Clustering analysis

Rhee, Je-Keun
Motivating Questions for Clustering

• What is the natural groupings in a set of data?
Motivating Questions for Clustering

Simpson's Family | School Employees | Females | Males

The way of grouping is not unique
Introduction to Weka

- **Weka: Data Mining Software in Java**
  - Weka is a collection of machine learning algorithms for data mining & machine learning tasks
  - What you can do with Weka?
    - data pre-processing, feature selection, **classification**, regression, **clustering**, association rules, and visualization
  - Weka is an open source software issued under the GNU General Public License
Dataset #1: Iris Flower

• Description
  – The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Aylmer Fisher (1936)
  – These were collected the data to quantify the geographic variation of Iris flowers in the Gaspé Peninsula

• Configuration of the data set
  – 4 attributes
    • Sepal length
    • Sepal width
    • Petal length
    • Petal width
Dataset #2: Handwritten Digits (MNIST)

• Description
  – The MNIST database of handwritten digits contains digits written by office workers and students
  – We will build a classifier and do clustering with the reduced set of MNIST

• Configuration of the data set
  – Attributes
    • pixel values in gray level in a 28x28 image
    • 784 attributes (all 0~255 integer)
  – Preprocessing for practice
    • Feature selection: 784 ➔ 400
  – Class value: 0~9, which represent digits from 0 to 9
Data format for Weka (.arff)

@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male }
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina }
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes }
@attribute class { present, not_present }

@data

63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present

Note: You can easily generate ‘arff’ file by adding a header to a usual CSV text file