Clustering

- Goal - Group data points into clusters having:
  - *Similar* records are grouped in the same cluster
  - *Non-similar* records are placed in different clusters
- Can be used for both:
  - Classification / Undirected automatic discovery
  - Visualization of high-dimensional data

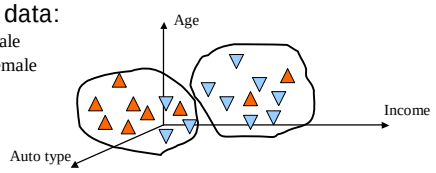
Clustering Techniques

- k-means
  - Number of clusters (k) is chosen
  - Members are iteratively moved between clusters
- Hierarchical (creates *dendrograms*)
  - Grouping by agglomeration
  - Grouping by division
- Self-organizing feature maps (SOFM)
  - No discrete clusters
  - Representation tries to preserve *proximity* information

Example (cont.)

- Low dimensionality so we could visualize the data:

▽- Male
△- Female



K-Means Approach

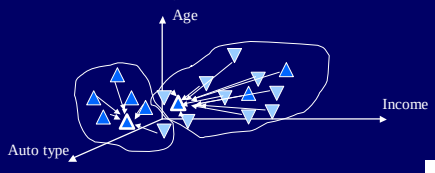
- One of the most popular clustering algorithms
- A *partitioning* method
- It has been around for a while
  - Has many variations
  - Has been given many different names (reinvented)
- Simple to understand and implement

K-Means Approach

- Start with *K* randomly generated *seed* clusters
  - These could be *K* randomly selected records
  - Best not to select very similar data points as seeds
- Iteratively update the *K* seeds to correspond to density centers

K-Means Approach

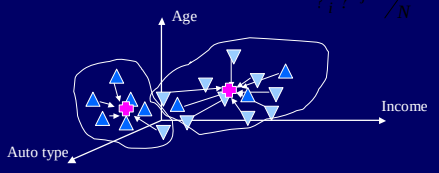
- Associate each of the records to one of the *K* cluster points (initial grouping) based on "similarity".



K-Means Approach (Cont.)

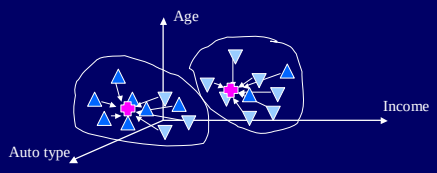
- Calculate the *mean* of each cluster

$$\frac{\sum_{i} x_{ij}}{N}$$



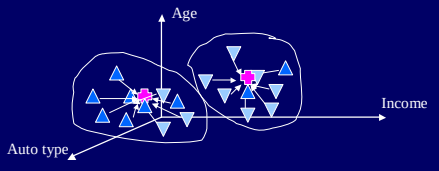
K-Means Approach (Cont.)

- Reassign each record to one of the *K* means



K-Means Approach (Cont.)

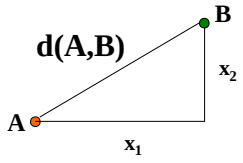
- Repeat the cycle until no changes
  - Recalculate new means
  - Recluster



K-Means Approach (Cont.)

- In addition to grouping, we have a cluster mean for each cluster
  - Average record for each cluster
- The dimensionality of the mean is the same as the data
  - Can t visualize the relationships between high-dimension clusters in low-dimension plot

How do we measure similarity (distance)?



Euclidean distance is the most basic measure.

$$d(A,B) = \sqrt{x_1^2 + x_2^2}$$

How to Measure Similarity

- Depends on the type of variable / attribute
  - Difficult to perform for categorical data
    - How similar is an apple to orange vs. an apple to a pear?
  - Can use heuristic functions for non-metric (non-true measures)
    - Use shape/size/weight of the fruit to establish a numeric feature
    - Use a predefined distance matrix