Data Mining

DONG Chengzu东承祖

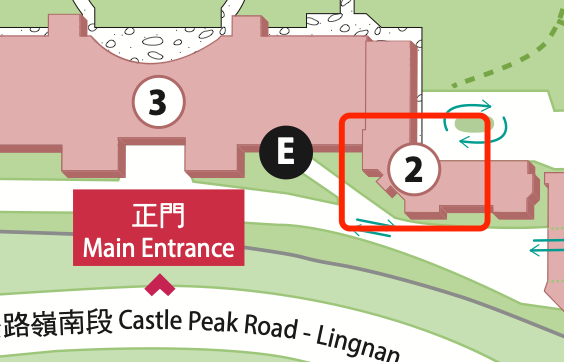
Email: [chengzudong@ln.edu.hk](mailto:chengzudong@ln.edu.hk)

Phone: 26168135

# General

meet him at office time at Wong Administration Building

Office Hour: all day long



## Assessment

attendance and participation 5%

Assignments 30%

Group Project 30%

Examination 35%

GAI is NOT permitted in assignment, case study and examination in this course. Files will be examined by Turnitin System to detect plagiarism.

## Required/Essential Readings

I think better to practice other than read of this course.

**目录**

[第1章 General - 1 -](#_Toc7930)

[1.1 Assessment - 1 -](#_Toc24835)

[1.2 Required/Essential Readings - 1 -](#_Toc2476)

[第2章 Week 1 Lecture 1 - 4 -](#_Toc3986)

[2.1 Why Data Mining - 4 -](#_Toc22961)

[2.2 Origin of data mining - 4 -](#_Toc5524)

[2.3 Data Mining Process - 4 -](#_Toc18968)

[2.4 Data Mining Tasks - 5 -](#_Toc3894)

[第3章 Week 2 Lecture 2 - 6 -](#_Toc1616)

[3.1 What is Data - 6 -](#_Toc16977)

[3.2 Types of Attributes and How to recognize attributes - 6 -](#_Toc3151)

[3.3 Important Characteristics of Data - 7 -](#_Toc20100)

[3.4 Types of data sets - 7 -](#_Toc6348)

[3.5 data quality problems - 7 -](#_Toc29421)

[3.6 Data Preprocessing: ways to deal data quality problems - 8 -](#_Toc13964)

[第4章 Week 3 Lecture 3 - 9 -](#_Toc16093)

[4.1 Classification Techniques - 9 -](#_Toc4301)

[4.2 Decision Tree Induction - 9 -](#_Toc27903)

[4.3 determine the Best Split - 9 -](#_Toc27221)

[4.4 Comparison among Impurity Measures - 10 -](#_Toc25790)

[4.5 Discussion of Decision Tree - 10 -](#_Toc26034)

[第5章 Week 4 Lecture 4 - 12 -](#_Toc31739)

[5.1 Decision Tree Issues - 12 -](#_Toc28836)

[5.2 Data Fragmentation - 12 -](#_Toc9037)

[5.3 Search Strategy - 12 -](#_Toc29653)

[5.4 Expressiveness - 12 -](#_Toc14689)

[5.5 Tree Replication - 12 -](#_Toc12613)

[5.6 Decision Boundary - 12 -](#_Toc26162)

[5.7 Oblique Decision Trees - 12 -](#_Toc20240)

[5.8 Classification Errors - 12 -](#_Toc9524)

[第6章 Week 5 Lecture 5 - 13 -](#_Toc14719)

[6.1 Association Rule Mining - 13 -](#_Toc7592)

[6.2 Frequent Itemset Definition - 13 -](#_Toc7808)

[6.3 Association Rule Definition - 13 -](#_Toc8878)

[6.4 Association Rule Mining Task - 14 -](#_Toc5622)

[6.5 Brute-force approach - 14 -](#_Toc16596)

[6.6 Apriori principle - 14 -](#_Toc25992)

[6.7 Rule Generation - 15 -](#_Toc11570)

[第7章 Week 6 Lecture 6 - 17 -](#_Toc976)

[7.1 Cluster Analysis - 17 -](#_Toc19029)

[7.2 What is not Cluster Analysis - 17 -](#_Toc24089)

[7.3 Types of Clusterings - 17 -](#_Toc22914)

[7.4 K-means Clustering - 19 -](#_Toc15639)

[7.5 means Objective Function - 19 -](#_Toc25372)

[7.6 Selecting Initial Points - 19 -](#_Toc18481)

[7.7 Limitations of K-means - 19 -](#_Toc30514)

[7.8 Overcoming K-means Limitations - 20 -](#_Toc25694)

[第8章 Week 7 Lecture 7 - 21 -](#_Toc12158)

[8.1 Hierarchical Clustering - 21 -](#_Toc21154)

[8.2 Strengths of Hierarchical Clustering - 21 -](#_Toc26205)

[8.3 Two main types of hierarchical clustering - 21 -](#_Toc21416)

[8.4 Basic algorithm - 21 -](#_Toc29319)

[8.5 How to Define Inter-Cluster Distance - 21 -](#_Toc277)

[8.6 Comparison - 22 -](#_Toc30102)

[8.7 Ward’s Method - 22 -](#_Toc21125)

[8.8 Time and Space requirements - 22 -](#_Toc31634)

[8.9 Problems and Limitations - 22 -](#_Toc13134)

[8.10 Euclidean Distance - 23 -](#_Toc23585)

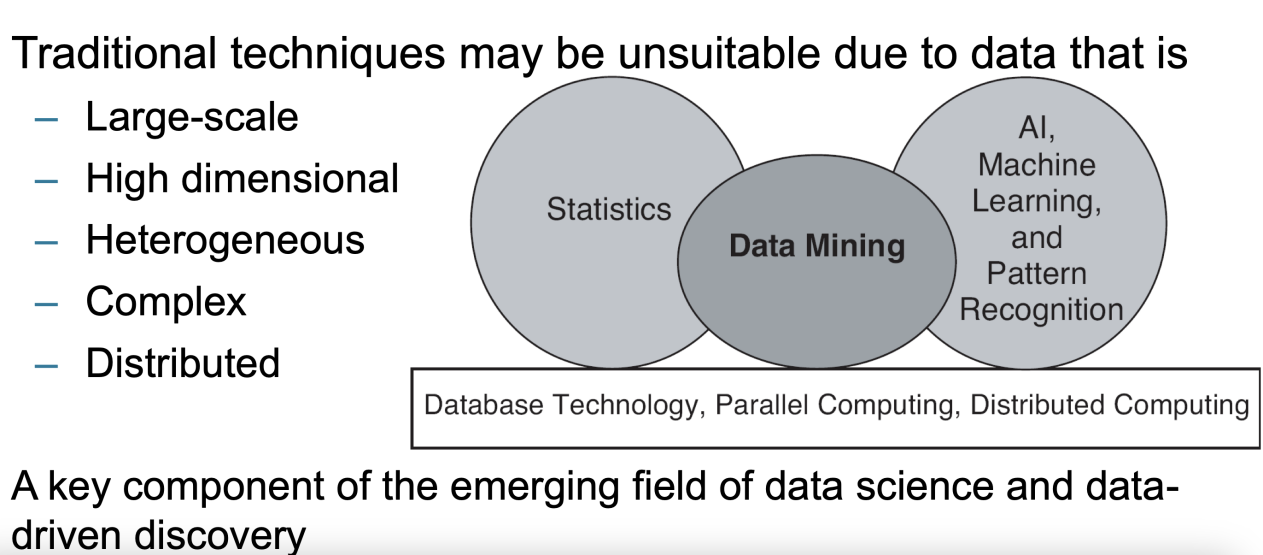
[第9章 Week 8 Lecture 8 - 24 -](#_Toc30590)

# Week 1 Lecture 1

## Why Data Mining

huge amount of data are collected and warehoused at enormous speeds, data mining helps to discover useful information and increase potential opportunities of business use.

## Origin of data mining



## Data Mining Process

|  |  |
| --- | --- |
| Business Understanding  Data Understanding  Data Preparation  Modeling  Evaluation  Deployment | WX20240927-080645@2x |

## Data Mining Tasks

### Prediction Methods

Use some variables to predict unknown or future values of other variables.

### Description Methods

Find human-interpretative patterns that describe the data.

### Methods Classification

Classification [Predictive]

Regression [Predictive]

Clustering [Descriptive]

Association Rule Discovery [Descriptive]

Sequential Pattern Discovery [Descriptive]

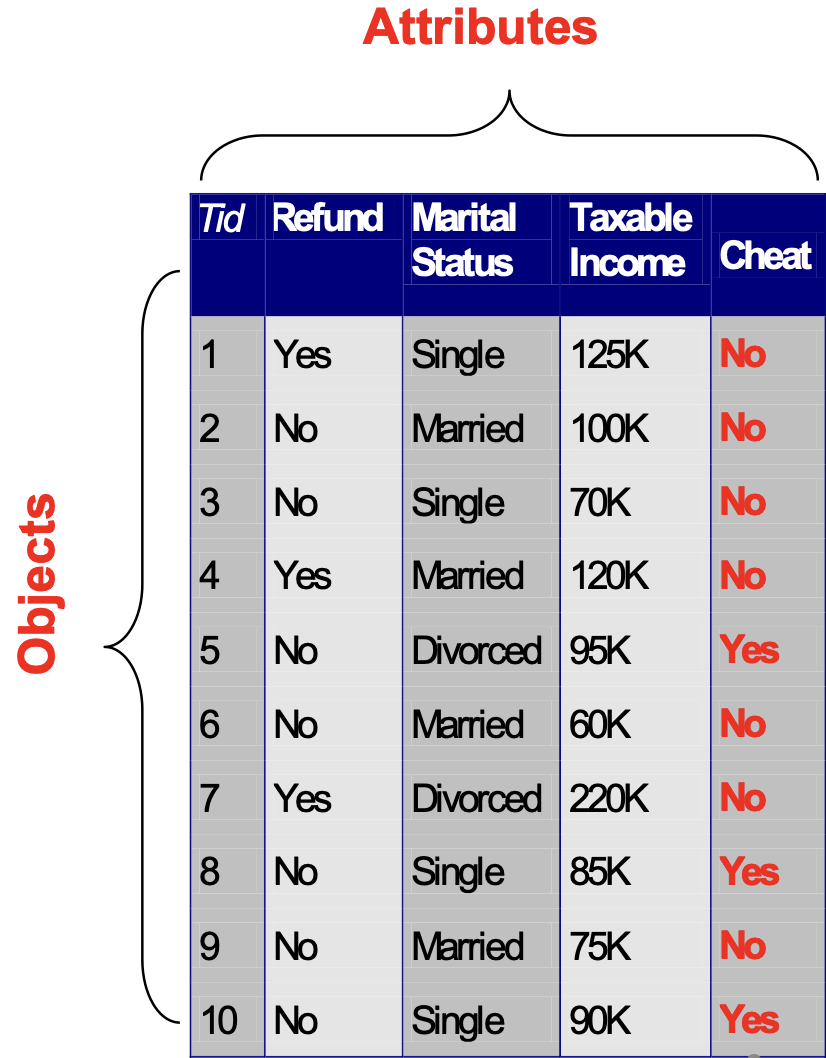
Deviation Detection [Predictive]

1. Week 2 Lecture 2

## What is Data

attribute: An attribute is a property or characteristic of an object

object: A collection of attributes describe an object



## Types of Attributes and How to recognize attributes

Distinctness: = 

Order: < >

Addition: + -

Multiplication: \* /

Nominal attribute: distinctness

Ordinal attribute: distinctness & order

Interval attribute: distinctness, order & addition

Ratio attribute: all 4 properties

## Important Characteristics of Data

Dimensionality: Number of attributes

Sparsity: Only presence counts

Resolution: Patterns depend on the scale

Size: Type of analysis may depend on size of data

## Types of data sets

### Record

Data Matrix

Document Data

Transaction Data

### Graph

World Wide Web

Molecular Structures

### Ordered

Spatial Data

Temporal Data

Sequential Data

Genetic Sequence Data

## data quality problems

Poor data quality negatively affects many data processing efforts

Noise and outliers

Wrong data

Fake data

Missing values

Duplicate data

## Data Preprocessing: ways to deal data quality problems

Aggregation

Sampling

Discretization and Binarization

Attribute Transformation

Dimensionality Reduction

Feature subset selection

Feature creation

1. Week 3 Lecture 3

## Classification Techniques

### Base Classifiers

Decision Tree based Methods

Rule-based Methods

Nearest-neighbor

Naïve Bayes and Bayesian Belief Networks

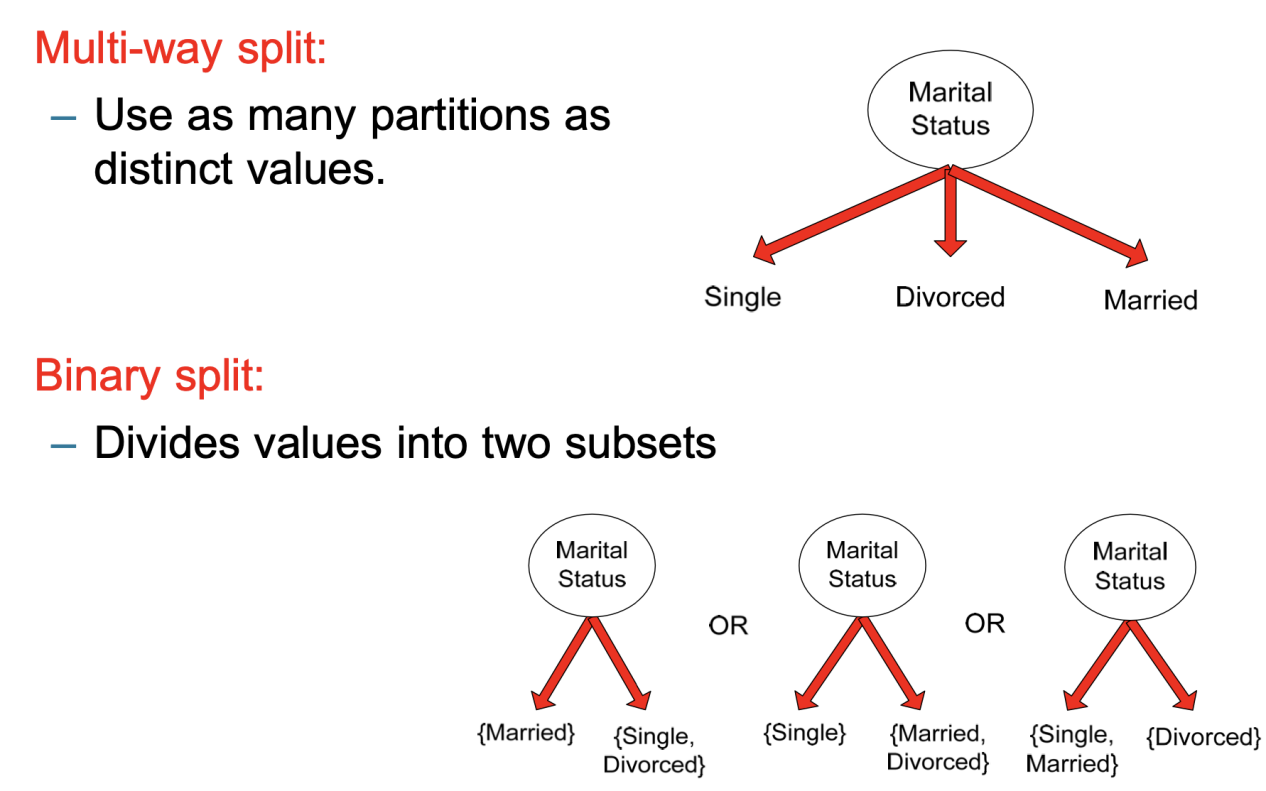
Support Vector Machines

Neural Networks, Deep Neural Nets

### Ensemble Classifiers

Boosting, Bagging, Random Forests

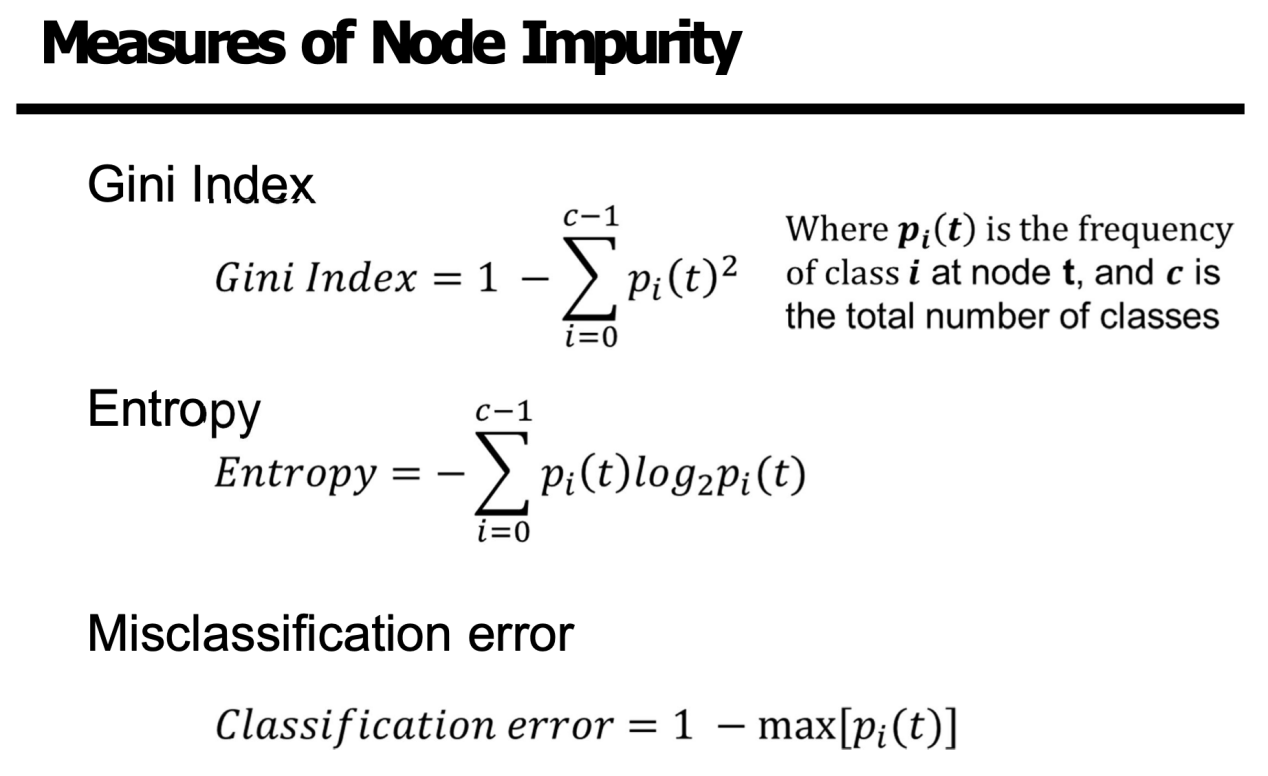
## Decision Tree Induction



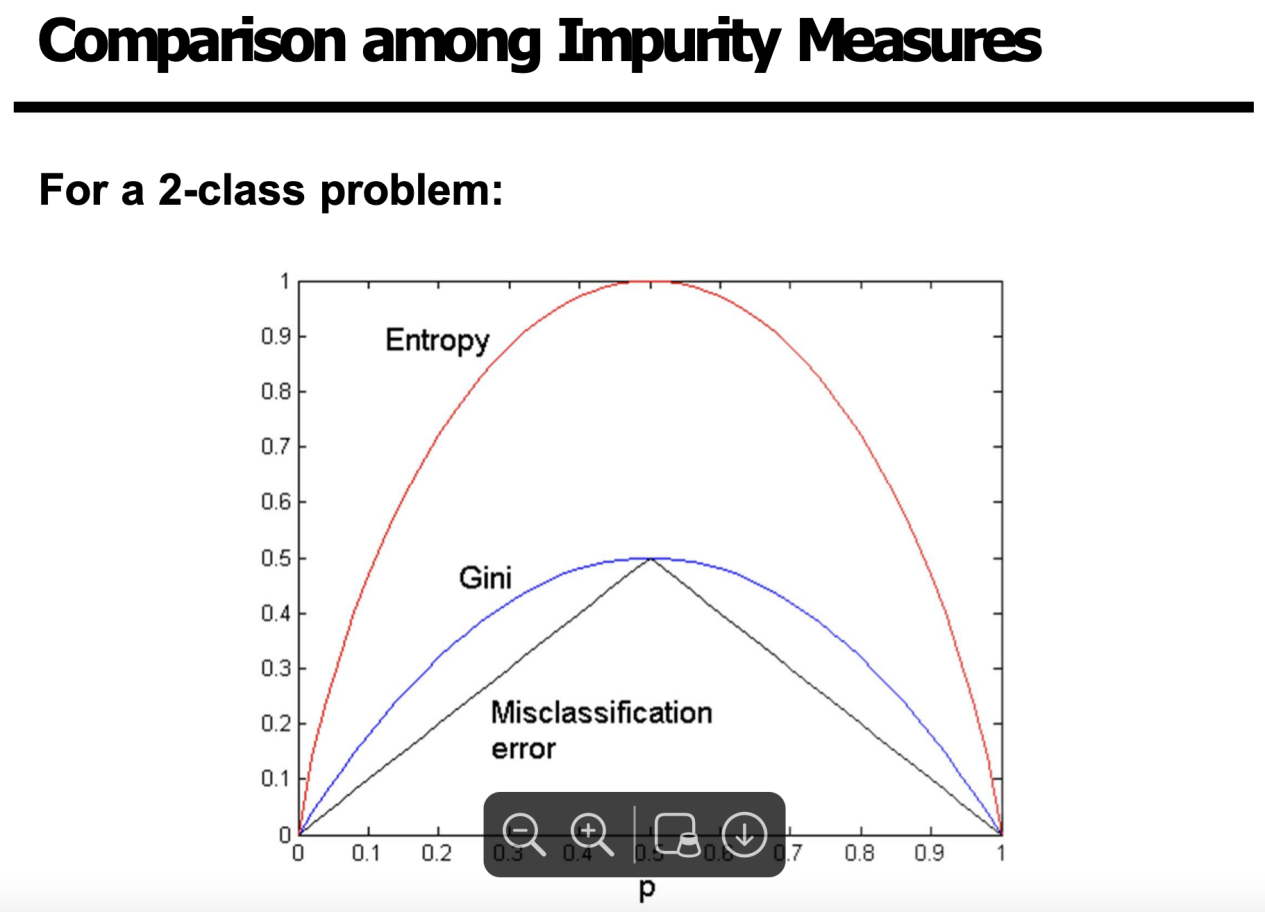
## determine the Best Split

measure of node impurity

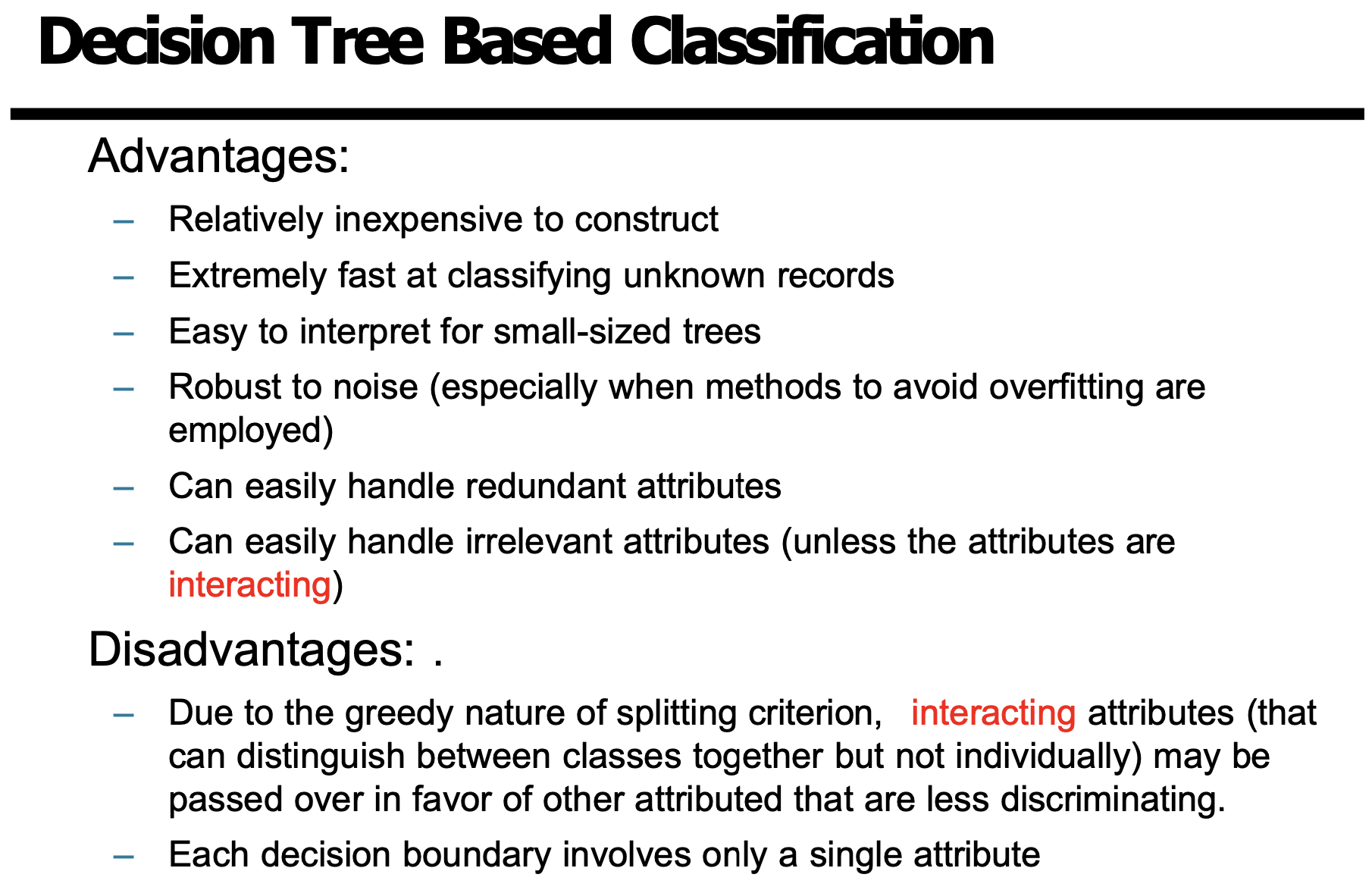
calculating node impurity the lower the better



## Comparison among Impurity Measures



## Discussion of Decision Tree



1. Week 4 Lecture 4

## Decision Tree Issues

Data Fragmentation | Search Strategy | Expressiveness

## Data Fragmentation

Number of instances at the leaf nodes could be too small to make any statistically significant decision

## Search Strategy

Finding an optimal decision tree is NP-hard

## Expressiveness

Decision tree provides expressive representation for learning discrete valued function Not expressive enough for modeling continuous variables

## Tree Replication

Same sub tree appears in multiple branches

## Decision Boundary

Decision boundary is parallel to axes because test condition involves a single attribute at-a-time

## Oblique Decision Trees

Test condition may involve multiple attributes

## Classification Errors

Training errors: Errors committed on the training set

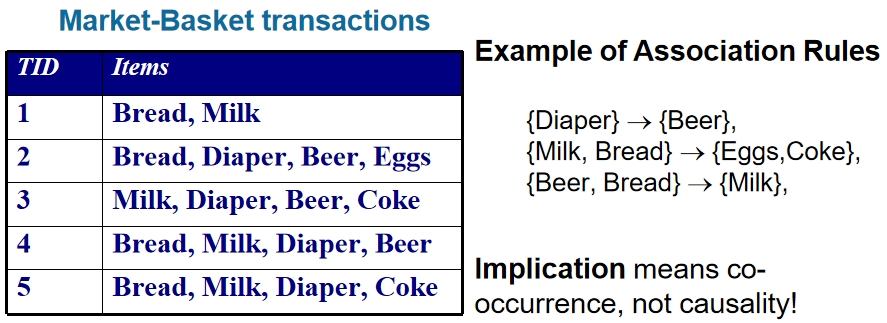
Test errors: Errors committed on the test set

Generalization errors: Expected error of a model over random selection of records from same distribution

1. Week 5 Lecture 5

## Association Rule Mining

Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction



## Frequent Itemset Definition

Itemset : A collection of one or more items Example: {Milk, Bread, Diaper}

k-itemset : An itemset that contains k items

k=5

Support count : Frequency of occurrence of an itemset

({Milk, Bread,Diaper}) = 2

Support : Fraction of transactions that contain an itemset

s({Milk, Bread, Diaper}) = 2/5

Frequent Itemset : An itemset whose support is greater than or equal to a minsup threshold

|  |  |
| --- | --- |
| TID | Items |
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

## Association Rule Definition

### Association Rule

An implication expression of the form X → Y, where X and Y are itemsets

Example : {Milk, Diaper} → {Beer}

### Rule Evaluation Metrics

Support (s) : Fraction of transactions that contain both X and Y

Confidence (c) : Measures how often items in Y appear in transactions that contain X

*s* = s (Milk, Diaper, Beer) = 2 = 0.4

| T | 5

*c* = s (Milk, Diaper, Beer) = 2 = 0.67

s (Milk,Diaper) 3

## Association Rule Mining Task

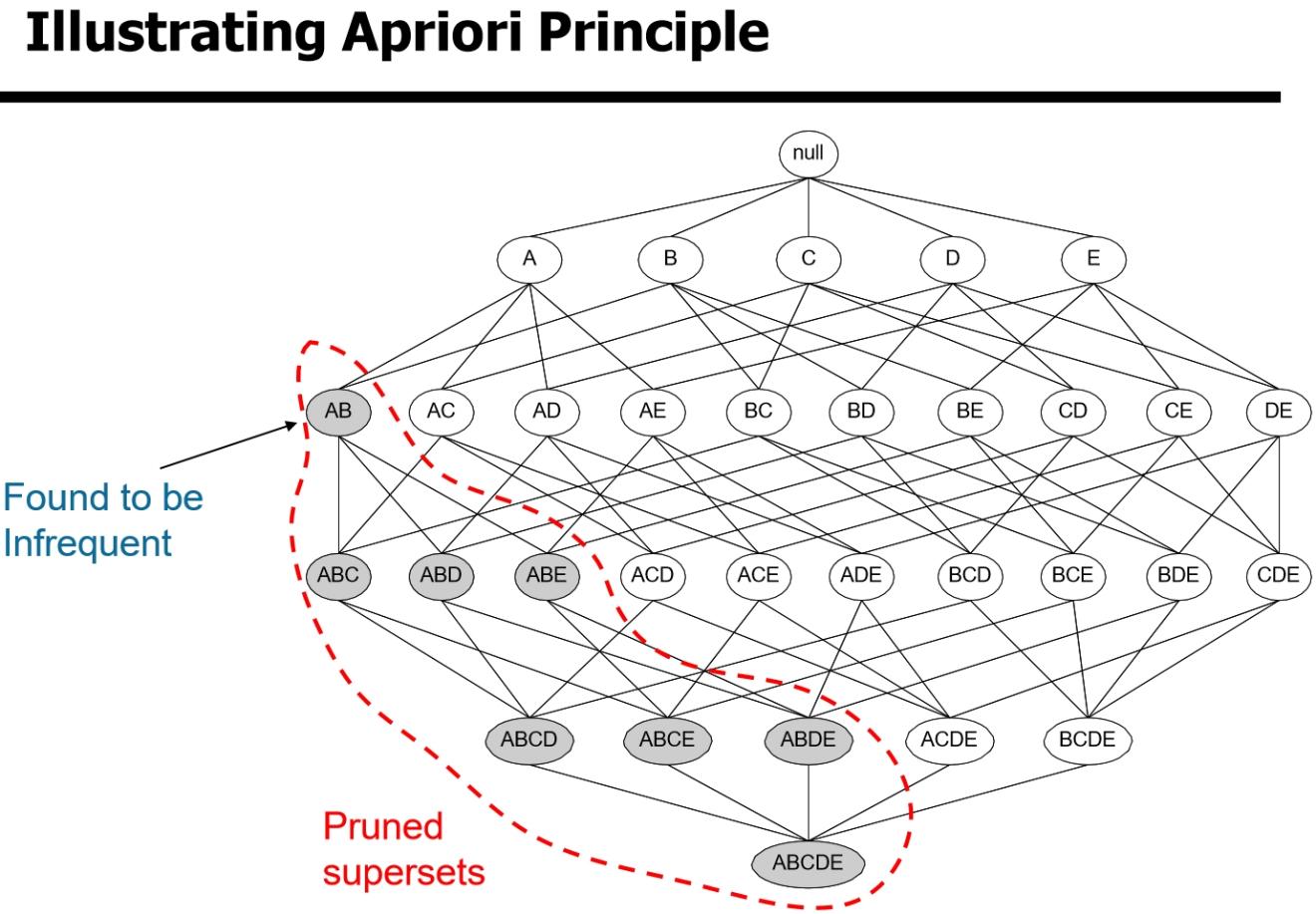
Given a set of transactions T, the goal of association rule mining is to find all rules having

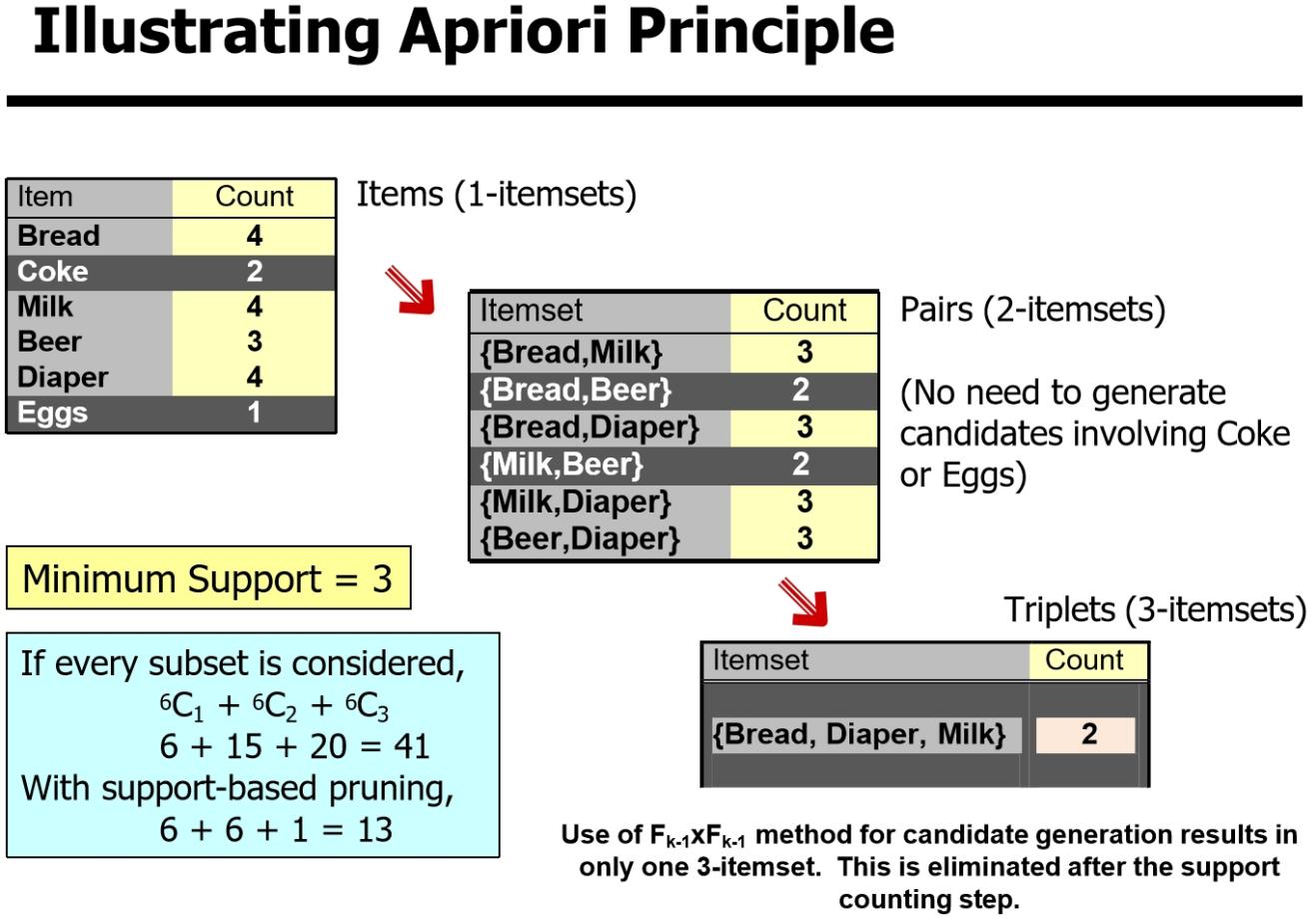
## Brute-force approach

List all possible association rules is Computationally prohibitive

## Apriori principle

If an itemset is frequent, then all of its subsets must also be frequent





## Rule Generation

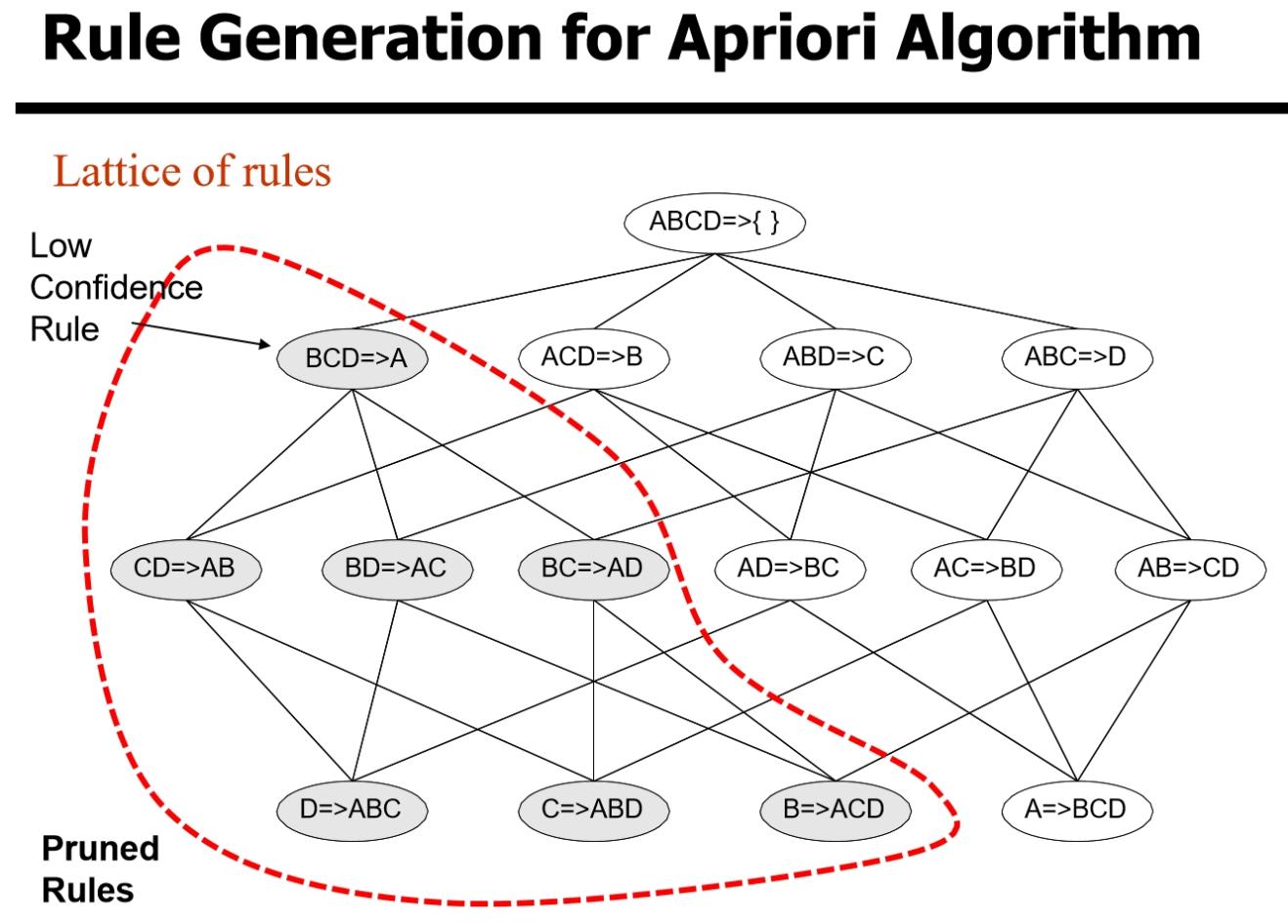
Given a frequent itemset L, find all non-empty subsets f⊂L such that f→L – f satisfies the minimum confidence requirement

If {A,B,C,D} is a frequent itemset, candidate rules:

|  |  |  |  |
| --- | --- | --- | --- |
| ABC D, | ABD C, | ACD B, | BCD A, |
| A BCD, | B ACD, | C ABD, | D ABC |
| AB CD, | AC  BD, | AD  BC, | BC AD, |
| BD AC, | CD AB, | L  ∅, | ∅L, |

If |L| = k, then there are 2k – 2 candidate association rules

(ignoring L→∅and ∅→L)



1. Week 6 Lecture 6

## Cluster Analysis

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Intra-cluster distances are minimized

Inter-cluster distances are maximized

## What is not Cluster Analysis

### Supervised classification

Have class label information

### Simple segmentation

Dividing students into different registration groups

alphabetically, by last name

### Results of a query

Groupings are a result of an external specification

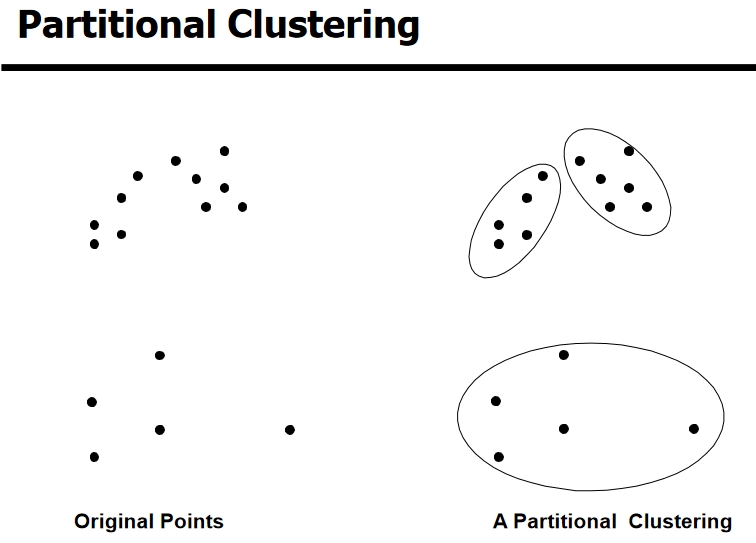
### Graph partitioning

Some mutual relevance and synergy, but areas are not identical

## Types of Clusterings

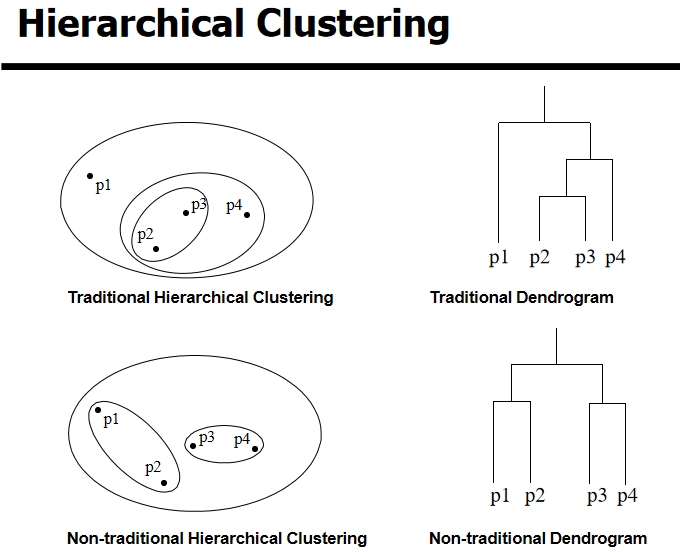
### Partitional Clustering

A division of data objects into non-overlapping subsets (clusters)

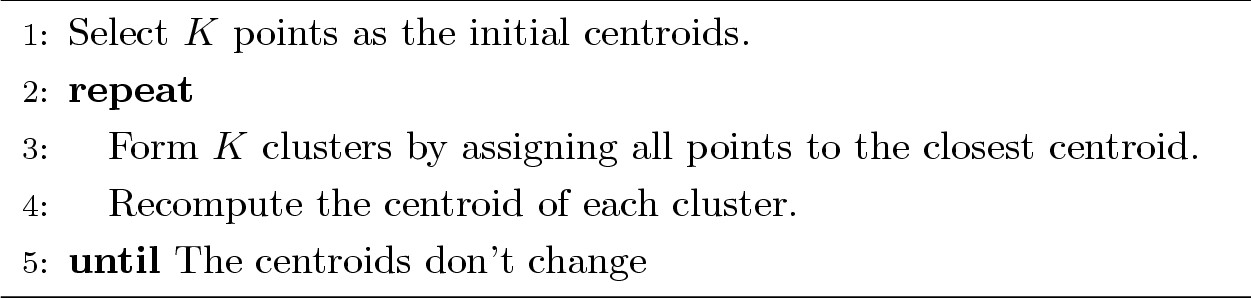


### Hierarchical Clustering

A set of nested clusters organized as a hierarchical tree



## K-means Clustering



Locking transaction has to be signed by k out of n pub keys

## means Objective Function

A common objective function (used with Euclidean distance measure) is Sum of Squared Error (SSE)



## Selecting Initial Points

### Multiple runs

Helps, but probability is not on your side

### Use some strategy to select the k initial centroids and then select among these initial centroids

Select most widely separated

Use hierarchical clustering to determine initial centroids

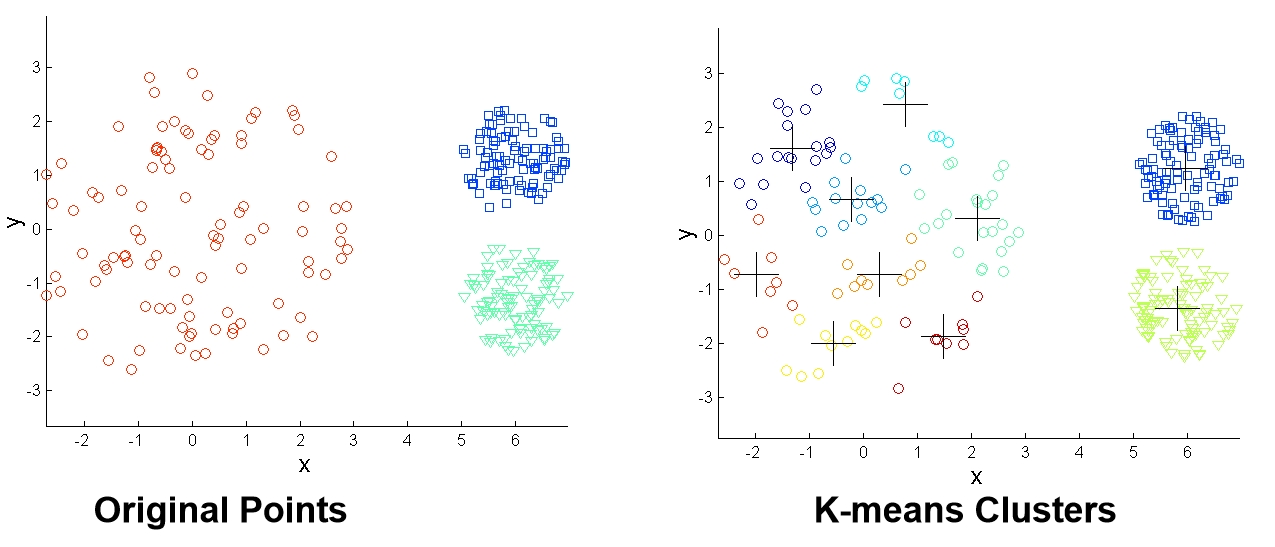
## Limitations of K-means

K-means has problems when clusters are of differing(Sizes, Densities, Non-globular shapes)

K-means has problems when the data contains outliers.

## Overcoming K-means Limitations

One solution is to find a large number of clusters such that each of them represents a part of a natural cluster. But these small clusters need to be put together in a post-processing step.



1. Week 7 Lecture 7

## Hierarchical Clustering

Produces a set of nested clusters organized as a hierarchical tree

Can be visualized as a dendrogram

## Strengths of Hierarchical Clustering

Do not have to assume any particular number of clusters

They may correspond to meaningful taxonomies

## Two main types of hierarchical clustering

### Agglomerative

Start with the points as individual clusters

At each step, merge the closest pair of clusters until only one cluster (or k clusters) left

### Divisive

Start with one, all-inclusive cluster

At each step, split a cluster until each cluster contains an individual

point (or there are k clusters)

## Basic algorithm

Compute the proximity matrix

Let each data point be a cluster

Repeat

Merge the two closest clusters

Update the proximity matrix

Until only a single cluster remains

## How to Define Inter-Cluster Distance

MIN

MAX

Group Average

Distance Between Centroids

## Comparison

|  |  |
| --- | --- |
| Strength | Limitation |
| 微信截图_20241024092140 | 微信截图_20241024092031 |
| 微信截图_20241024092158 | 微信截图_20241024092109 |

## Ward’s Method

Similarity of two clusters is based on the increase in squared error when two clusters are merged

Less susceptible to noise

Biased towards globular clusters

Hierarchical analogue of K-means

## Time and Space requirements

O(N2) space since it uses the proximity matrix.

O(N3) time in many cases

## Problems and Limitations

Once a decision is made to combine two clusters, it cannot be undone

No global objective function is directly minimized

Different schemes have problems with one or more of the following:

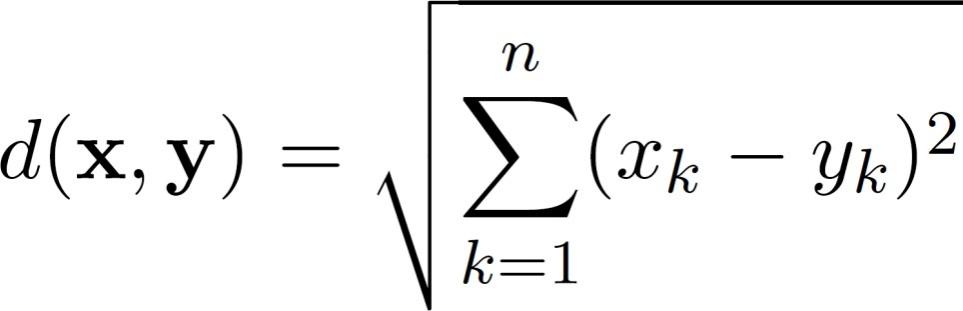
Sensitivity to noise

Breaking large clusters

Difficulty handling clusters of different sizes and non- globular shapes

## Euclidean Distance

Standardization is necessary, if scales differ.



1. Week 8 Lecture 8