

ICS 421 Spring 2010

Data Mining 2

Asst. Prof. Lipyeow Lim
Information & Computer Science Department
University of Hawaii at Manoa

Beyond Co-Occurrences

How do you obtain rules to predict future high risk customers?

- Suppose you have a car-insurance company and you have collected the following historical data
- Example: if $16 < \text{age} < 25$ AND cartype is sports or truck, then risk is high

Age	CarType	HighRisk
23	Sedan	F
30	Sports	F
36	Sedan	F
25	Truck	T
30	Sedan	F
23	Truck	T
30	Truck	F
25	Sports	T
18	Sedan	F

Predictor attributes

Dependent attribute

numerical attributes

categorical attributes

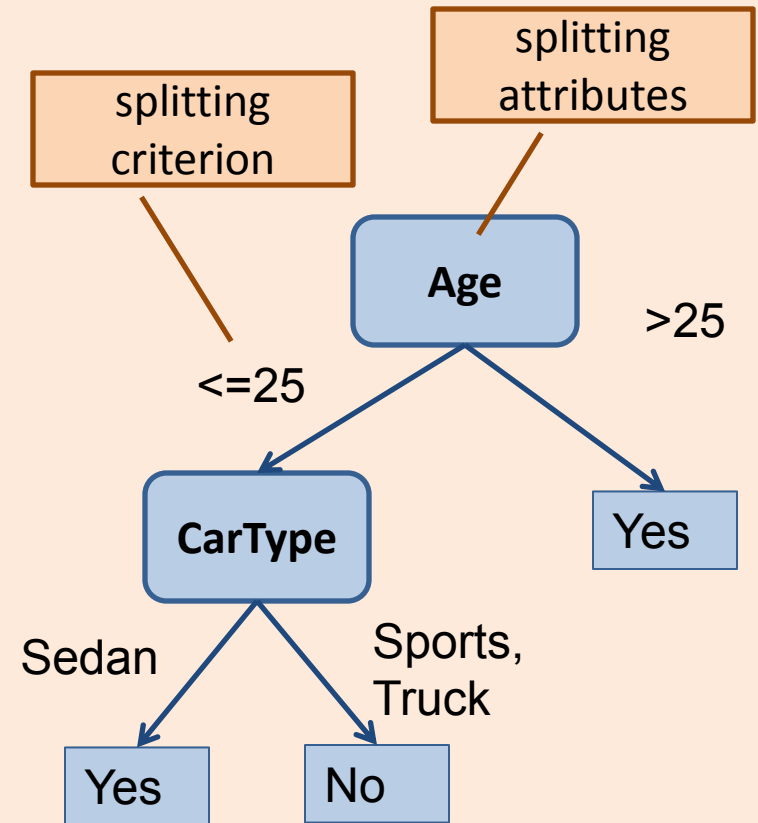
Classification & Regression Rules

- **Classification rules**
 - Dependent attribute is categorical
- **Regression rules**
 - Dependent attribute is numerical
- **Support for $C1 \rightarrow C2$**
 - Percentage of tuples that satisfy $C1$ and $C2$
- **Confidence for $C1 \rightarrow C2$**
 - Percentage of tuples satisfying $C1$ that also satisfy $C2$.

Age	CarType	Expenditure
23	Sedan	200
30	Sports	150
36	Sedan	300
25	Truck	220
30	Sedan	400
23	Truck	80
30	Truck	100
25	Sports	125
18	Sedan	500

Decision Trees

- Classification rules that can be structured as trees.
- Trees for regression rules are called regression trees
- A *decision tree* T encodes d (a classifier or regression function) in form of a tree.
- A node t in T without children is called a *leaf node*. Otherwise t is called an *internal node*.

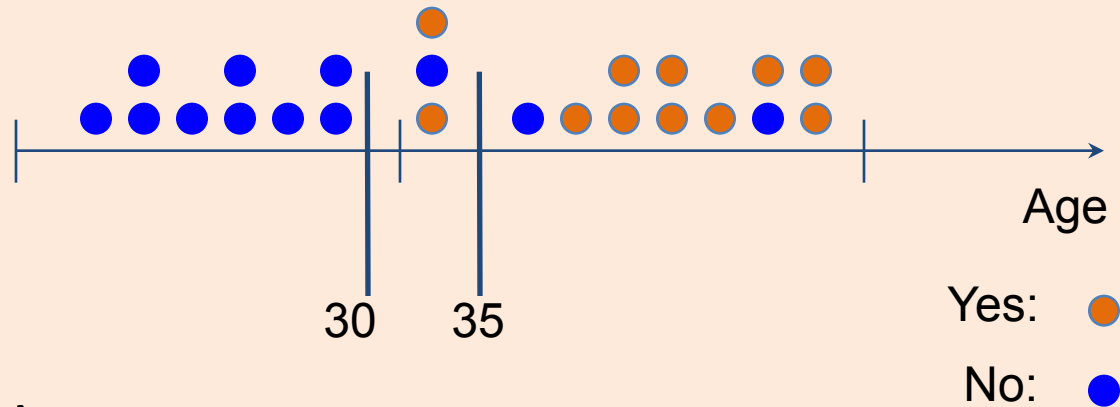


Top-down Decision Tree Algorithm

- Examine training database and find best splitting predicate for the root node
- Partition training database
- Recurse on each child node

```
BuildTree(Node  $t$ , Training database  $D$ ,  
           Split Selection Method  $\mathbf{S}$ )  
(1) Apply  $\mathbf{S}$  to  $D$  to find splitting criterion  
(2) if ( $t$  is not a leaf node)  
(3)     Create children nodes of  $t$   
(4)     Partition  $D$  into children partitions  
(5)     Recurse on each partition  
(6) endif
```

Split Selection Method

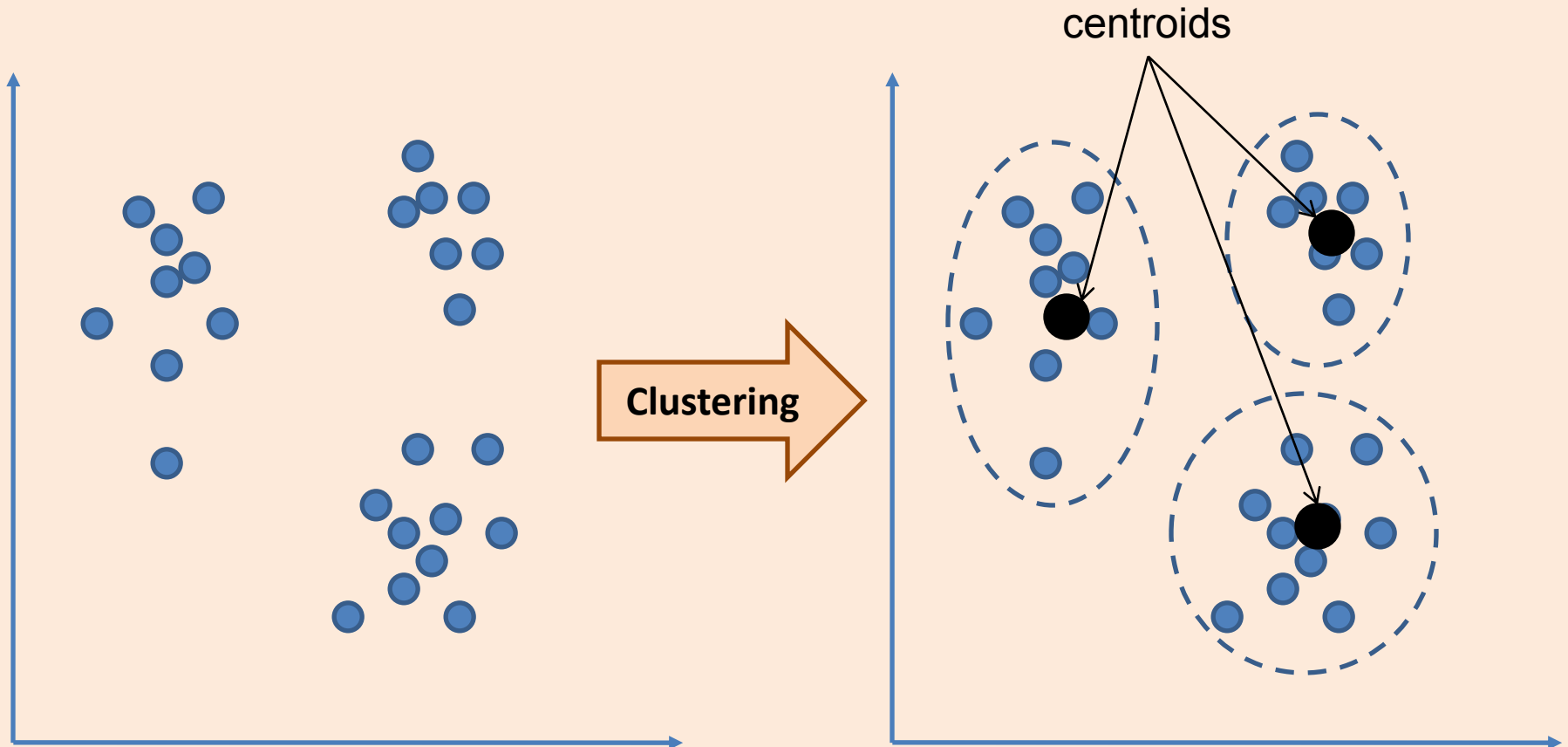


- Two decisions:
 - What is the splitting attribute
 - What is the splitting criterion
- Numerical or ordered attributes:
 - Find a split point that separates the (two) classes
- Categorical attributes:
 - evaluate all possible partitions

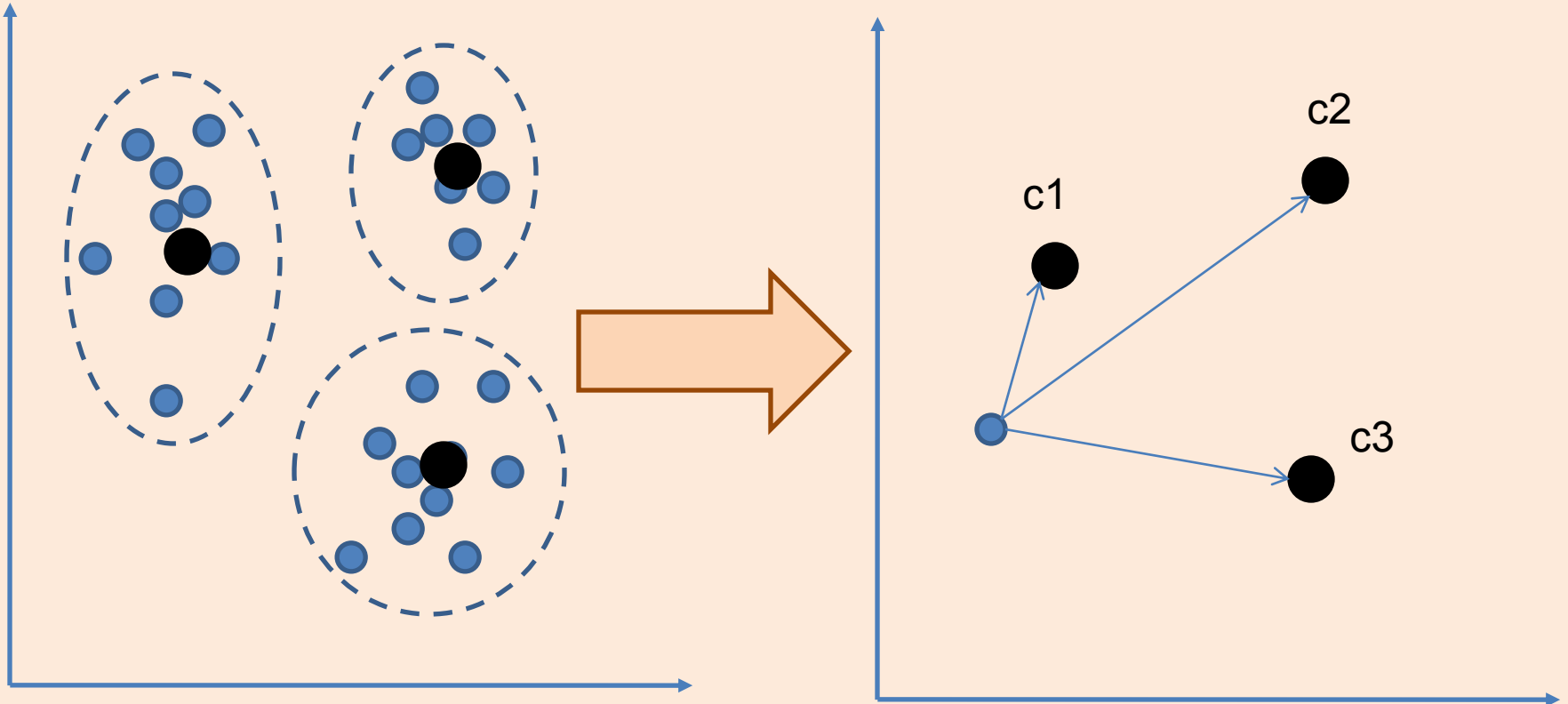
Clustering

- Decision trees learn to predict class labels given predictor attributes -- **supervised learning**.
- What if no class labels are available ?
 - Find patterns via **unsupervised learning**
- **Clustering** is the process of organizing objects into groups whose members are similar in some way
 - Given:
 - Data Set D (training set)
 - Similarity/distance metric/information
 - Find:
 - Partitioning of data
 - Groups of similar/close items

Clustering Visually



Why are clusters useful ?



Applications

- *Marketing*
 - finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- *Biology*
 - classification of plants and animals given their features;
- *Insurance*
 - identifying groups of motor insurance policy holders with a high average claim cost
 - identifying frauds;
- *Earthquake studies*
 - clustering observed earthquake epicenters to identify dangerous zones;
- *WWW*
 - document classification; clustering weblog data to discover groups of similar access patterns.

Minkowski Distance (L_p Norm)

- Consider two records $x=(x_1, \dots, x_d)$, $y=(y_1, \dots, y_d)$:

$$d(x, y) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \dots + |x_d - y_d|^p}$$

Special cases:

- $p=1$: Manhattan distance

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_p - y_p|$$

- $p=2$: Euclidean distance

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_d - y_d)^2}$$

Properties of Distances: Metric Spaces

- A metric space is a set S with a global distance function d . For every two points x, y in S , the distance $d(x,y)$ is a nonnegative real number.
- A metric space must also satisfy
 - $d(x,y) = 0$ iff $x = y$
 - $d(x,y) = d(y,x)$ (**symmetry**)
 - $d(x,y) + d(y,z) \geq d(x,z)$ (**triangle inequality**)

Clustering: Informal Problem Definition

Input:

- A data set of N records each given as a d -dimensional data feature vector.

Output:

- Determine a natural, useful “partitioning” of the data set into a number of (k) clusters and noise such that we have:
 - High similarity of records within each cluster (**intra-cluster similarity**)
 - Low similarity of records between clusters (**inter-cluster similarity**)

Clustering Algorithms

- Partitioning-based clustering
 - K-means clustering
 - K-medoids clustering
 - EM (expectation maximization) clustering
- Hierarchical clustering
 - Divisive clustering (top down)
 - Agglomerative clustering (bottom up)
- Density-Based Methods
 - Regions of dense points separated by sparser regions of relatively low density

K-means Clustering Algorithm

Initialize k cluster centers

Do

1. **Assignment step**: Assign each data point to its closest cluster center
2. **Re-estimation step**: Re-compute cluster centers

While (there are still changes in the cluster centers)

Visualization at:

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

Issues with K-means

Why is K-Means working:

- How does it find the cluster centers?
- Does it find an optimal clustering
- What are good starting points for the algorithm?
- What is the right number of cluster centers?
- How do we know it will terminate?