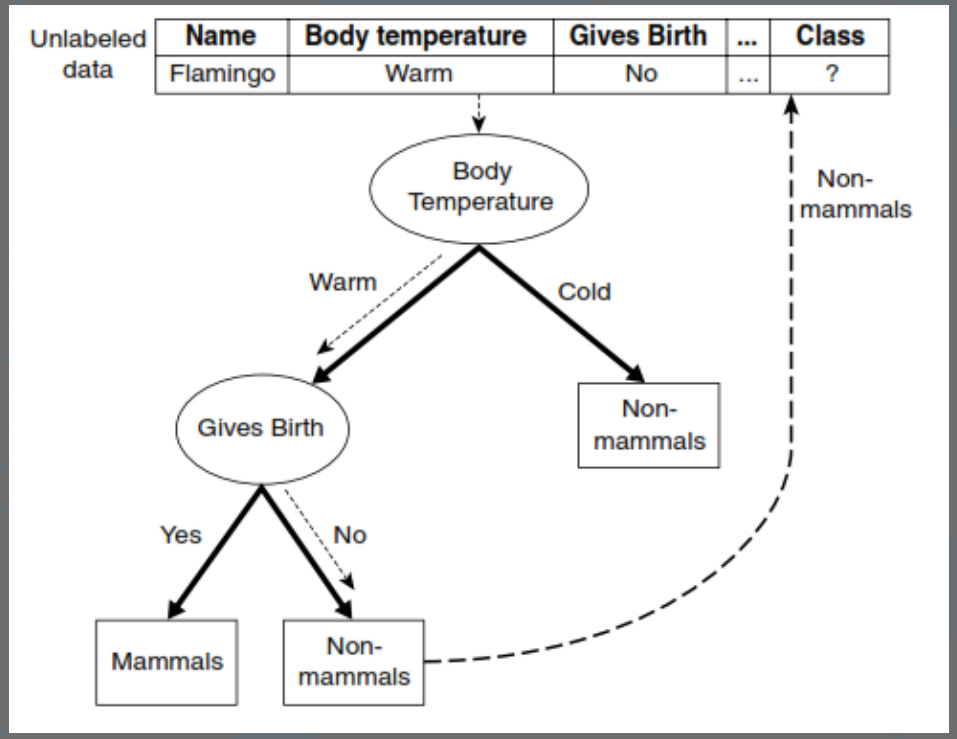


Table 4.1. The vertebrate data set.

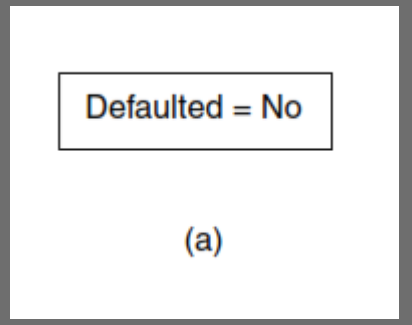
Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Building a Decision Tree

- Hunt's algorithm
  - Step 1: If all records in an attribute value belongs to the same class, then this attribute value is labeled with that class.
  - Step 2: If there are more than one class, a child node is created for each outcome of test condition.

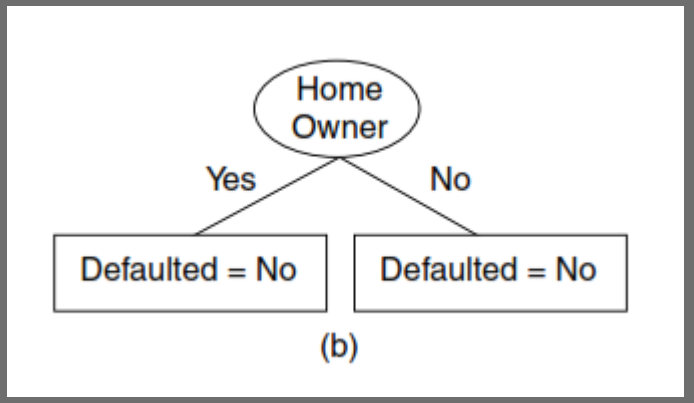
Example

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	binary	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



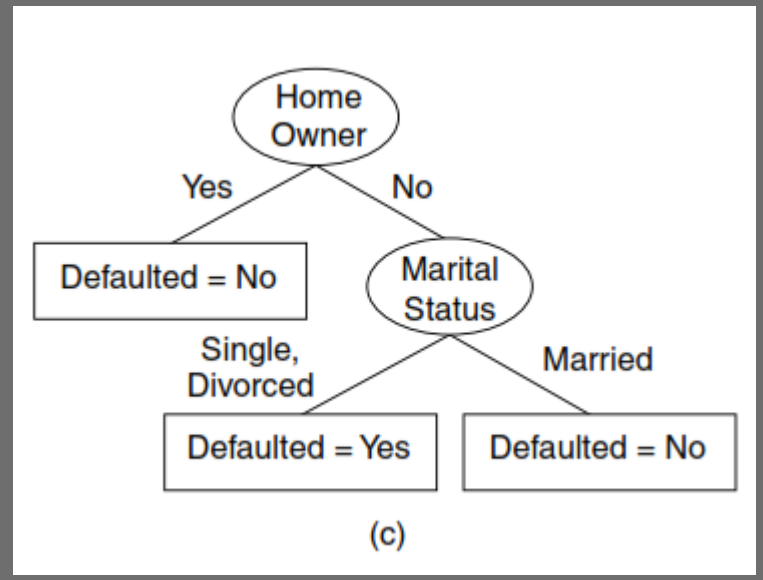
Example

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



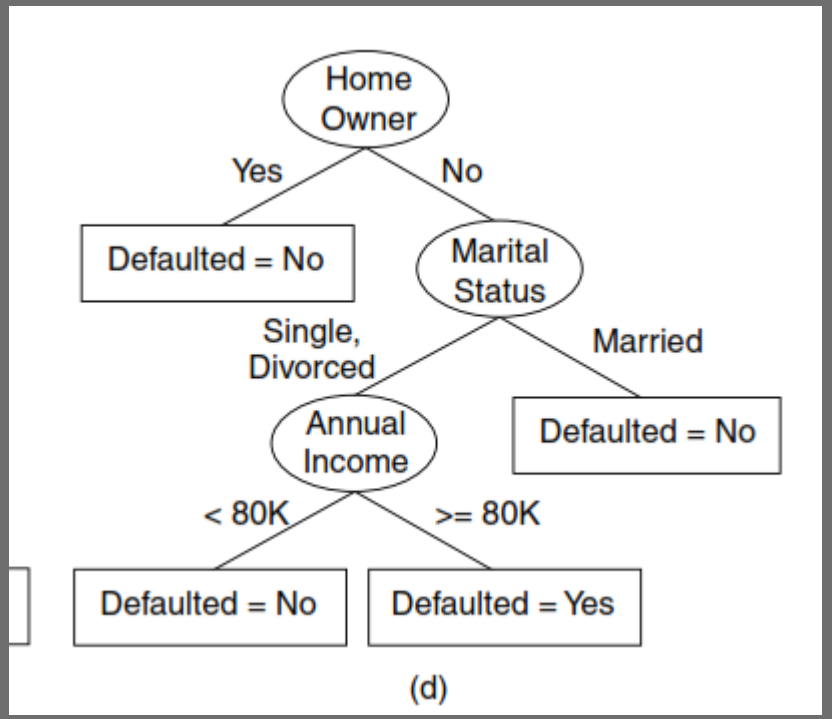
Example

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Example

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	binary	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Attribute Types in Decision Trees

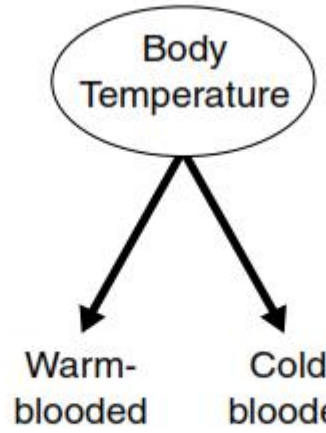
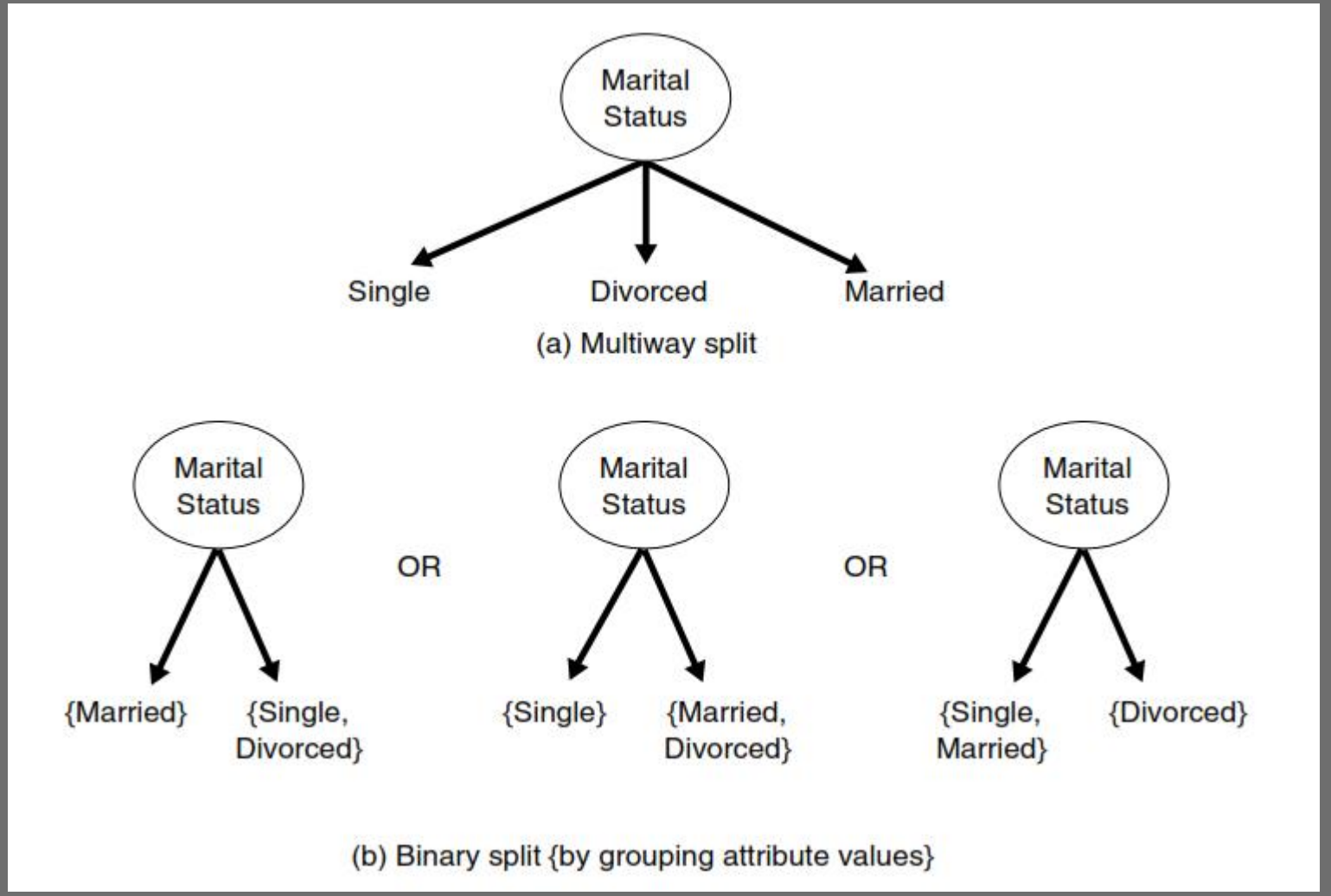


Figure 4.8. Test condition for binary attributes.

Nominal and Ordinal



Attribute Types in Decision Trees

For interval and ratio, the test condition can be expressed as a comparison test:

Only two option:
A < v? or A > v?

Or more...

$$v_i \leq A < v_{i+1}, \text{ for } i = 1, \ldots, k.$$

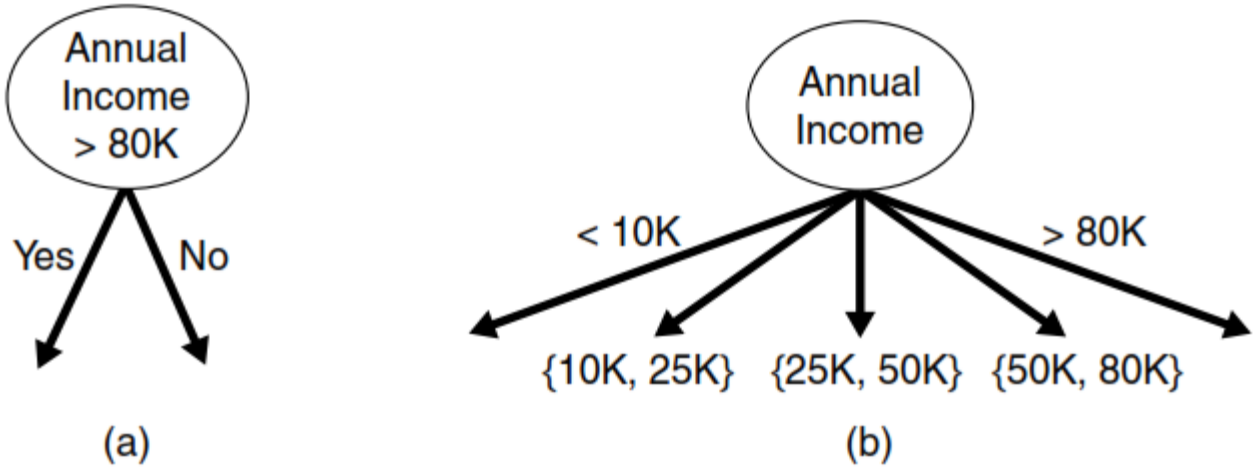


Figure 4.11. Test condition for continuous attributes.

About decision trees...

- In complex data sets, it is difficult to build a decision tree by hand.
- Which attribute must come first? (as a root node) and how we will split the data? This needs a search and calculation.
- To determine how well a test condition performs, we need to compare the degree of impurity of the parent node (before splitting) with the degree of impurity of the child nodes (after splitting). The larger their difference, the better the test condition.
- The gain, Δ, is a criterion that can be used to determine the goodness of a split:

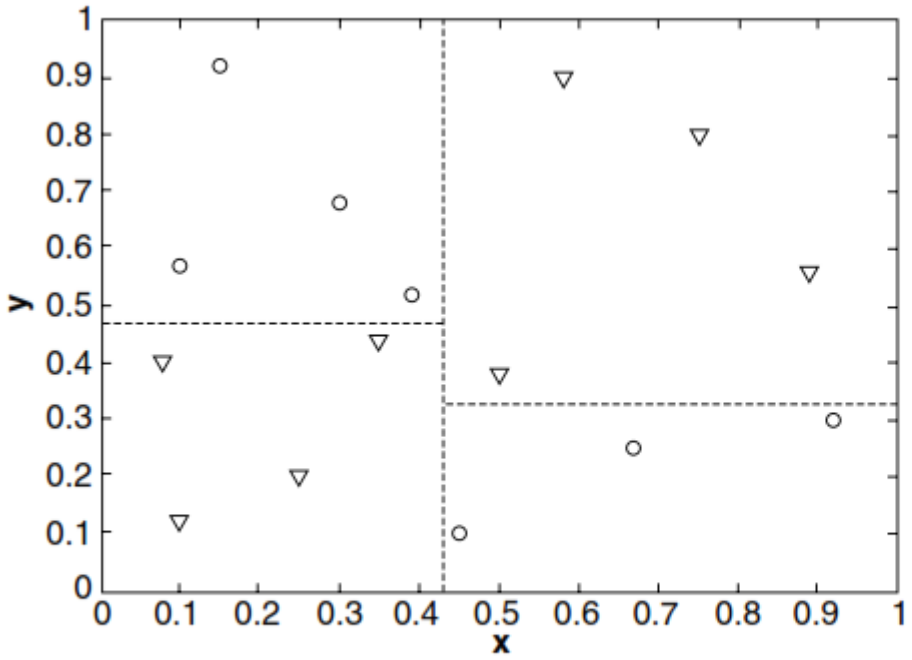
$$\begin{aligned}
\text{Entropy}(t) &= -\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t), \\
\text{Gini}(t) &= 1 - \sum_{i=0}^{c-1}[p(i|t)]^2, \\
\text{Classification error}(t) &= 1 - \max_i[p(i|t)],
\end{aligned}$$

Impurity Measures

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j),$$

Gain

Decision Boundaries



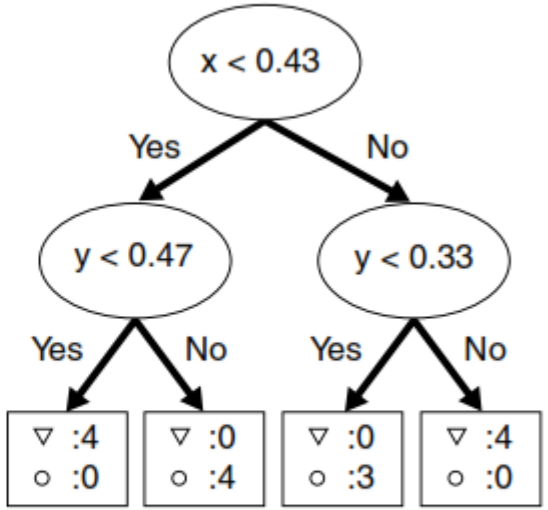
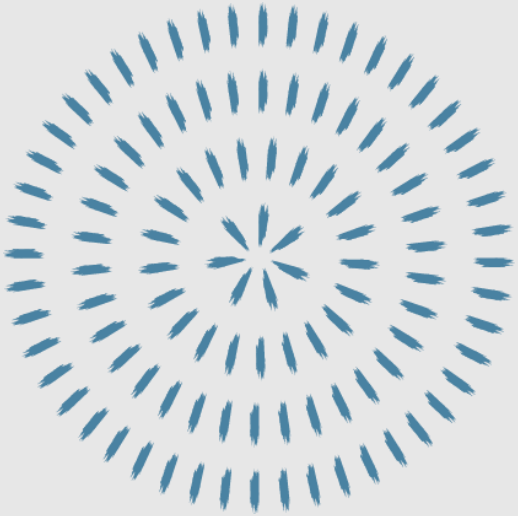


Figure 4.20. Example of a decision tree and its decision boundaries for a two-dimensional data set.



Application

https://colab.research.google.com/drive/103A2NWthcMpQHG2R9_0YpOwCacWjP1B9?usp=sharing