Data Warehouse and Data Mining

Dhruv Gupta

dhruv.gupta@ntnu.no

07-March-2023

NTNU | Norwegian University of Science and Technology

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Classification
   - Introduction
   - Decision Trees
     - Definitions
     - Decision Tree Induction
     - Underfitting and Overfitting
     - Evaluation
   - Nearest Neighbor Classifiers
   - Other Classification Methods

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

Administrative

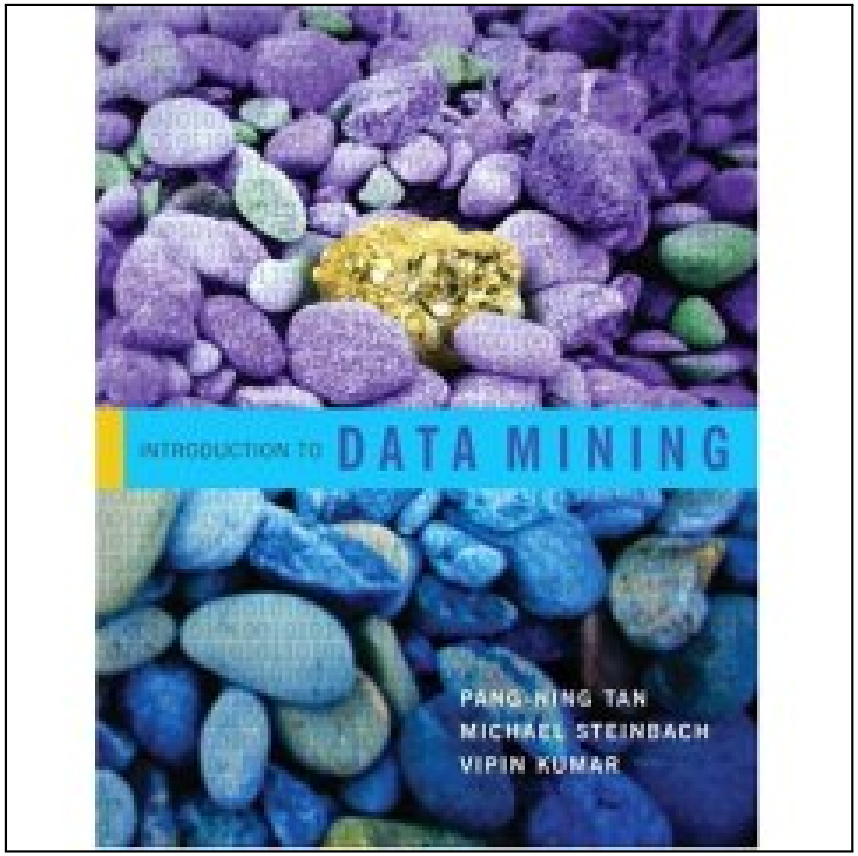
1. Third Assignment
   - Due by 9.March.2023.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

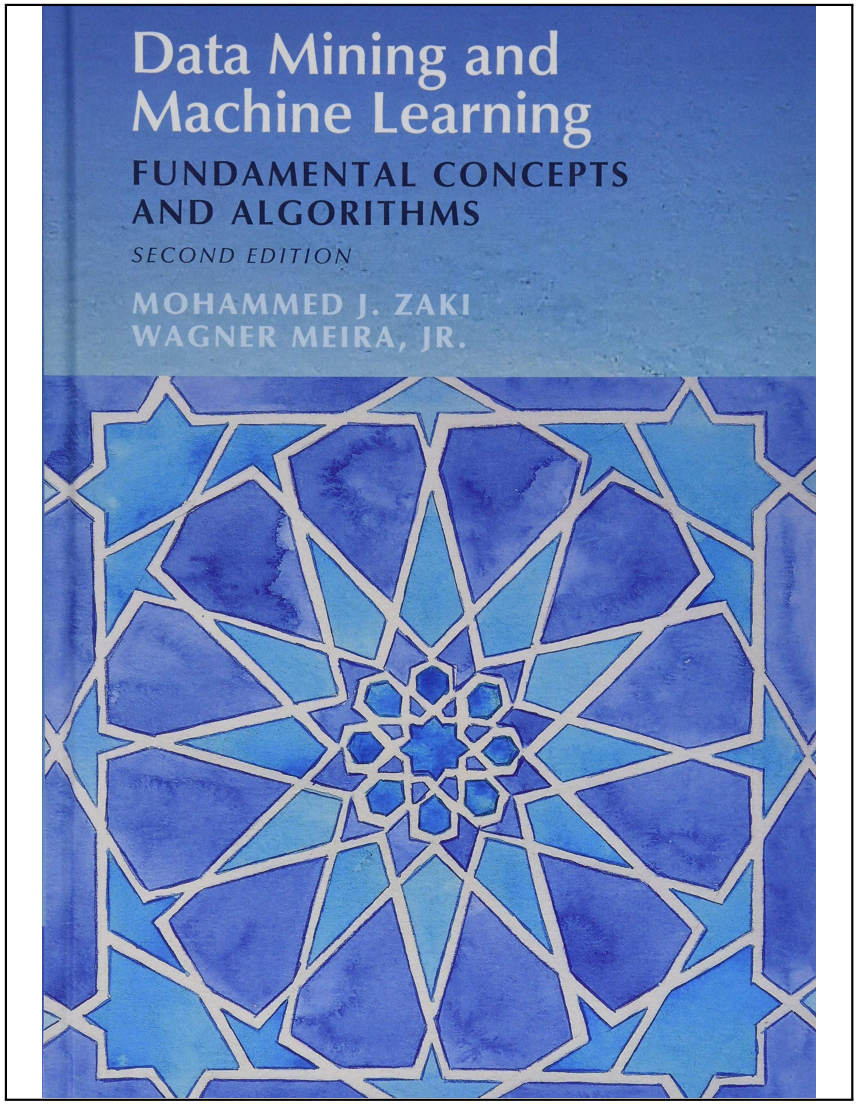
References for "Classification"

1. Book: Tan et al. *"Introduction to Data Mining"*, 1st Edition, 2006, Pearson Education Inc.
2. Text and images for majority of slides in "Classification" are based on the book by Tan et al.



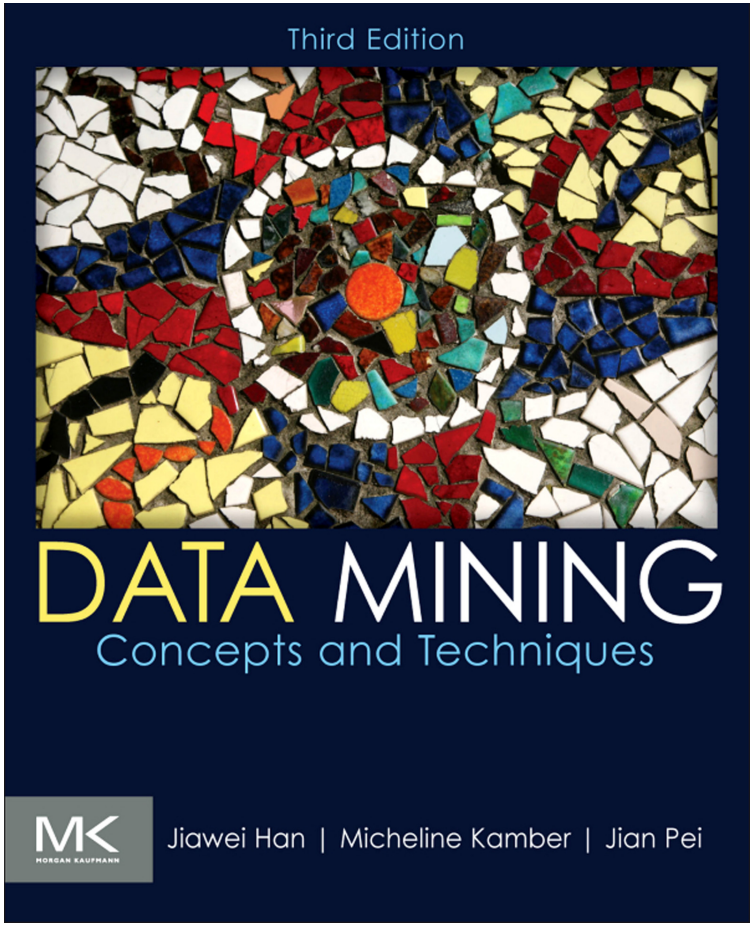
References for "Classification"

1. Book: Zaki and Meira. *"Data Mining and Machine Learning: Fundamental Concepts and Algorithms"*, 2nd Edition, 2020, Cambridge University Press.
2. All text and images for some slides in "Classification" are based on the book by Zaki and Meira et al.



References for "Classification"

1. Book: Han et al. *"Data Mining Concepts and Techniques"*, 3rd Edition, 2012, Morgan Kaufmann Publishers.
2. All text and images for some slides in "Classification" are based on the book by Han et al.



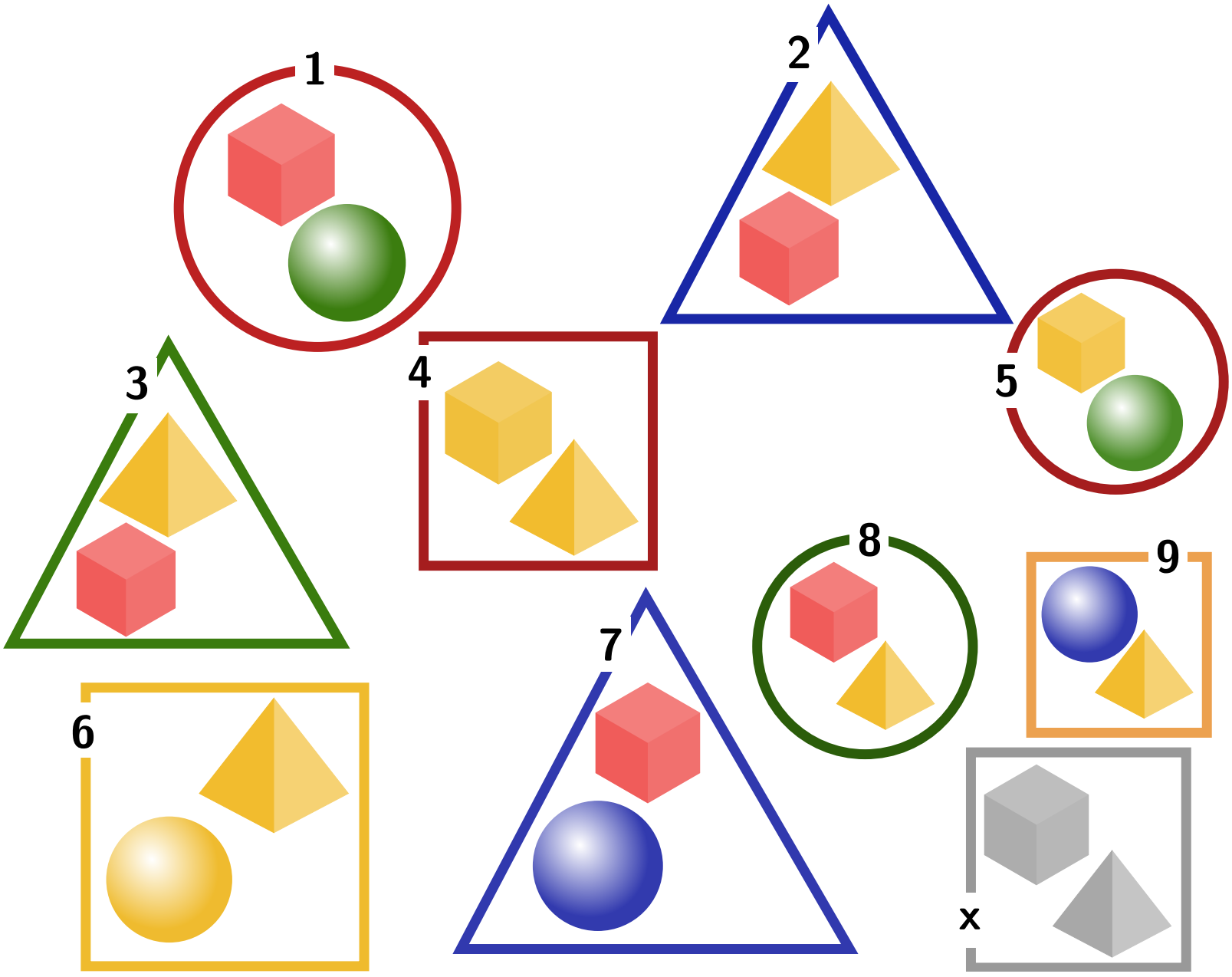
1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Classification
   - Introduction
   - Decision Trees
     - Definitions
     - Decision Tree Induction
     - Underfitting and Overfitting
     - Evaluation
   - Nearest Neighbor Classifiers
   - Other Classification Methods

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

Classification — Toy Example



Classification — Supervised Learning

- Consider the example observations from earlier.
- Given this set of data can we learn something, so that when a new observation is made (e.g., "x") we can predict its other characteristics?

ID	Shape	Shape Color	Object_1	Color_1	Object_2	Color_2
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow
x	Square	?	Cube	?	Pyramid	?

Classification — Definition

- Given a collection of records (training set):
  - Each record is by characterized by a tuple (x, y), where x is the attribute set and y is the class label:
    - x: attribute, predictor, independent variable, or input.
    - y: class, response, dependent variable, or output.
- Task:
  - Learn a model that maps each attribute set x into one of the predefined class labels y.

Figure 3.2. A schematic illustration of a classification task.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

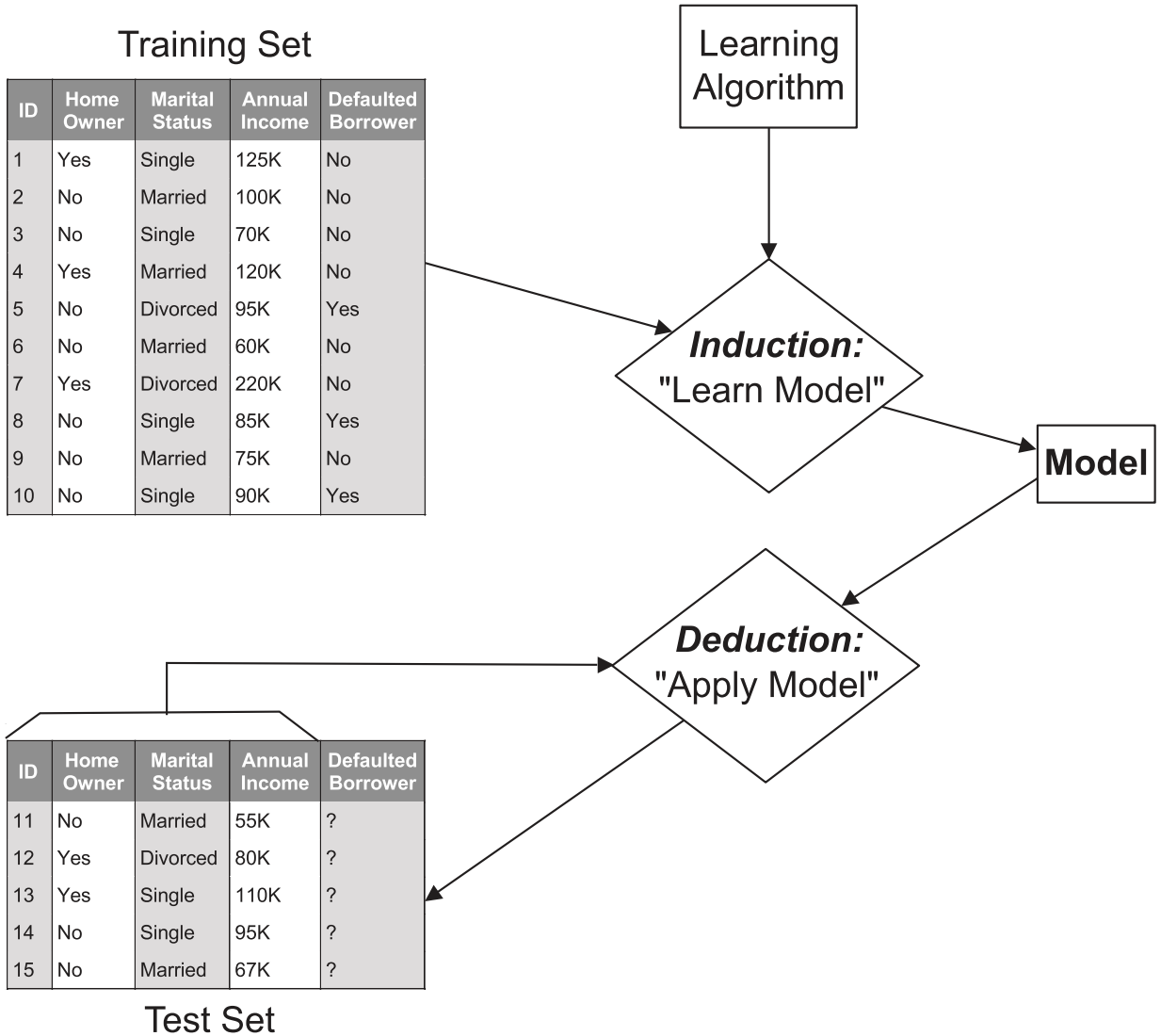
Examples of Classification Task

Task	**Attribute Set** x	**Class Label** y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from x-rays or MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies

Figure 3.2. A schematic illustration of a classification task.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

General Approach for Building a Classification Model



ID	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

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
11	No	Married	55K	?
12	Yes	Divorced	80K	?
13	Yes	Single	110K	?
14	No	Single	95K	?
15	No	Married	67K	?

Figure 3.3. General framework for building a classification model.

Classification Techniques

- Base Classifiers
  - Decision Tree based Methods.
  - Rule-based Methods.
  - Nearest-Neighbor.
  - Naïve Bayes and Bayesian Belief Networks.
  - Support Vector Machines.
  - Neural Networks, Deep Neural Nets.
- Ensemble Classifiers
  - Boosting, Bagging, Random Forests.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Classification Techniques

- In curriculum (in detail):
  - Decision Tree based Methods.
  - Nearest-Neighbor.
- Can be used as black-boxes without knowing to many details in this introductory course:
  - Rule-based Methods.
  - Naïve Bayes and Bayesian Belief Networks.
  - Support Vector Machines.
  - Neural Networks, Deep Neural Nets.
  - Boosting, Bagging, Random Forests.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

Example of a Decision Tree — 1

Table 3.2. A sample data for the vertebrate classification problem.

Vertebrate 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

Vertebrate 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	?

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Example of a Decision Tree — 1

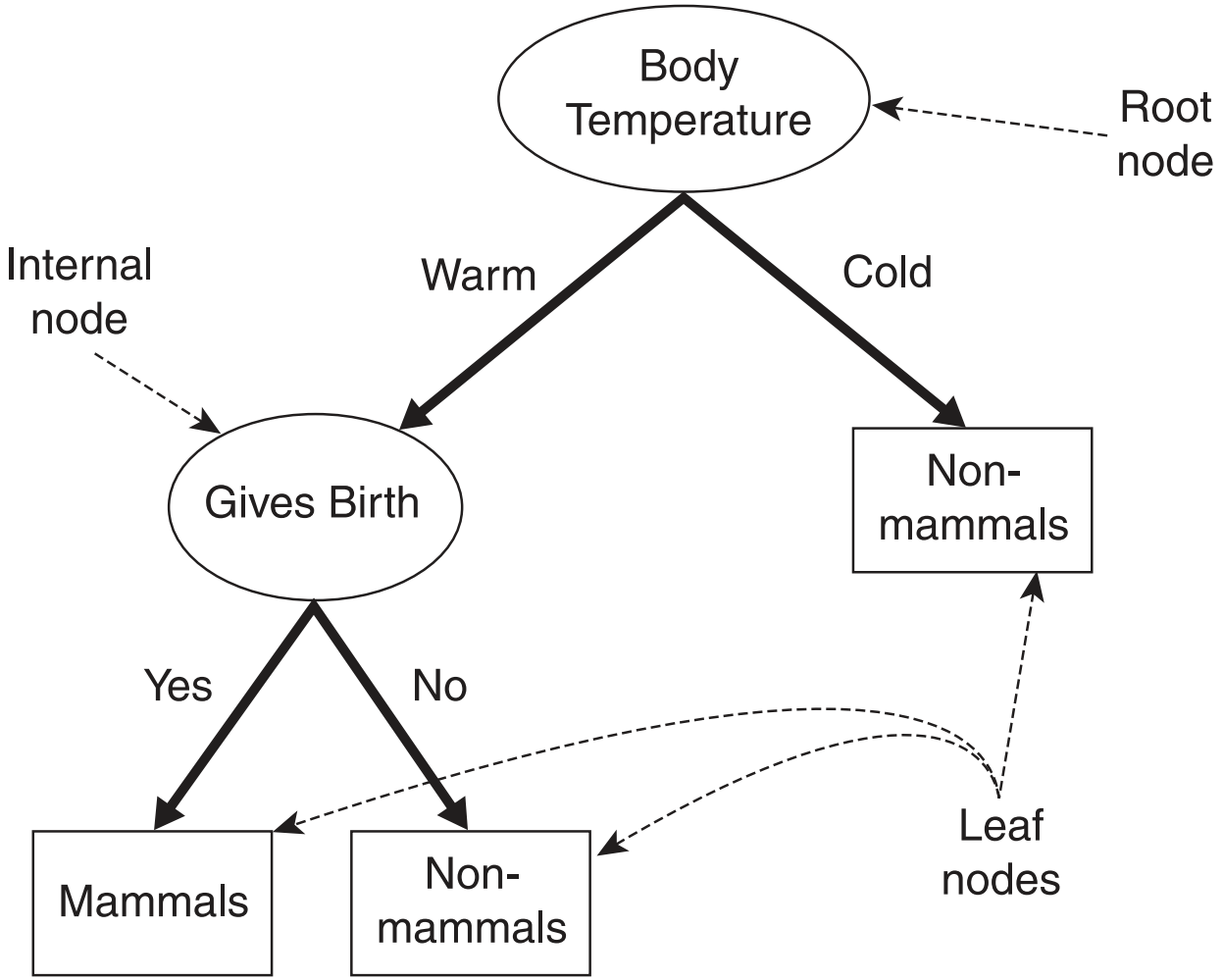


Figure 3.4. A decision tree for the mammal classification problem.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Example of a Decision Tree — 1

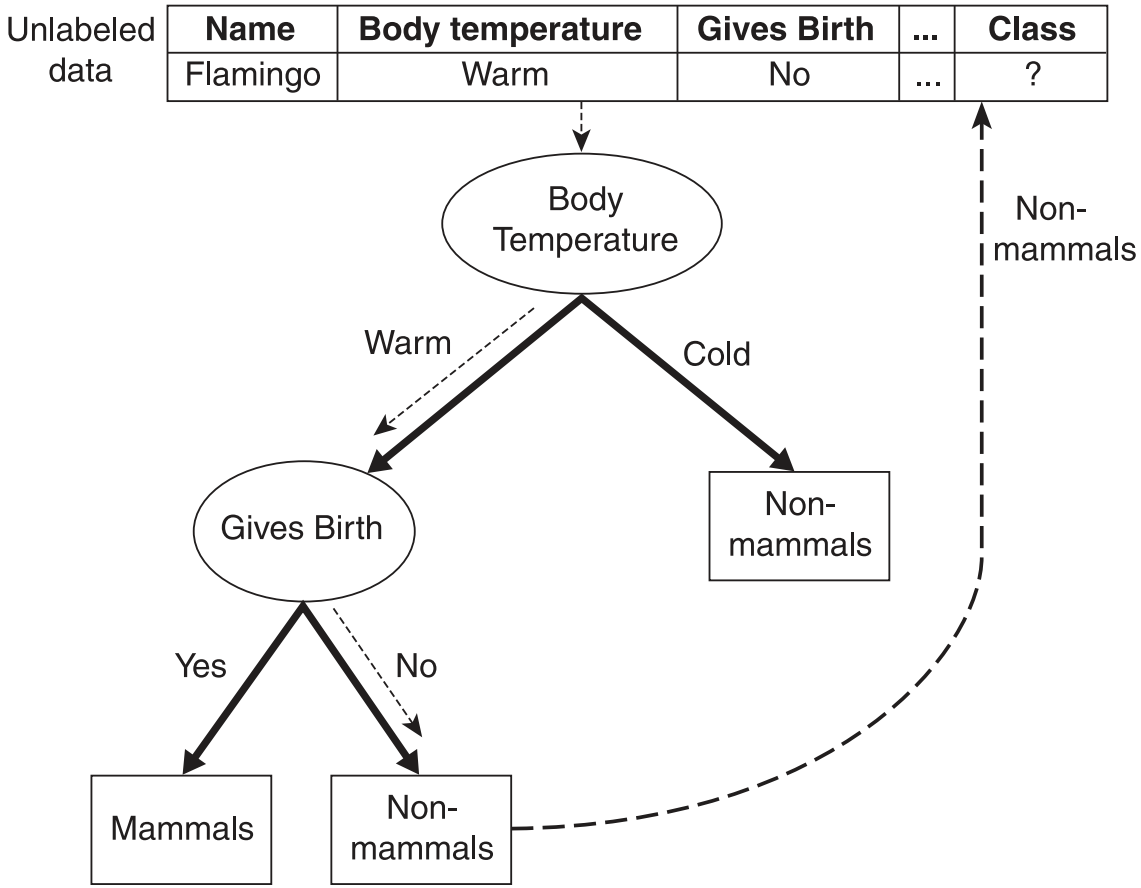


Figure 3.5. Classifying an unlabeled vertebrate. The dashed lines represent the outcomes of applying various attribute test conditions on the unlabeled vertebrate. The vertebrate is eventually assigned to the `Non-mammals` class.

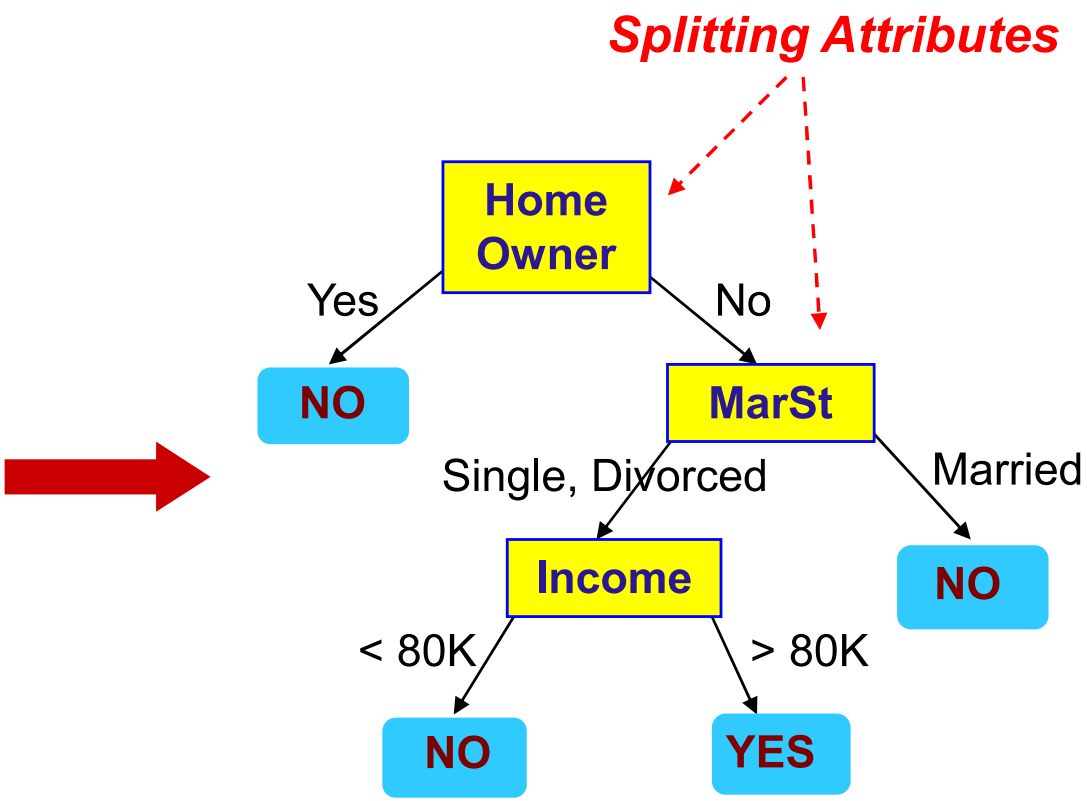
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Example of a Decision Tree — 2

categorical categorical continuous class

ID	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

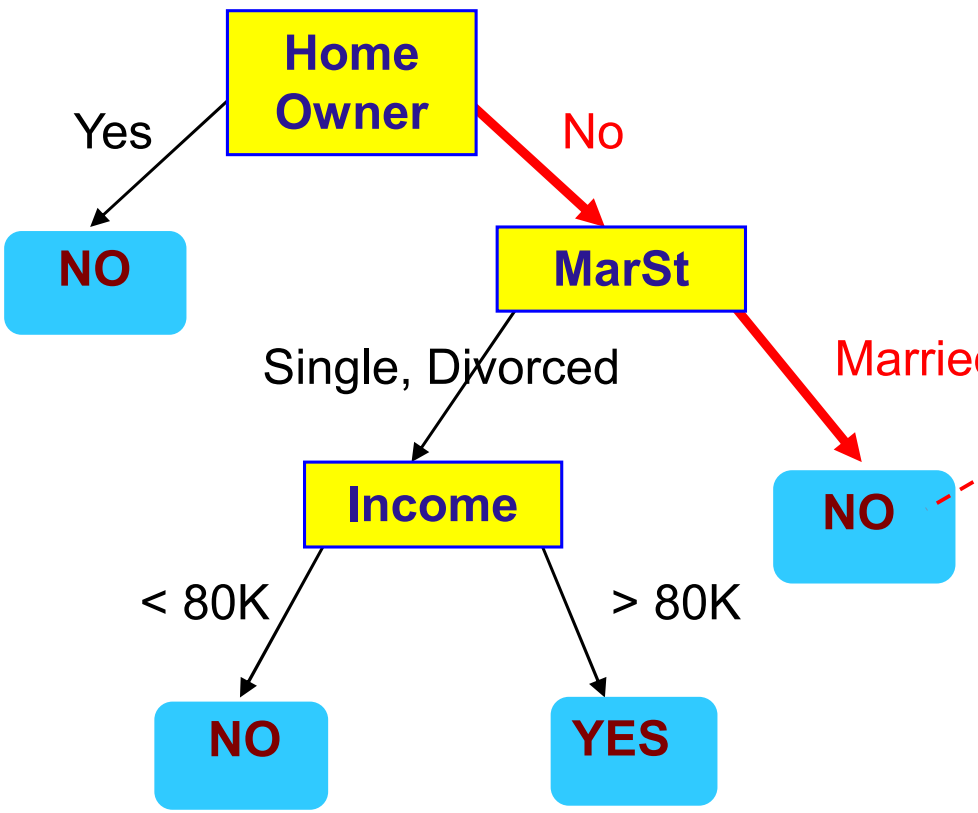
Training Data



Model: Decision Tree

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Example of a Decision Tree — 2



Test Data

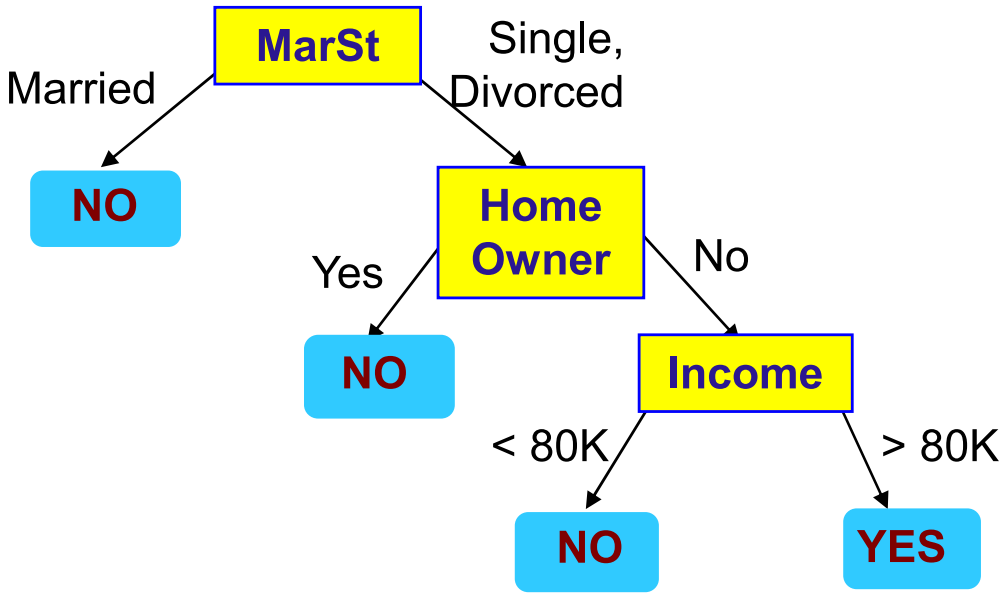
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Assign Defaulted to "No"

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Example of a Decision Tree — 2

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
		categorical	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



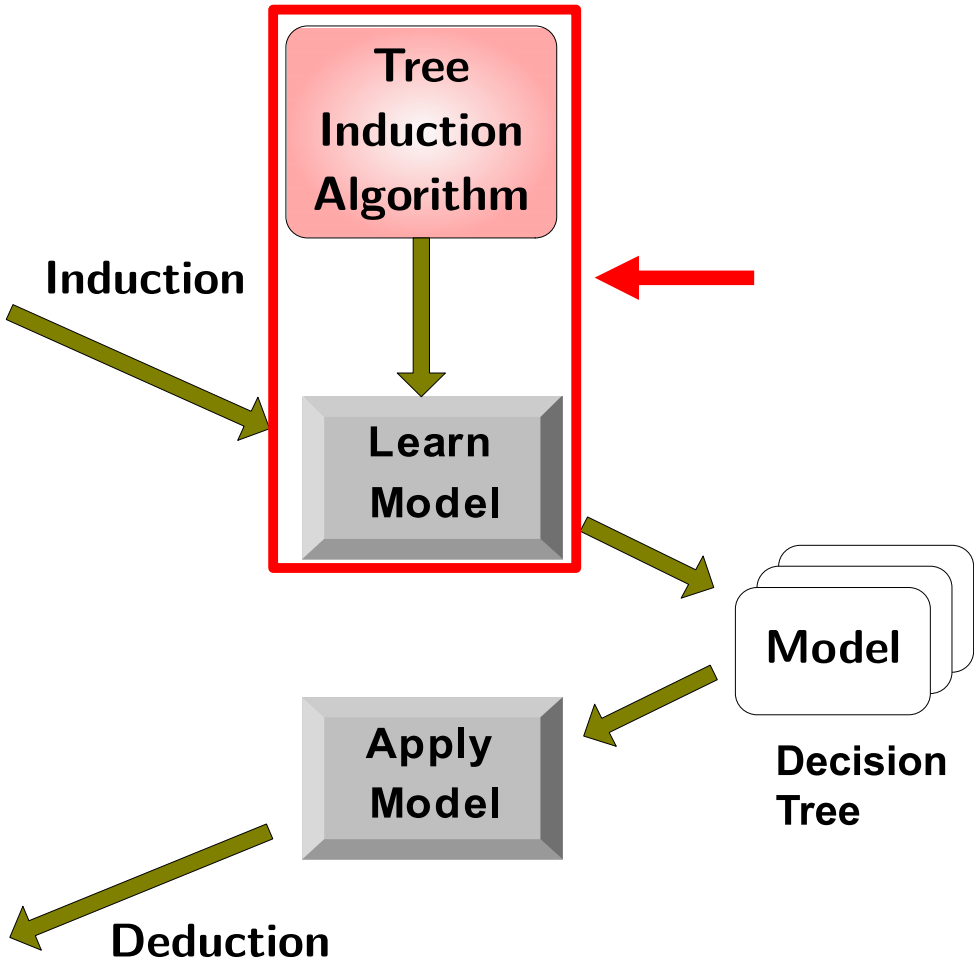
There could be more than one tree that fits the same data!

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Decision Tree Classification Task

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

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	?



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

Decision Tree Induction

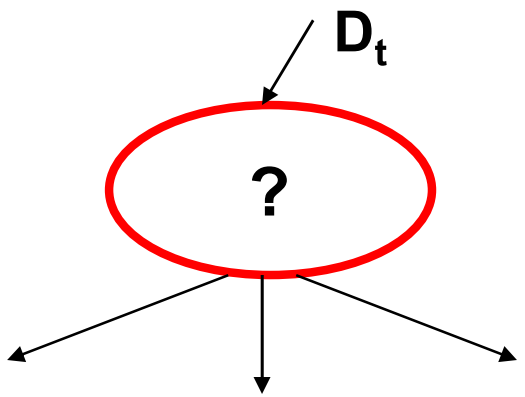
- Many algorithms:
  - Hunt's Algorithm (one of the earliest)
  - CART
  - ID3, C4.5
  - SLIQ, SPRINT

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Hunt's Algorithm

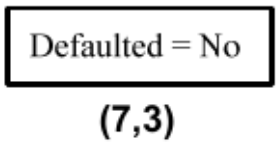
1. Let D_t be the set of training records that reach a node t.
2. General Procedure:
   - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t.
   - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

ID	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

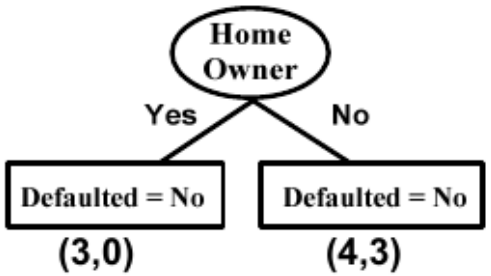


All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

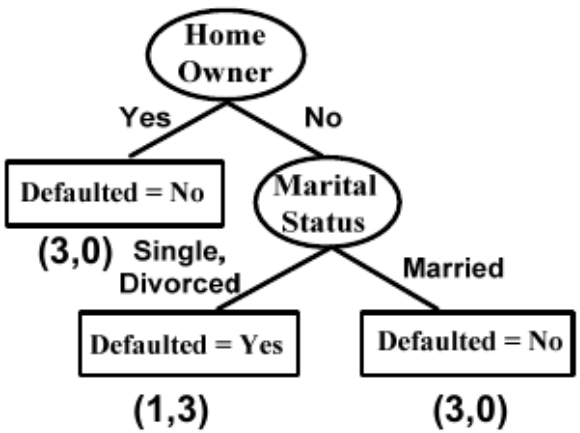
Hunt's Algorithm



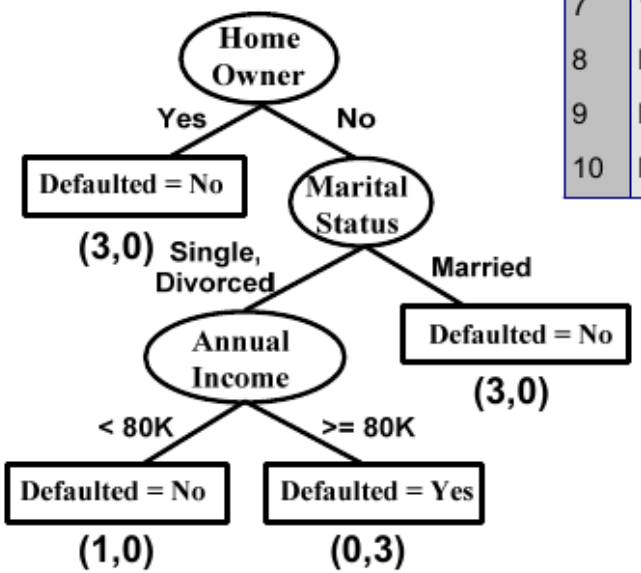
(a)



(b)



(c)



(d)

ID	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

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Design issues of Decision Tree Induction

- How should training records be split?
  - Method for expressing test condition.
    - Depending on attribute types.
  - Measure for evaluating the goodness of a test condition.
- How should the splitting procedure stop?
  - Stop splitting if all the records belong to the same class or have identical attribute values.
  - Early termination.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

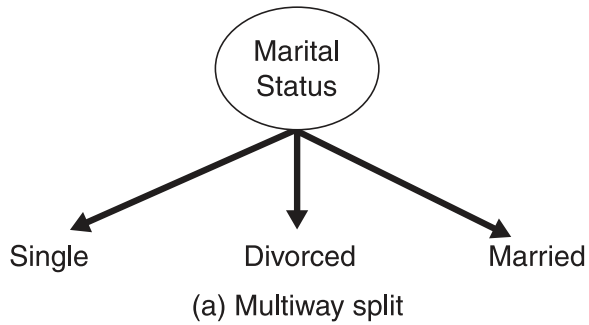
Methods for Expressing Test Conditions

- Depends on attribute types:
  - Binary
  - Nominal
  - Ordinal
  - Continuous

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Test Condition for Nominal Attributes

1. Multi-way Split:
   - Use as many partitions as distinct values.
2. Binary Split:
   - Divides values into two subsets.



(a) Multiway split

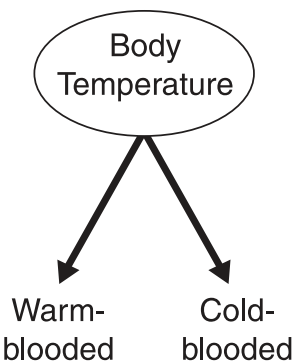
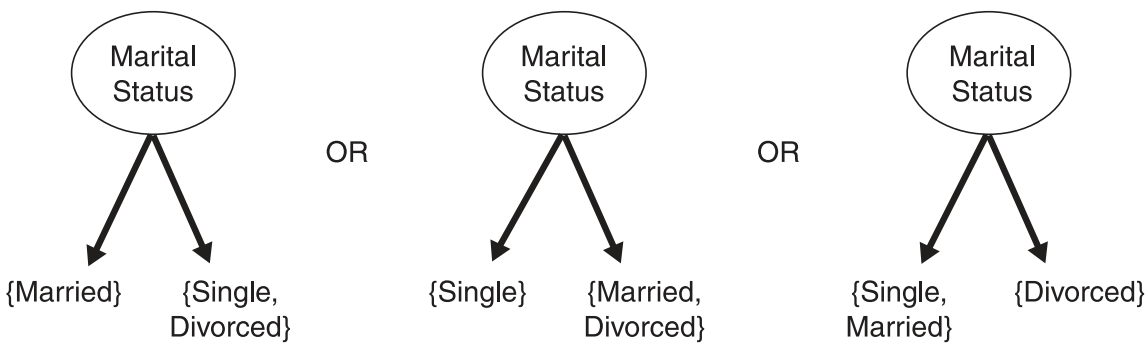


Figure 3.7. Attribute test condition for a binary attribute.

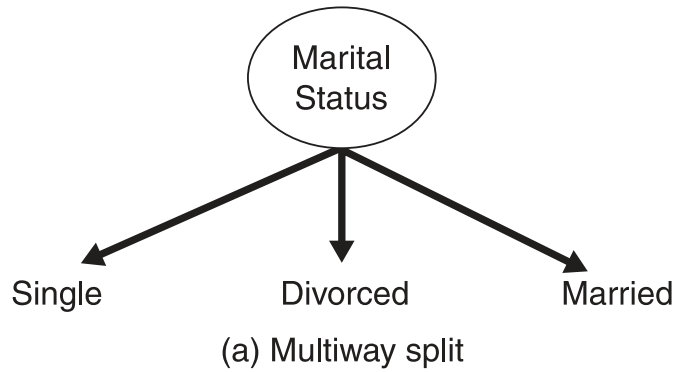


(b) Binary split {by grouping attribute values}

Figure 3.8. Attribute test conditions for nominal attributes.

Test Condition for Ordinal Attributes

1. Multi-way Split:
   - Use as many partitions as distinct values.
2. Binary Split:
   - Divides values into two subsets.
   - Preserve order property among attribute values.



(a) Multiway split

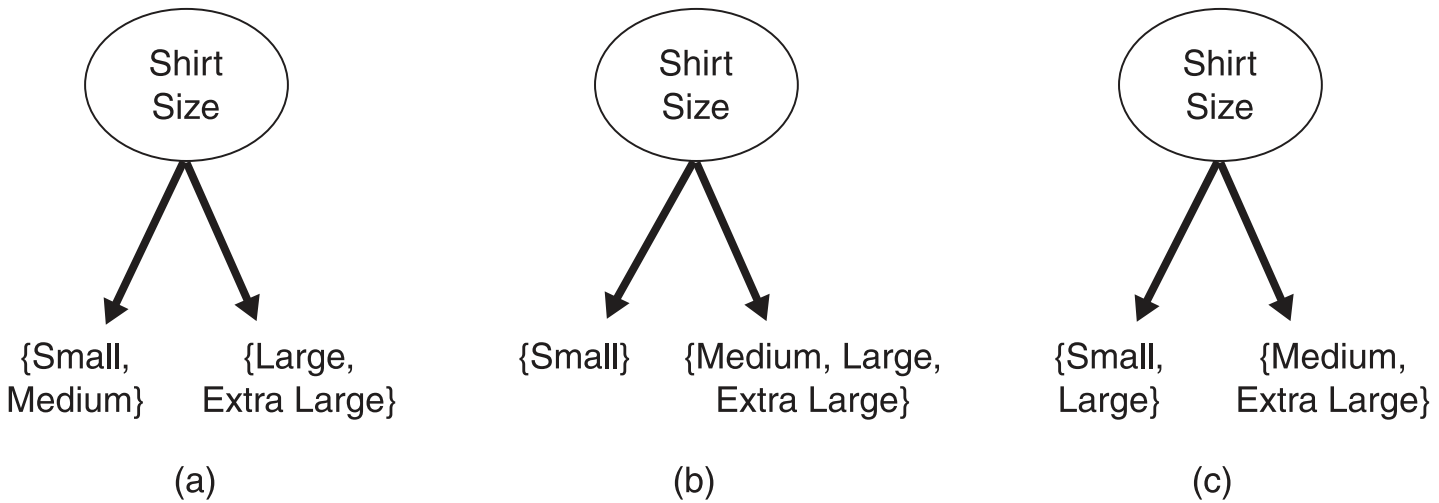


Figure 3.9. Different ways of grouping ordinal attribute values.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Test Condition for Continuous Attributes

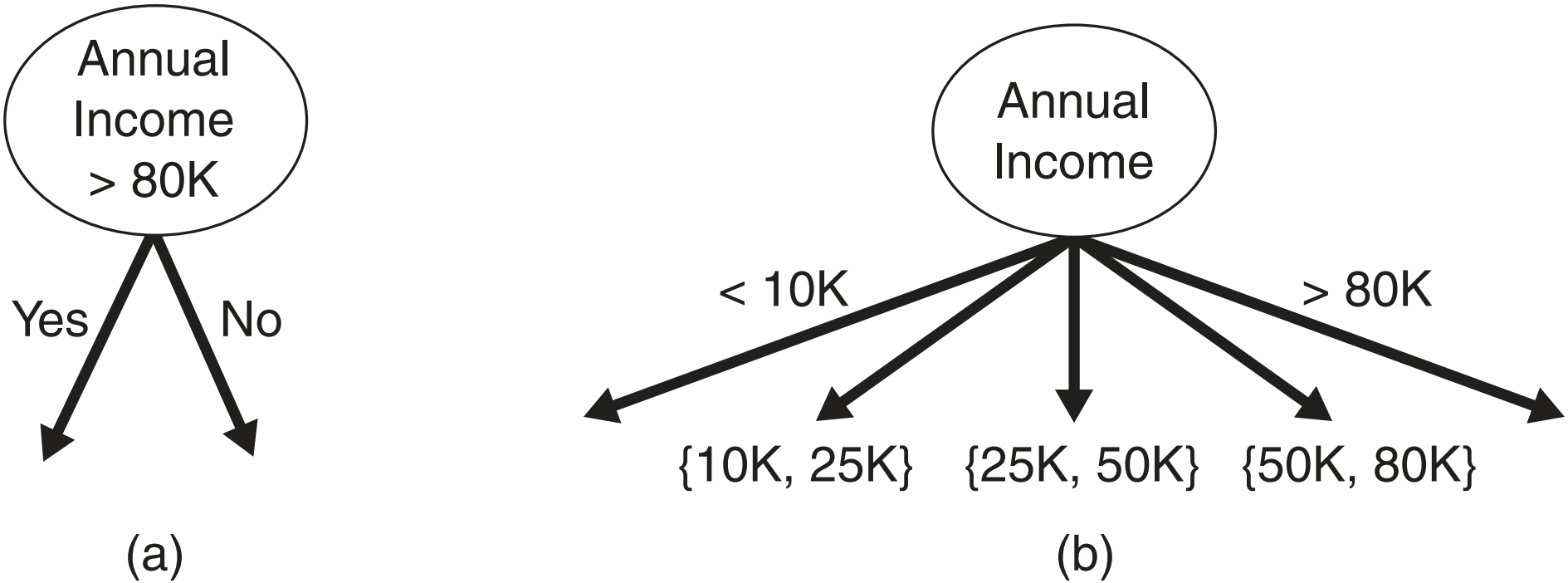


Figure 3.10. Test condition for continuous attributes.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Splitting Based on Continuous Attributes

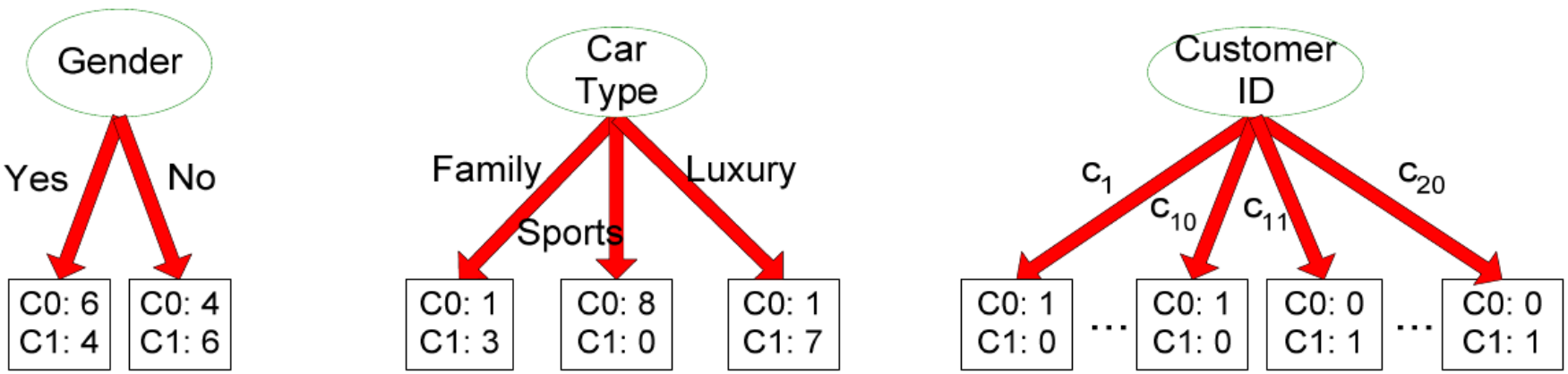
- Different ways of handling.
  - Discretization to form an ordinal categorical attribute.
  - Ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.
    - Static – discretize once at the beginning.
    - Dynamic – repeat at each node.
- Binary Decision: $(A < v)$ or $(A \neq v)$.
  - Consider all possible splits and finds the best cut.
  - Can be more compute intensive.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to determine the Best Split

Customer Id	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Before Splitting: 10 records of class 0,
10 records of class 1



Which test condition is the best?

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to determine the Best Split

- Greedy approach:
  - Nodes with purer class distribution are preferred.
- Need a measure of node impurity:

C0: 5 C1: 5	C0: 9 C1: 1
High Degree of Impurity	Low Degree of Impurity

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Measures of Node Impurity

- Gini Index:

$$\text{Gini Index} = 1 - \sum_{i=0}^{c-1} \Big[p_i(t)\Big]^2. \tag{1}$$

- Entropy:

$$\text{Entropy} = -\sum_{i=0}^{c-1} p_i(t) \cdot \log_2(t). \tag{2}$$

- Misclassification Error:

$$\text{Classification Error} = 1 - \max\Big[p_i(t)\Big]. \tag{3}$$

- Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Finding the Best Split

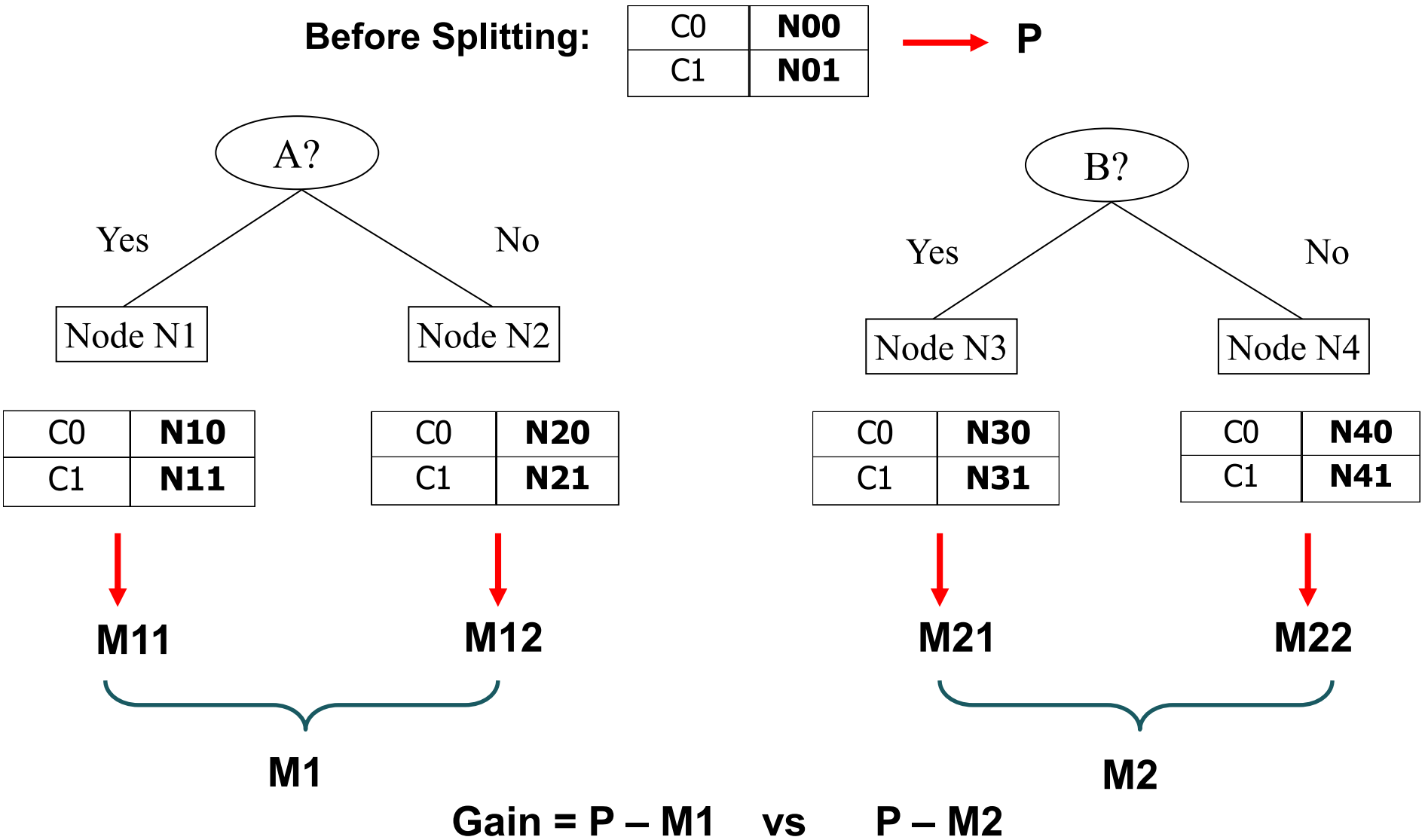
1. Compute impurity measure (P) before splitting.
2. Compute impurity measure (M) after splitting.
   - Compute impurity measure of each child node.
   - M is the weighted impurity of child nodes.
3. Choose the attribute test condition that produces the highest gain.

$$\text{Gain} = P - M. \tag{4}$$

4. Or, equivalently, lowest impurity measure after splitting (M).

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Finding the Best Split



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Measure of Impurity — GINI

- Gini Index for a given node t

$$\text{Gini Index} = 1 - \sum_{i=0}^{c-1} \left[p_i(t) \right]^2. \tag{5}$$

- Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes.
- Maximum of $1 - \frac{1}{c}$ when records are equally distributed among all classes, implying the least beneficial situation for classification.
- Minimum of 0 when all records belong to one class, implying the most beneficial situation for classification.
- Gini index is used in decision tree algorithms such as CART, SLIQ, SPRINT.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Measure of Impurity — GINI

- Gini Index for a given node t

$$\text{Gini Index} = 1 - \sum_{i=0}^{c-1} \left[p_i(t)\right]^2. \tag{6}$$

- Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes.
- For 2-class problem $(p, 1-p)$:

$$\text{GINI} = 1 - p^2 - (1-p)^2 = 2 \cdot p(1-p). \tag{7}$$

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Measure of Impurity — GINI

- Gini Index for a given node t

$$\text{Gini Index} = 1 - \sum_{i=0}^{c-1} \left[p_i(t) \right]^2. \tag{8}$$

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

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

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

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

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

Gini = 1 – (1/6)² – (5/6)² = 0.278

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

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

Gini = 1 – (2/6)² – (4/6)² = 0.444

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Computing GINI Index for a Collection of Nodes

- When a node p is split into k partitions (children):

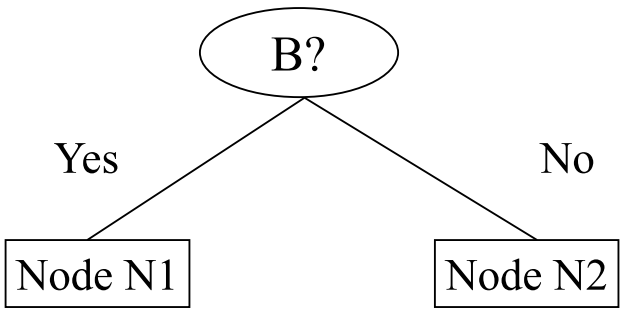
$$\text{GINI}_{\text{split}} = \sum_{i=1}^{k} \frac{n_i}{n} \text{GINI}(i). \tag{9}$$

- Where n_i is the number of records at child i and n is the number of records at parent node p.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Binary Attributes — Computing GINI Index

- Splits into two partitions (child nodes).
- Effect of Weighing partitions:
  - Larger and purer partitions are sought.



	Parent
C1	**7**
C2	**5**
Gini = 0.486	

Gini(N1)
= 1 – (5/6)² – (1/6)²
= 0.278

Gini(N2)
= 1 – (2/6)² – (4/6)²
= 0.444

	N1	N2
C1	**5**	**2**
C2	**1**	**4**
Gini=0.361		

Weighted Gini of N1 N2
= 6/12 * 0.278 +
6/12 * 0.444
= 0.361

Gain = 0.486 – 0.361 = 0.125

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Categorical Attributes — Computing GINI Index

- For each distinct value, gather counts for each class in the dataset.
- Use the count matrix to make decisions.

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	0.163		

Two-way split
(find best partition of values)

	CarType	
	{Sports, Luxury}	{Family}
C1	9	1
C2	7	3
Gini	0.468	

	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10
Gini	0.167	

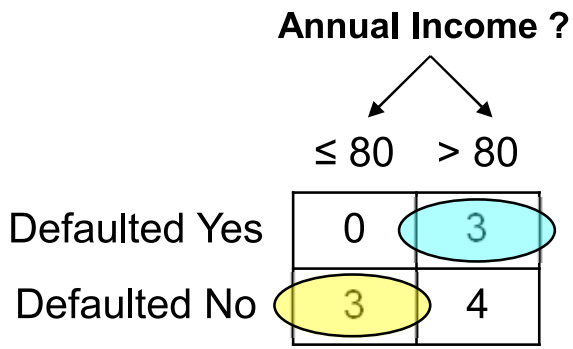
Which of these is the best?

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Continuous Attributes — Computing GINI Index

- Use Binary Decisions based on one value.
- Several Choices for the splitting value:
  - Number of possible splitting values = Number of distinct values.
- Each splitting value has a count matrix associated with it.
  - Class counts in each of the partitions, $A \leq v$ and $A > v$.
- Simple method to choose best v:
  - For each v, scan the database to gather count matrix and compute its Gini index.
  - Computationally Inefficient! Repetition of work.

ID	Home Owner	Marital Status	Annual Income	Defaulted
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



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Continuous Attributes: Computing GINI Index

- For efficient computation, for each attribute:
  1. Sort the attribute on values.
  2. Linearly scan these values, each time updating the count matrix and computing gini index.
  3. Choose the split position that has the least gini index.

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Annual Income										
Sorted Values →	60	70	75	85	90	95	100	120	125	220
Split Positions →	55	65	72	80	87	92	97	110	122	172	230

	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		***0.300***		0.343		0.375		0.400		0.420	

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Measure of Impurity — Entropy

- Entropy:

$$\text{Entropy} = -\sum_{i=0}^{c-1} p_i(t) \cdot \log_2\Big[p_i(t)\Big]. \tag{10}$$

- Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes.
- Maximum of $\log_2(c)$ when records are equally distributed among all classes, implying the least beneficial situation for classification.
- Minimum of 0 when all records belong to one class, implying most beneficial situation for classification
- Entropy based computations are quite similar to the GINI index computations.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Computing Entropy of a Single Node

- Entropy:

$$\text{Entropy} = -\sum_{i=0}^{c-1} p_i(t) \cdot \log_2\left[p_i(t)\right]. \quad (11)$$

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

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

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

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

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

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

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

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

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Computing Information Gain after Splitting

- Information Gain:

$$\text{Gain}_{\text{split}} = \text{Entropy}(p) - \sum_{i=1}^{k} \frac{n_i}{n} \text{Entropy}(i). \tag{12}$$

- Parent Node, p is split into k partitions (children) n is number of records in child node i
- Choose the split that achieves most reduction (maximizes GAIN).
- Used in the ID3 and C4.5 decision tree algorithms.
- Information gain is the mutual information between the class variable and the splitting variable.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Gain Ratio

- Gain Ratio:

$$\text{Gain Ratio} = \frac{\text{Gain}_{\text{split}}}{\text{Split Info}} \tag{13}$$

$$\text{Split Info} = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \left[\frac{n_i}{n} \right]. \tag{14}$$

- Parent Node, p is split into k partitions (children), n_i is the number of records in child node i.
- Adjusts Information Gain by the entropy of the partitioning (Split Info).
  - Higher entropy partitioning (large number of small partitions) is penalized!
- Used in C4.5 algorithm.
- Designed to overcome the disadvantage of Information Gain.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Gain Ratio

- Gain Ratio:

$$\text{Gain Ratio} = \frac{\text{Gain}_{\text{split}}}{\text{Split Info}} \tag{15}$$

$$\text{Split Info} = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \left[\frac{n_i}{n} \right]. \tag{16}$$

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	0.163		

SplitINFO = 1.52

	CarType	
	{Sports, Luxury}	{Family}
C1	9	1
C2	7	3
Gini	0.468	

SplitINFO = 0.72

	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10
Gini	0.167	

SplitINFO = 0.97

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Measure of Impurity — Classification Error

- Classification error at a node t:

$$\text{Error}(t) = 1 - \max_i \Big[p_i(t) \Big]. \tag{17}$$

- Maximum of $1 - \frac{1}{c}$ when records are equally distributed among all classes, implying the least interesting situation.
- Minimum of 0 when all records belong to one class, implying the most interesting situation.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Computing Error of a Single Node

- Classification error at a node t:

$$\text{Error}(t) = 1 - \max_i \left[p_i(t) \right]. \tag{18}$$

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

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

Error = 1 – max (0, 1) = 1 – 1 = 0

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

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

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

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

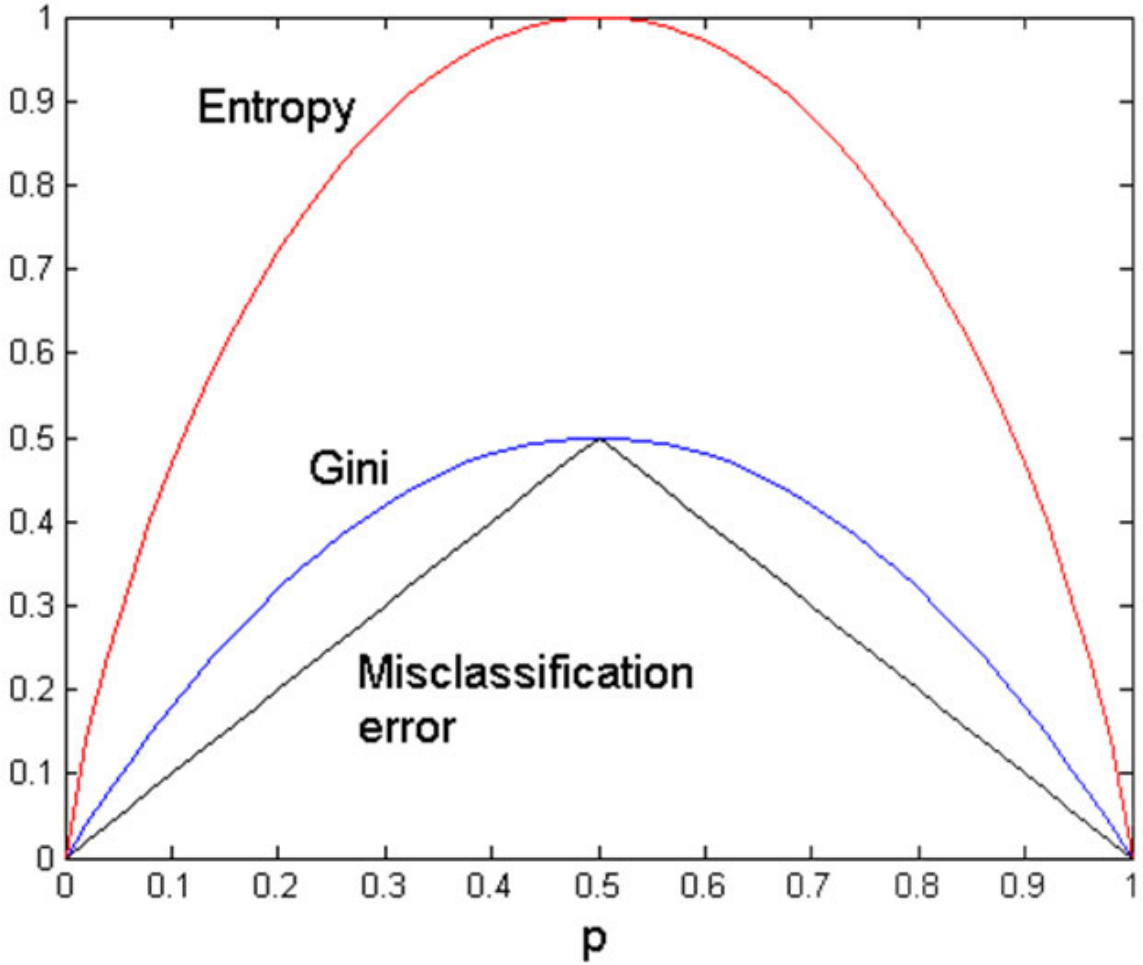
P(C1) = 2/6 P(C2) = 4/6

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

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Comparison among Impurity Measures

- For a 2-class problem.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Tree Induction

- Greedy strategy:
  - Split the records based on an attribute test that optimizes certain criterion.
- Issues:
  - Determine how to split the records.
    - How to specify the attribute test condition?
    - How to determine the best split?
    - Determine when to stop splitting.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Stopping Criteria for Tree Induction

- Stop expanding a node when all the records belong to the same class.
- Stop expanding a node when all the records have similar attribute values.
- Early termination (to be discussed later).

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Decision Tree Based Classification

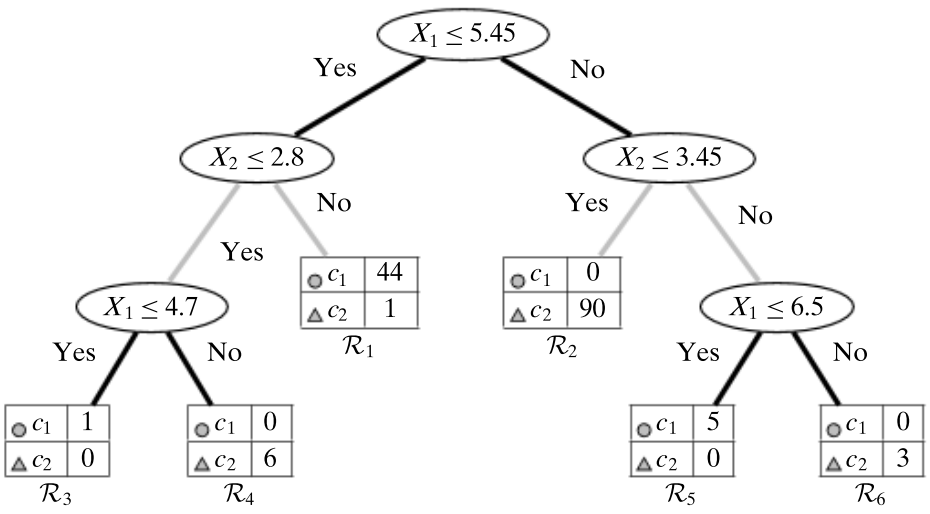
- Advantages:
  - Inexpensive to construct.
  - Extremely fast at classifying unknown records.
  - Easy to interpret for small-sized trees.
  - Accuracy is comparable to other classification techniques for many simple data sets.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

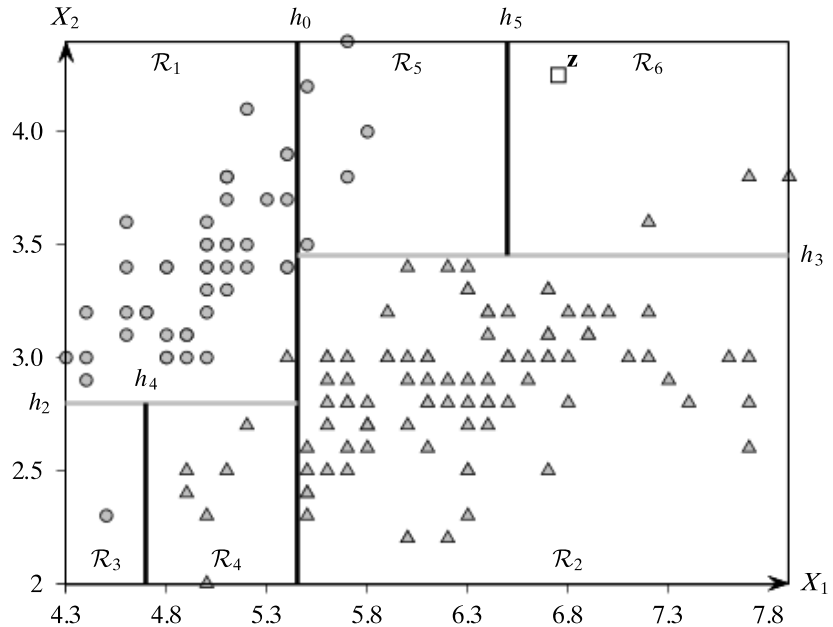
1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

Recursive Partitioning via Axis-Parallel Hyperplanes



(b) Decision Tree



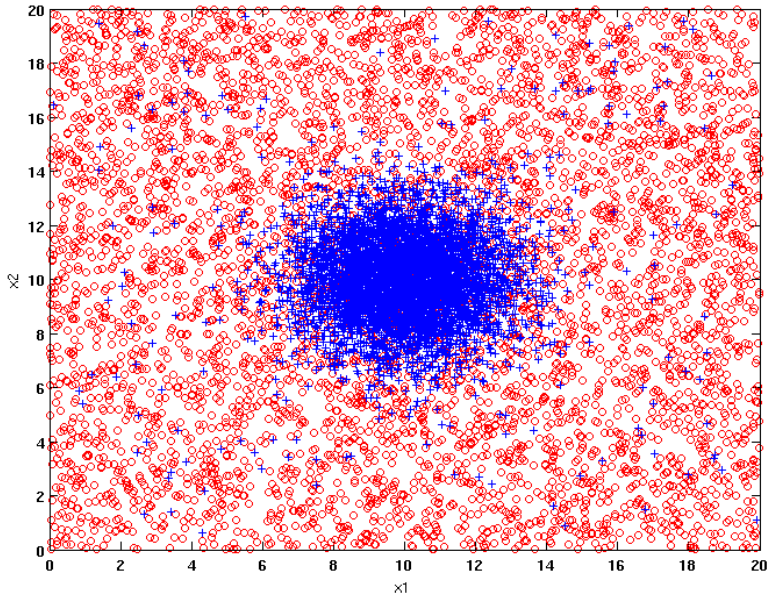
(a) Recursive Splits

Figure 19.1. Decision trees: recursive partitioning via axis-parallel hyperplanes.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

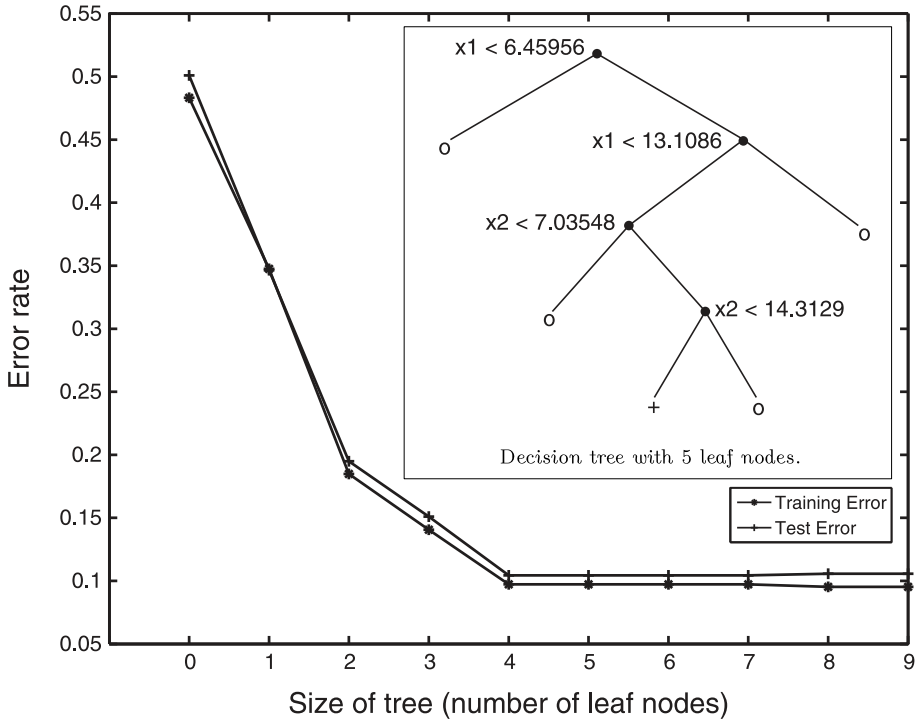
Example Data Set

- Consider a two class problem.
- + : 5400 instances
  - 5000 instances generated from a Gaussian centered at (10,10).
  - 400 noisy instances added.
- o : 5400 instances
  - Generated from a uniform distribution.
- 10% of the data used for training and 90% of the data used for testing.

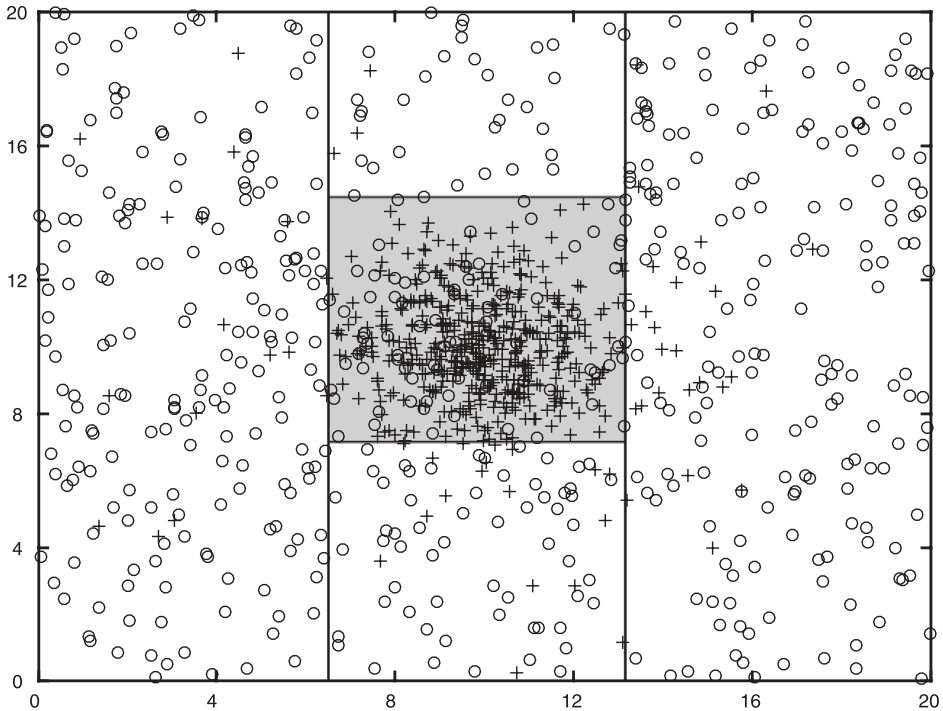


All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Underfitting and Overfitting



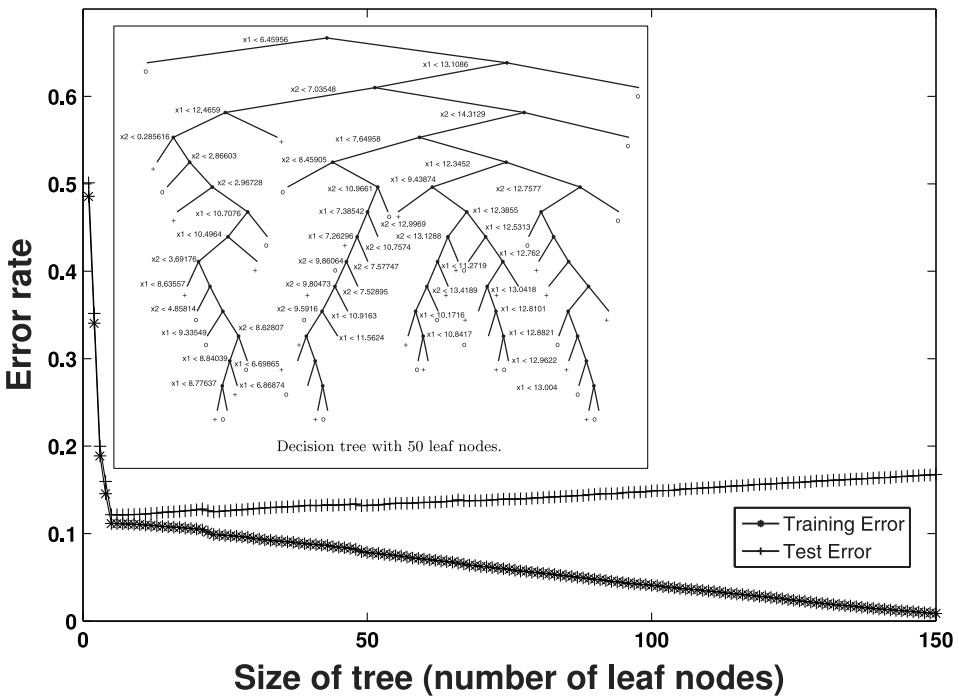
(a) Varying tree size from 1 to 8.



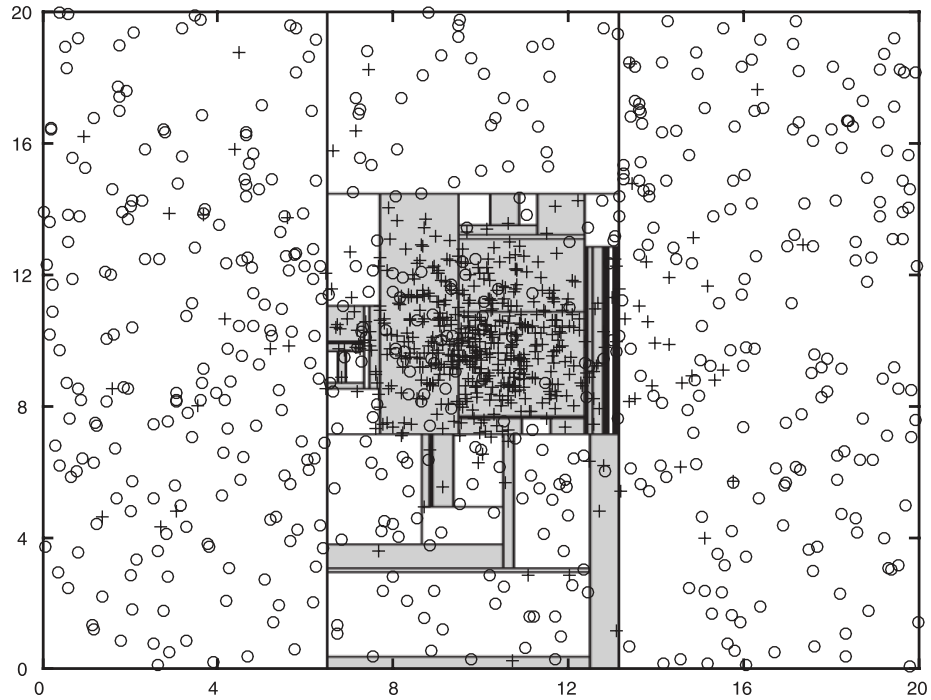
Decision boundary for tree with 5 leaf nodes.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Underfitting and Overfitting



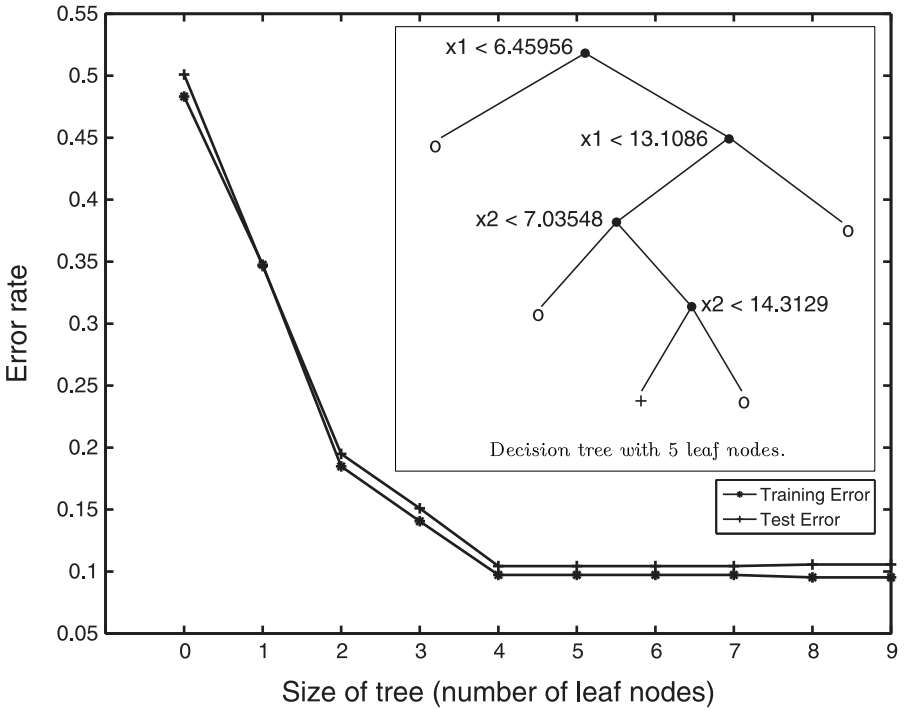
(b) Varying tree size from 1 to 150.

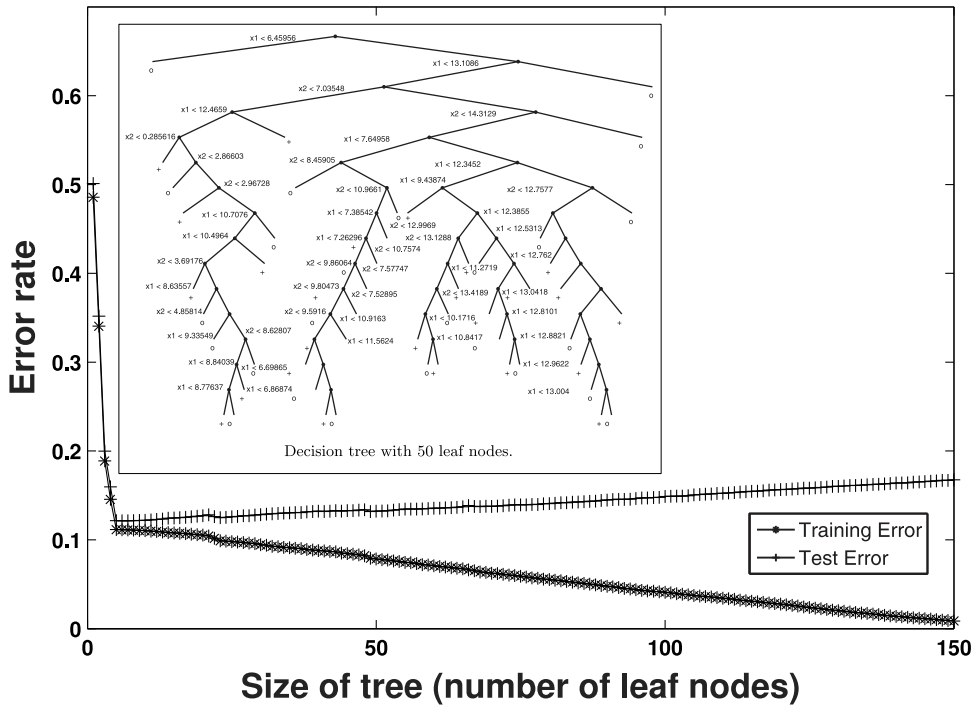


Decision boundary for tree with 50 leaf nodes.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Underfitting and Overfitting

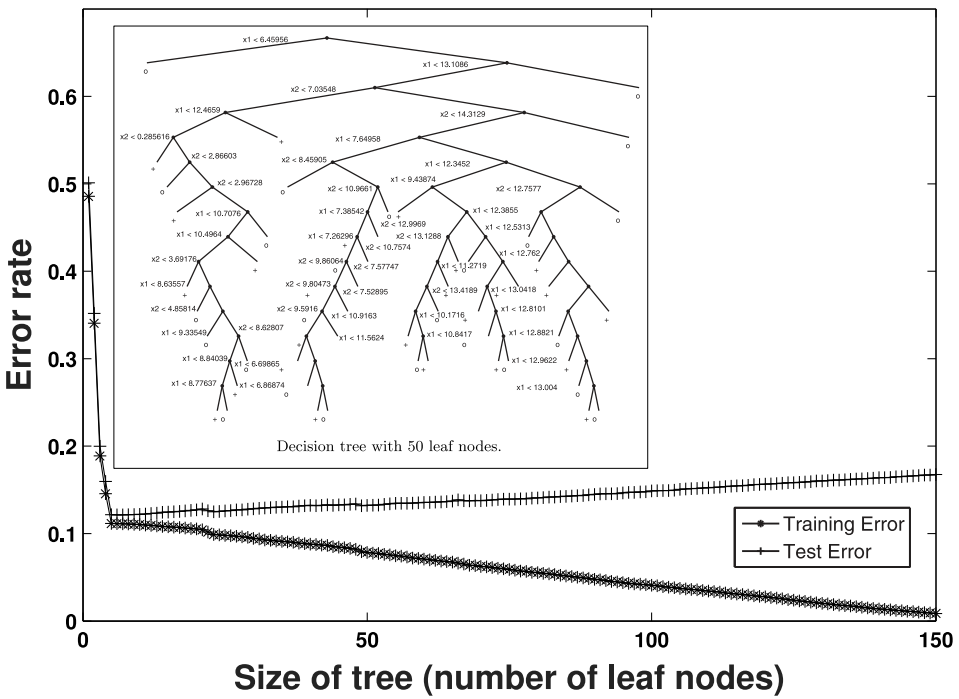




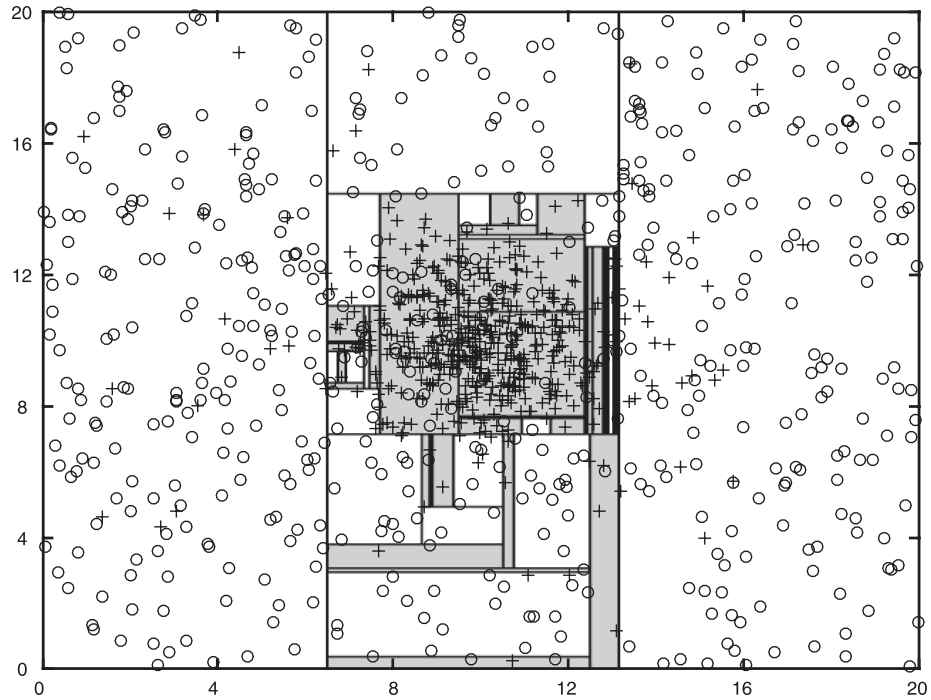
- As the model becomes more and more complex, test errors can start increasing even though training error may be decreasing.
- Underfitting: when model is too simple, both training and test errors are large.
- Overfitting: when model is too complex, training error is small but test error is large.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Address Overfitting



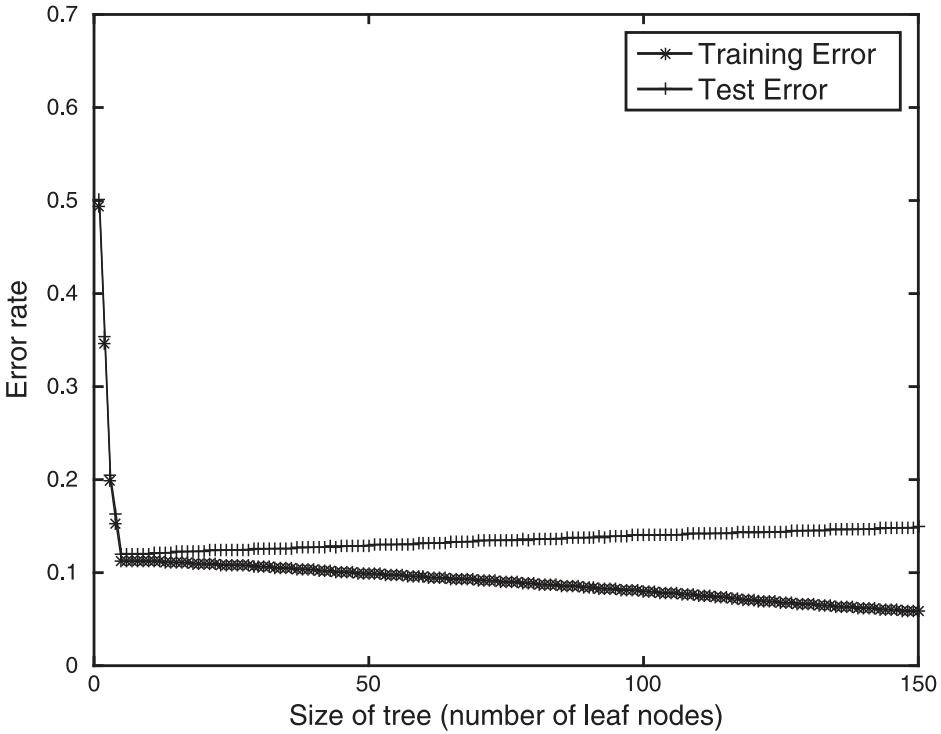
(b) Varying tree size from 1 to 150.



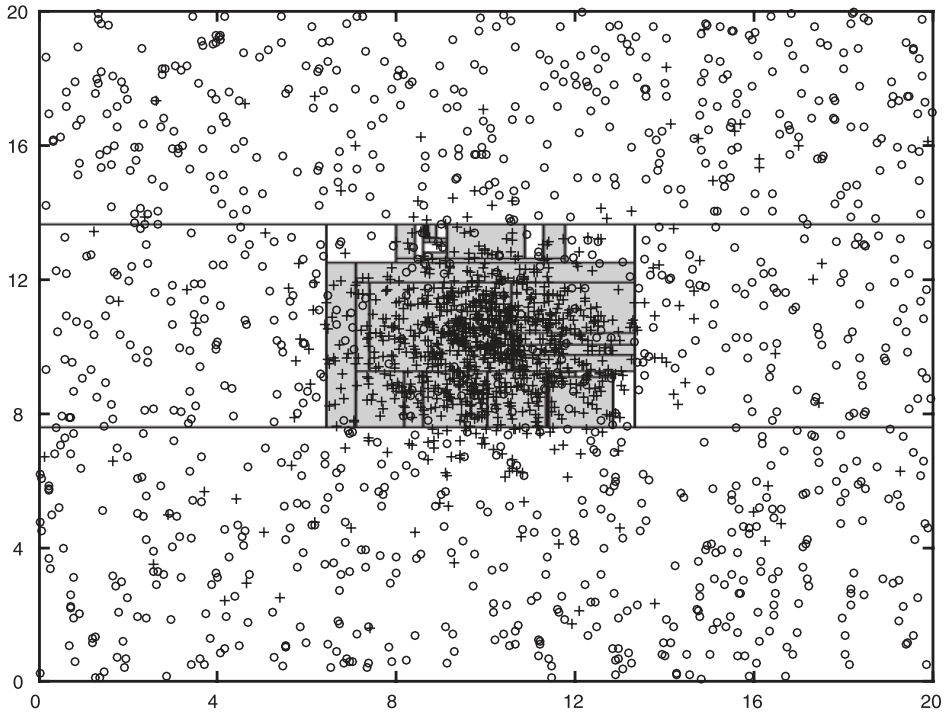
Decision boundary for tree with 50 leaf nodes.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Address Overfitting



Training and test error rates using 20% data for training



Decision boundary for tree with 50 leaf nodes using 20% nodes using 20% data for training.

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Address Overfitting

- Pre-Pruning (Early Stopping Rule):
  - Stop the algorithm before it becomes a fully-grown tree.
  - Typical stopping conditions for a node:
    - Stop if all instances belong to the same class.
    - Stop if all the attribute values are the same.
  - More restrictive conditions:
    - Stop if number of instances is less than some user-specified threshold.
    - Stop if class distribution of instances are independent of the available features (e.g., using χ^2 test).
    - Stop if expanding the current node does not improve impurity measures (e.g., Gini or information gain).

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Address Overfitting

- Post-pruning:
  - Grow decision tree to its entirety.
  - Trim the nodes of the decision tree in a bottom-up fashion.
  - If generalization error improves after trimming, replace sub-tree by a leaf node.
  - Class label of leaf node is determined from majority class of instances in the sub-tree.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

Evaluation Metrics

$$\begin{aligned}\text{Accuracy} &= \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= \frac{a + d}{a + b + c + d}.\end{aligned}$$

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a (True Positive)	b (False Negative)
	Class = No	c (False Positive)	d (True Negative)

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Evaluation Metrics

$$\text{Error Rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{FN + FP}{TP + TN + FP + FN} = \frac{b + c}{a + b + c + d}.$$

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a (True Positive)	b (False Negative)
	Class = No	c (False Positive)	d (True Negative)

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Evaluation Metrics

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{a}{a + c}.$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{a}{a + b}.$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot a}{2 \cdot a + b + c}.$$

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	a (True Positive)	b (False Negative)
	Class = No	c (False Positive)	d (True Negative)

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Methods of Estimation

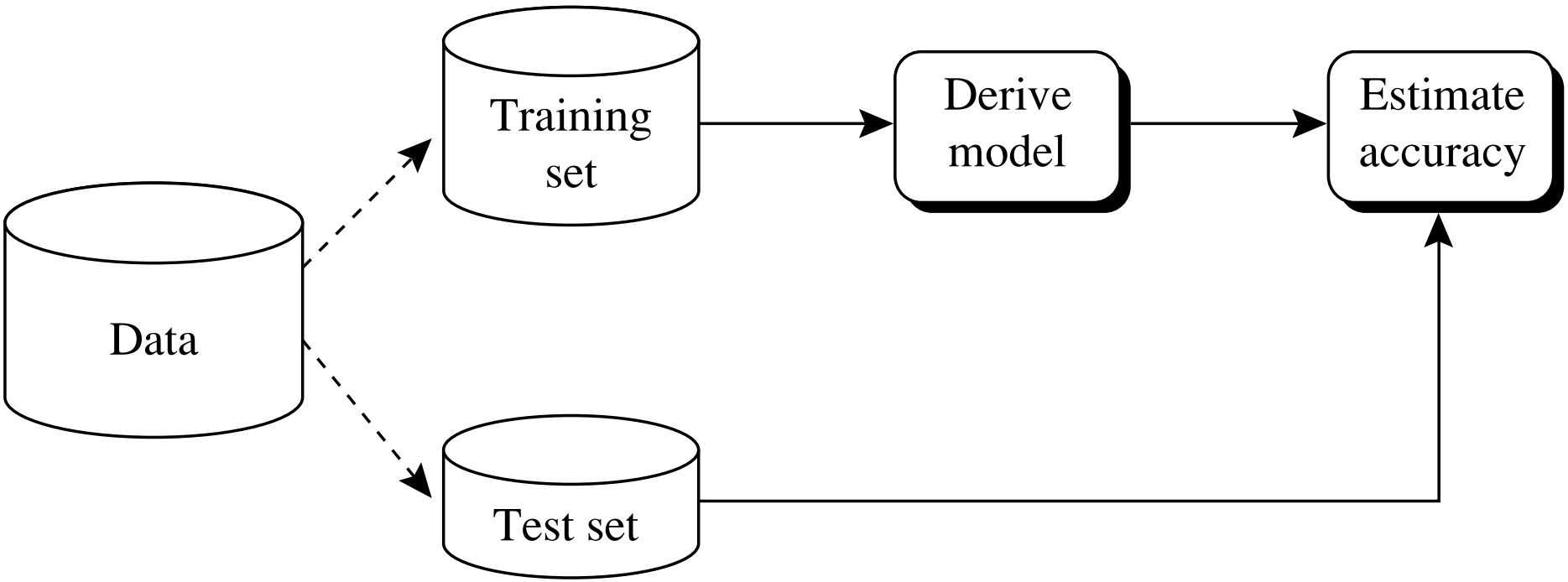


Figure 8.17 Estimating accuracy with the holdout method.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Methods of Estimation

- Holdout:
  - Reserve 2/3 for training and 1/3 for testing.
- Random subsampling:
  - Repeated holdout.
- Cross validation:
  1. Partition data into k disjoint subsets.
  2. k-fold: train on $k-1$ partitions, test on the remaining one.
  - Leave-one-out: $k = n$.
- Bootstrap
  - Sampling with replacement.

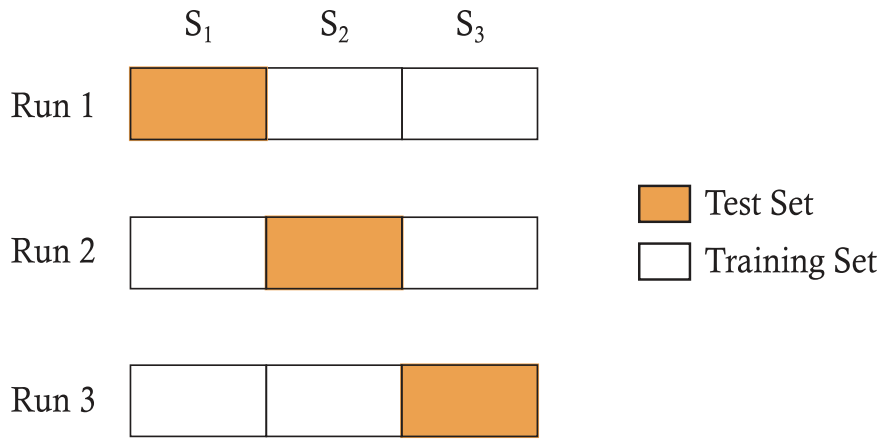


Figure 3.33. Example demonstrating the technique of 3-fold cross-validation.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

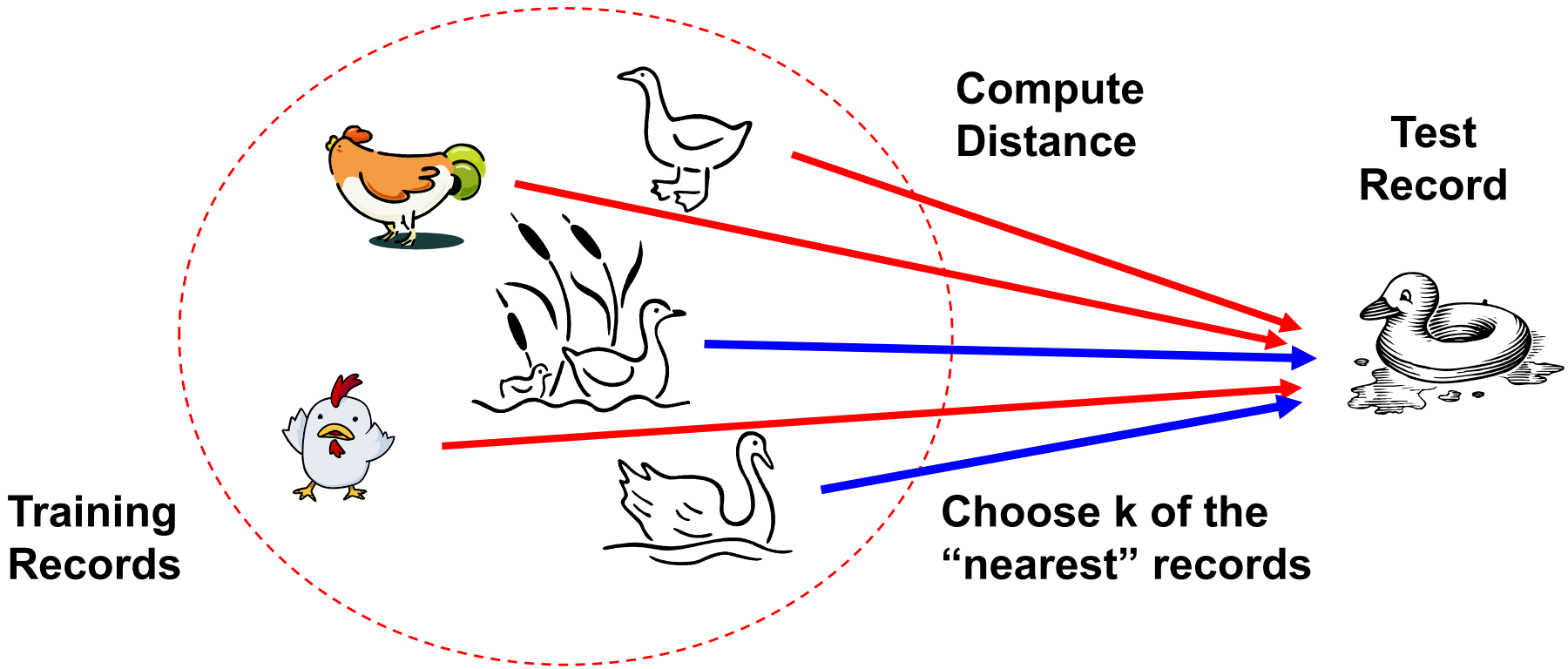
Nearest Neighbor Classifiers

- Eager Learners: Learn a model that maps the input attributes to the class label from the training data.
  - Decision Trees.
- Lazy Learners: Delay modeling the training data until classification of test examples is needed.
  - Rote Classifier: Memorize entire training data and perform classification only if the attributes of a test instance match one of the training examples exactly.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classifiers

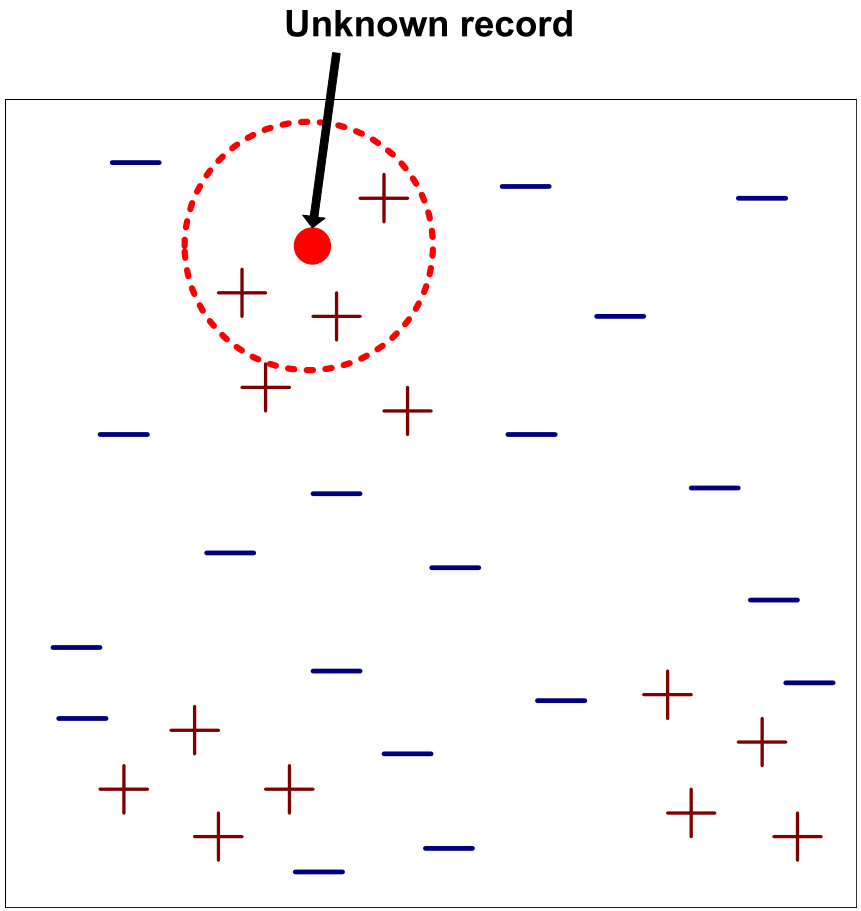
- Basic idea: "If it walks like a duck, quacks like a duck, then it's probably a duck."



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classifiers

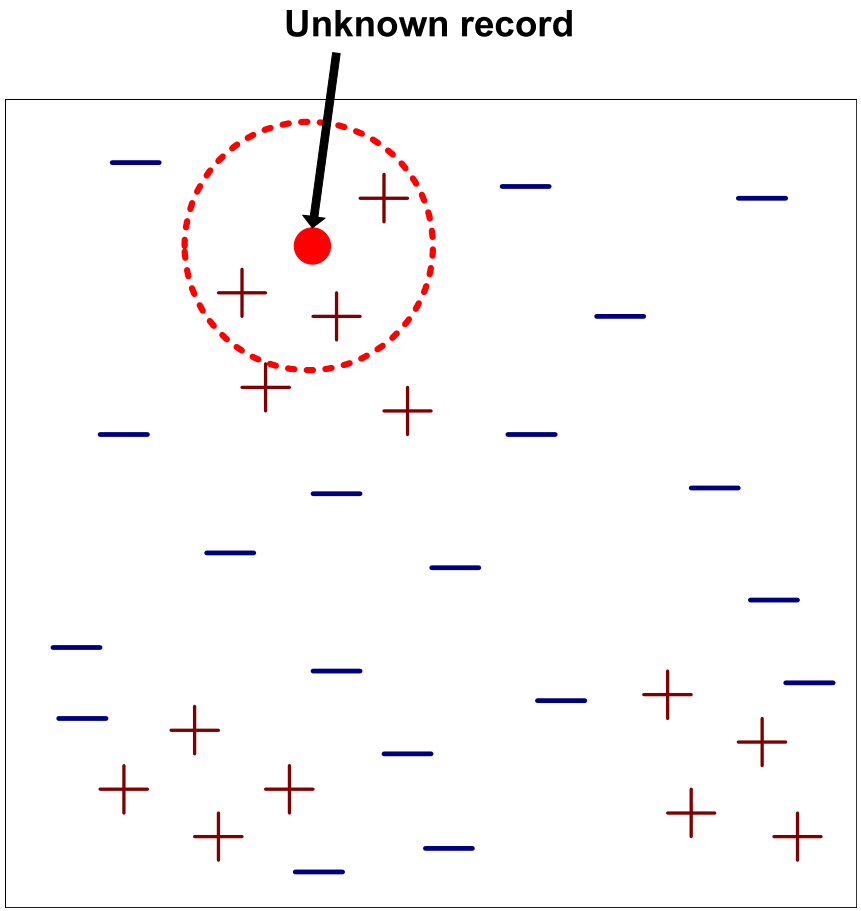
- Requires the following:
  - A set of labeled records.
  - Proximity metric to compute distance/similarity between a pair of records (e.g., Euclidean distance).
  - The value of k, the number of nearest neighbors to retrieve.
  - A method for using class labels of k nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote).



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

How to Determine the Class Label of a Test Sample?

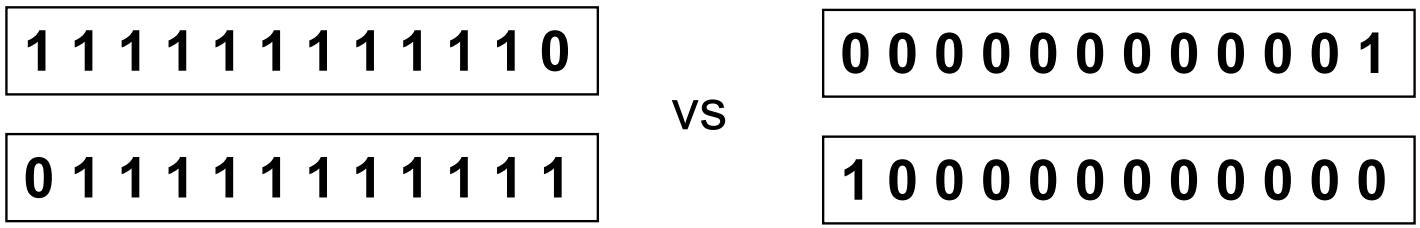
- Take the majority vote of class labels among the k-nearest neighbors.
- Weight the vote according to distance (e.g., with a weight factor, $w = \frac{1}{d^2}$).



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Choice of Proximity Measure Matters

- For documents, cosine is better than correlation or Euclidean.



Euclidean distance = 1.4142 for both pairs, but the cosine similarity measure has different values for these pairs.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

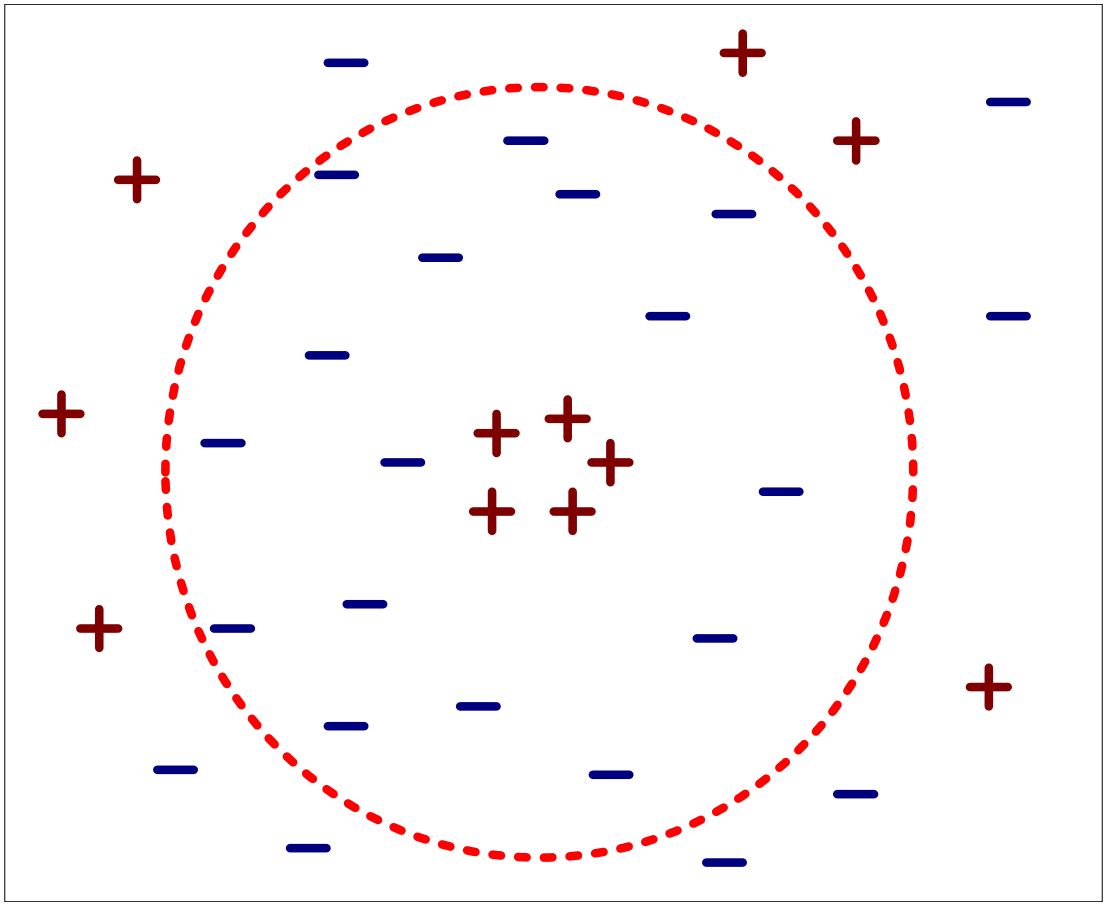
Nearest Neighbor Classification

- Data preprocessing is often required:
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes.
  - Example:
    - Height of a person may vary from 1.5 m to 1.8 m.
    - Weight of a person may vary from 90 lb to 300 lb.
    - Income of a person may vary from 10 K to 1 M.
  - Time series are often standardized to have
    0 means a standard deviation of 1.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classification

- Choosing the value of k:
  - If k is too small, sensitive to noise points.
  - If k is too large, neighborhood may include points from other classes.

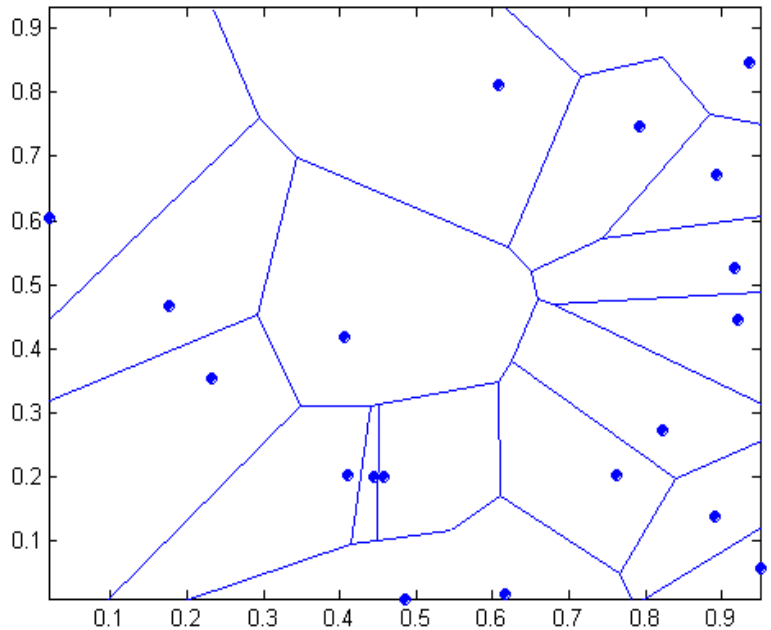


All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Nearest Neighbor Classification

- Nearest neighbor classifiers are local classifiers.
- They can produce decision boundaries of arbitrary shapes.

1-NN decision boundary is a Voronoi Diagram.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1 Announcements and References
- Administrative
- References for Today's Lecture

2 Classification
- Introduction
- Decision Trees
  - Definitions
  - Decision Tree Induction
  - Underfitting and Overfitting
  - Evaluation
- Nearest Neighbor Classifiers
- Other Classification Methods

Other Classification Methods — Rule-Based Classifier

- Classify records by using a collection of "if...then..." rules.
- Rule: $(\text{Condition}) \rightarrow y$
  - Condition is a conjunction of tests on attributes.
  - y is the class label.
- Examples of classification rules:
  - $(\text{BloodType} = \text{Warm}) \wedge (\text{LayEggs} = \text{Yes}) \rightarrow \text{Birds}$.
  - $(\text{TaxableIncome} < 50\text{K}) \wedge (\text{Refund} = \text{Yes}) \rightarrow (\text{Evade} = \text{No})$.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Classification Methods — Naïve Bayes

- In many applications, the class label of a test record cannot be predicted with certainty even though its attribute set is identical to some of the training examples.
- This situation may arise because of noisy data or the presence of certain confounding factors that affect classification but are not included in the analysis.
- Naïve Bayes is based on the Bayes Theorem (with additional assumption on independence):

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}. \tag{19}$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Classification Methods — Support Vector Machines

- Support vector machines (SVMs), is a method for the classification of both linear and nonlinear data.
- It uses a nonlinear mapping to transform the original training data into a higher dimension.
- Within this new dimension, it searches for the linear optimal separating hyperplane (i.e., a "decision boundary" separating the tuples of one class from another).
- With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.
- The SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors).

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Other Classification Methods — Support Vector Machines

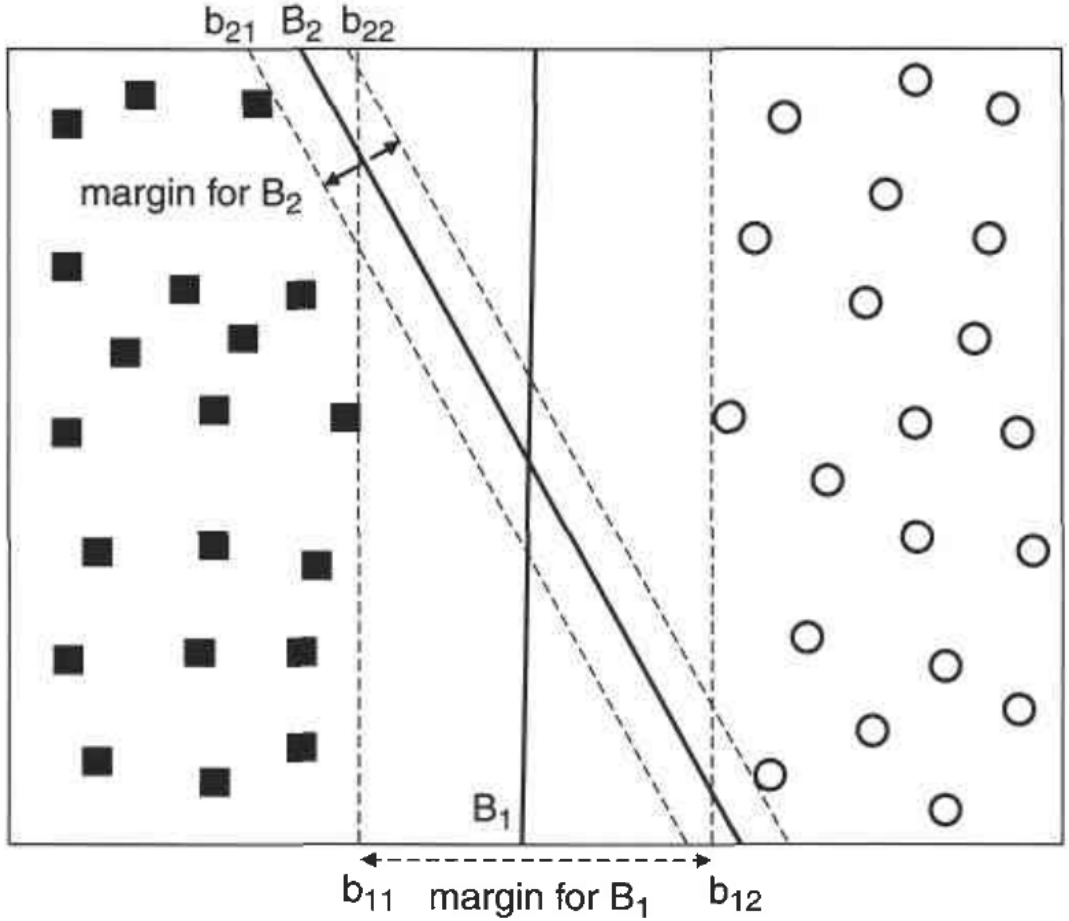


Figure 5.22. Margin of a decision boundary.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Other Classification Methods — Artificial Neural Networks

- Artificial neural networks (ANNs) or simply neural networks are inspired by biological neuronal networks.
- ANNs are comprised of abstract neurons that try to mimic real neurons at a very high level.
- ANNs can be described via a weighted directed graph $G = (V, E)$, with each node representing a neuron, and each directed edge representing a synaptic to dendritic connection.
- The weight of the edge denotes the synaptic strength.

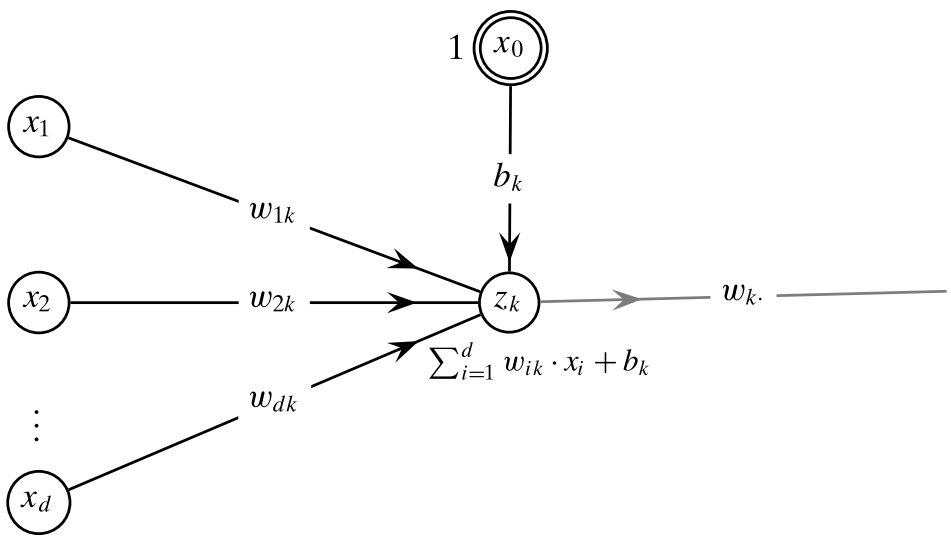


Figure 25.1. Artificial neuron: aggregation and activation.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

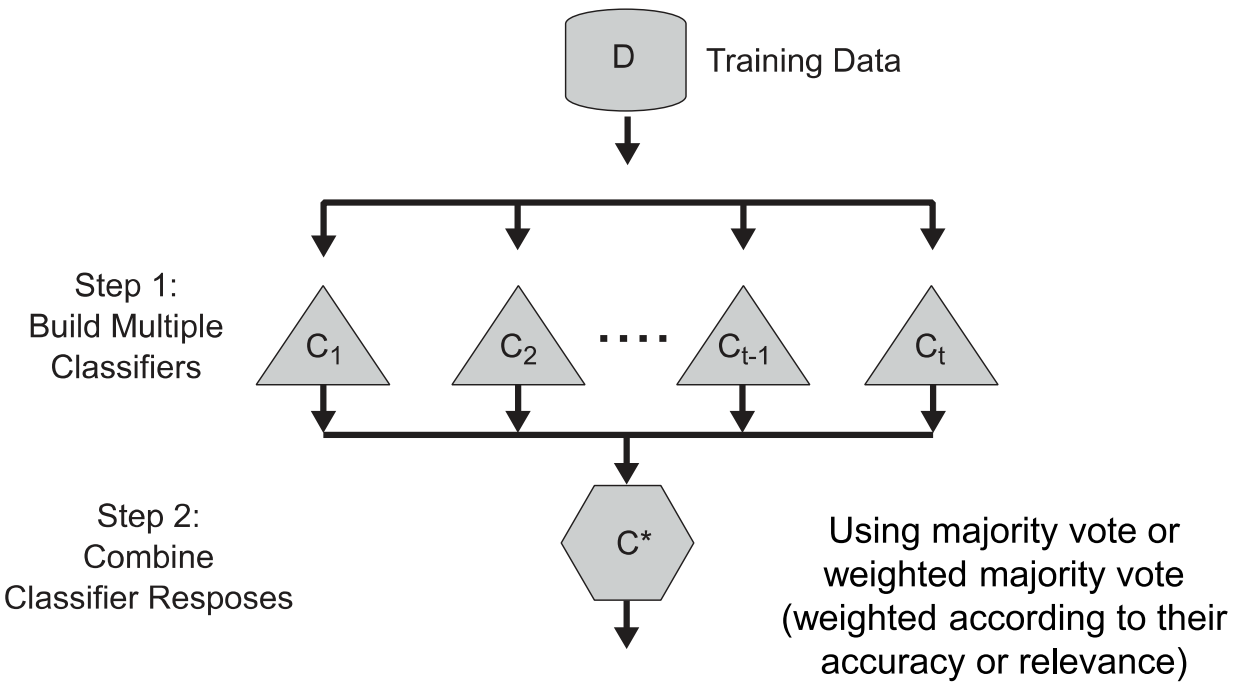
Other Classification Methods — Artificial Neural Networks

- ANNs are characterized by the type of activation function used to generate an output, and the architecture of the network in terms of how the nodes are interconnected.
- For example,
  - Is the graph is a directed acyclic graph or has cycles?
  - Is the graph is layered or not, and so on.
- It is important to note that a neural network is designed to represent and learn information by adjusting the synaptic weights.
- ANNs given enough hidden units and enough training samples, can closely approximate any function.

All from: 1. Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press. 2. Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Other Classification Methods — Ensemble Methods

- Ensemble methods improve classification accuracy by aggregating the predictions of multiple classifiers.
- An ensemble method constructs a set of base classifiers from training data and performs classification by taking a vote on the predictions made by each base classifier.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.