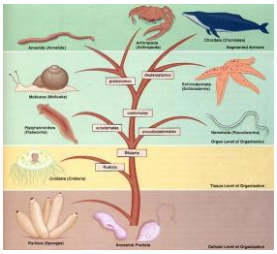
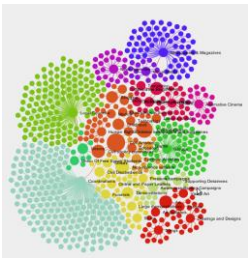




Generic Machine Learning

Clustering and Classification

TCTI-VKAAI-17: Applied Artificial Intelligence

1

K-Means Clustering Algorithm

Also called 'Centroids'

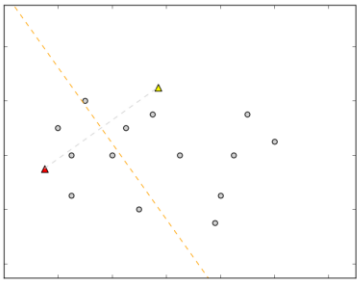
Amount of clusters

- Input: K, set of points $x_1, \ldots, x_n$
- Place centroids $c_1, \ldots, c_k$ at random locations
- Repeat until convergence: (Until nothing changes)
  - For each point x_i:
    - Find the nearest centroid c_j $\quad \arg\min_j D(x_i, c_j)$
    - Assign the point x_i to cluster j
  - For each cluster $j = 1, \ldots, K$:
    - Calculate new centroid c_j = mean of all points x_i assigned to cluster j $\quad c_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a)$ for $a = 1, \ldots, d$
- Stop when none of the cluster assignments changes

Distance (e.g., Euclidian) between instance x_i and c_j

15

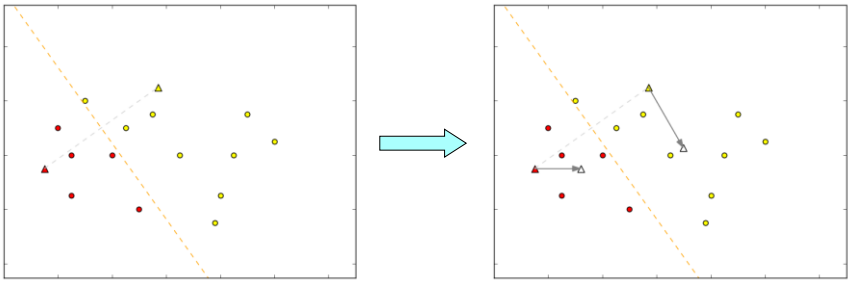
K-Means Example



- Input K, points $x_1, \ldots, x_n$
- Place centroids randomly
- Repeat until convergence:
  - For each point x_i:
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j:
    - Compute new centroid c_j

16

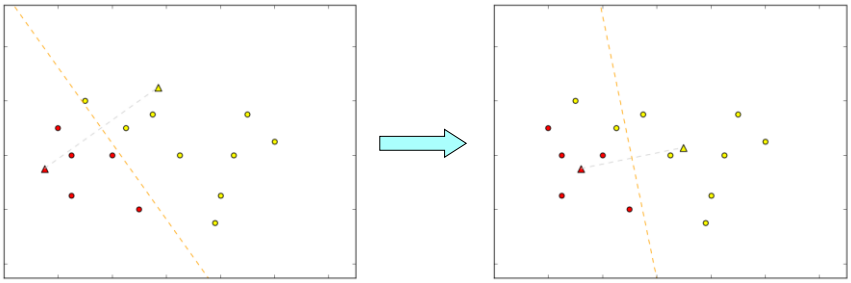
K-Means Example



- Input K, points $x_1, \ldots, x_n$
- Place centroids randomly
- Repeat until convergence:
  - For each point x_i:
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j:
    - Compute new centroid c_j

17

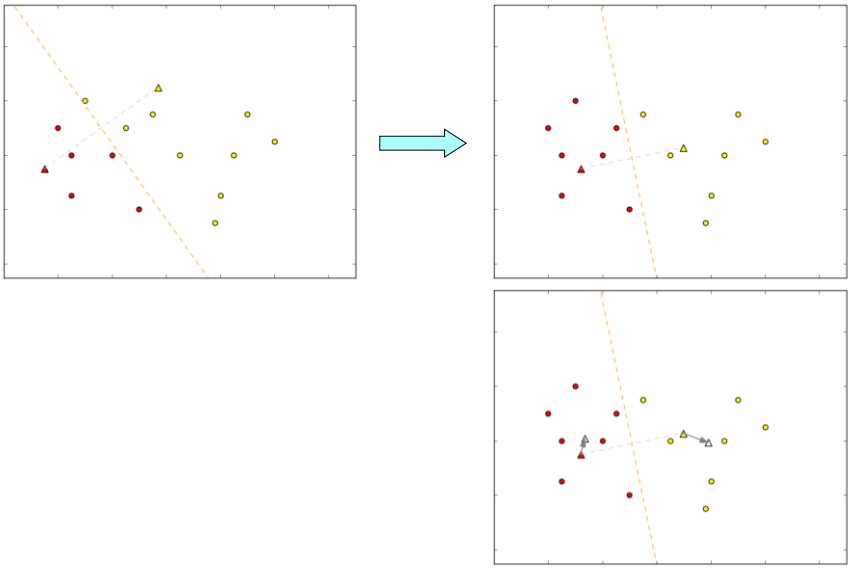
K-Means Example



- Input K, points $x_1, \ldots, x_n$
- Place centroids randomly
- Repeat until convergence:
  - For each point x_i:
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j:
    - Compute new centroid c_j

18

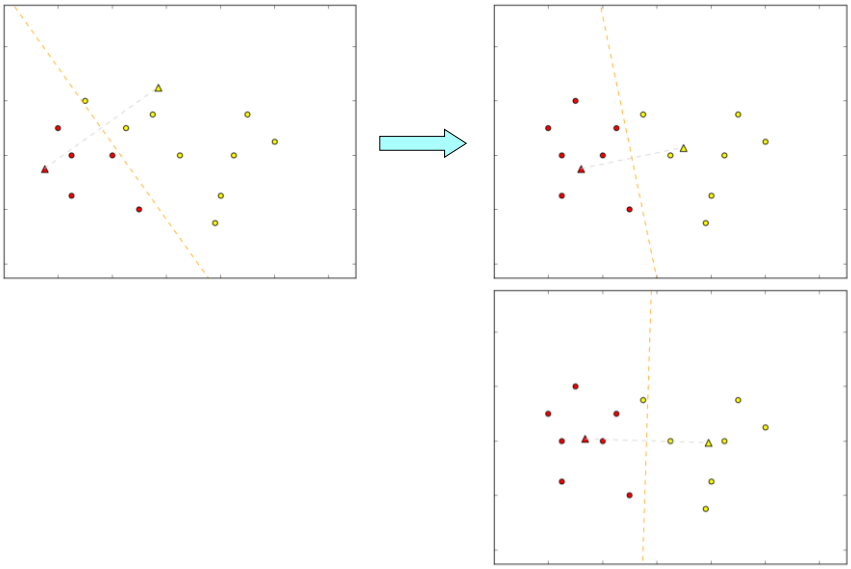
K-Means Example



- Input K, points $x_1, \ldots, x_n$
- Place centroids randomly
- Repeat until convergence:
  - For each point x_i:
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j:
    - Compute new centroid c_j

19

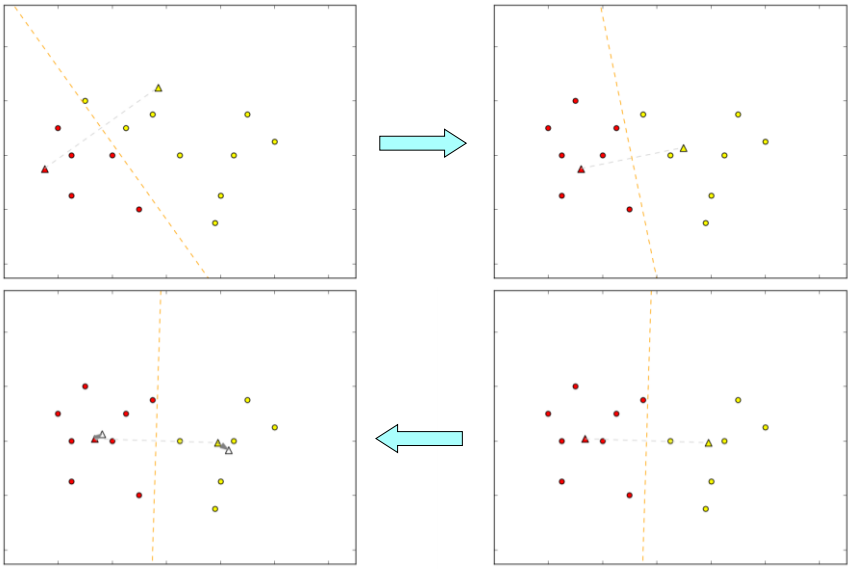
K-Means Example



- Input K, points $x_1, \ldots, x_n$
- Place centroids randomly
- Repeat until convergence:
  - For each point x_i:
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j:
    - Compute new centroid c_j

20

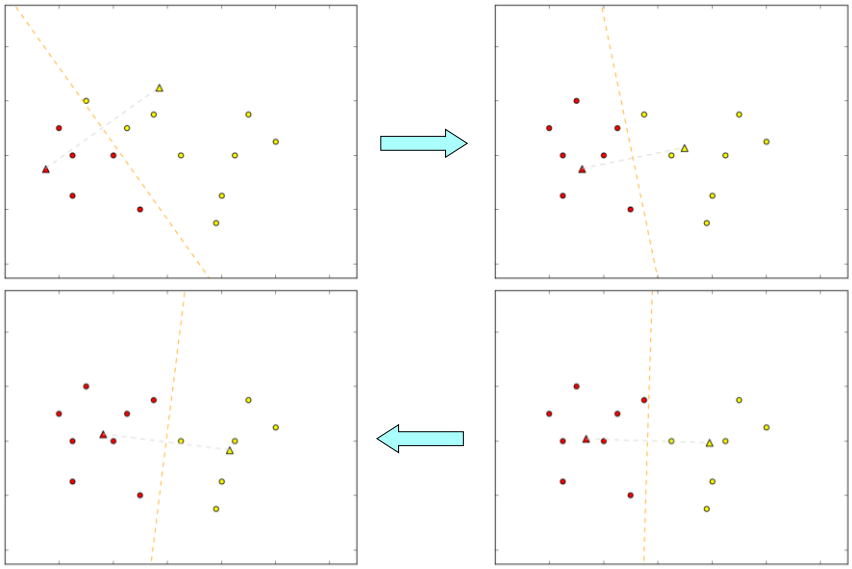
K-Means Example



- Input K, points $x_1, \ldots, x_n$
- Place centroids randomly
- Repeat until convergence:
  - For each point x_i:
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j:
    - Compute new centroid c_j

21

K-Means Example



- Input K, points $x_1, \ldots, x_n$
- Place centroids randomly
- Repeat until convergence:
  - For each point x_i:
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j:
    - Compute new centroid c_j

22

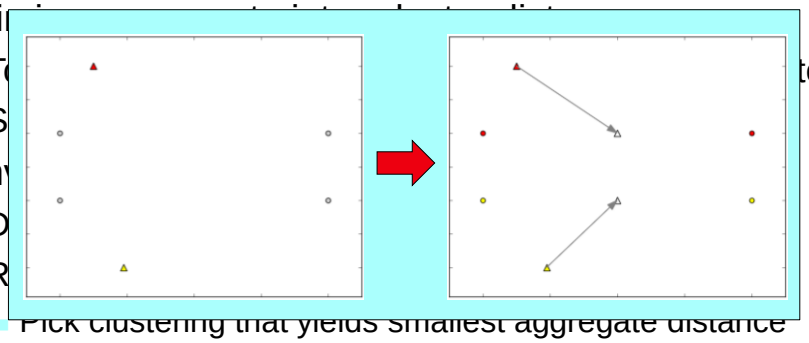
K-Means Properties

- Minimizes aggregate intra-cluster distance
  - Total squared distance from point to centre of its cluster
  - Same as variance if Euclidian distance is used
- Converges to local minimum
  - Different starting points → very different results
  - Run several times with random starting points
    - Pick clustering that yields smallest aggregate distance
- Nearby points might not end up in the same cluster
  - The following clustering is a stable local minimum

23

K-Means Properties

- Mini
  - T
  - S
- Con
  - D
  - R
    - Pick clustering that yields smallest aggregate distance
- Nearby points might not end up in the same cluster
  - The following clustering is a stable local minimum

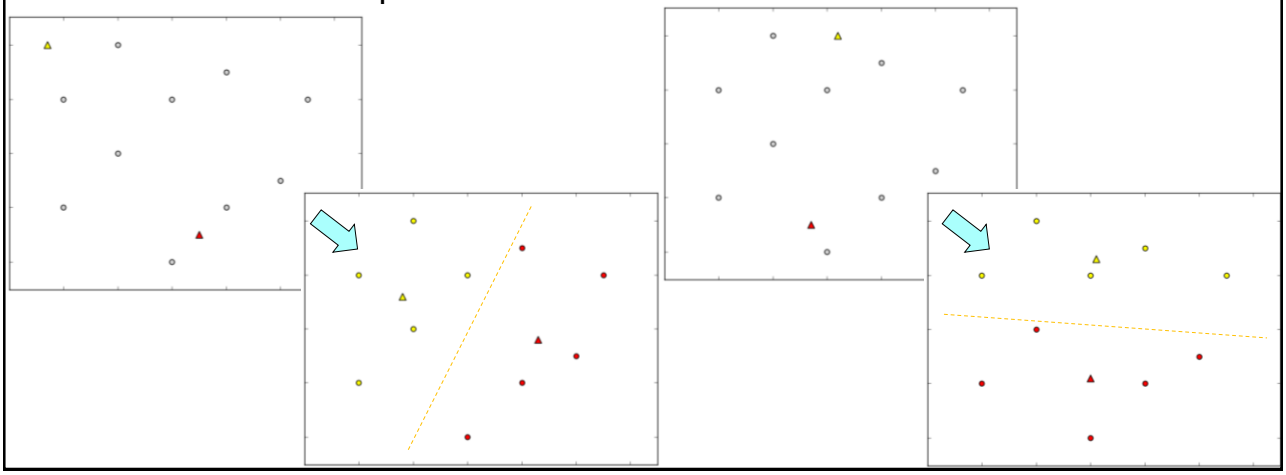


ter

24

Convergence

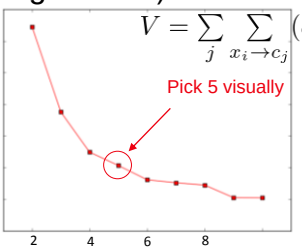
- K-Means converges to a ***local minimum***
  - Solution depends on initial values of K's

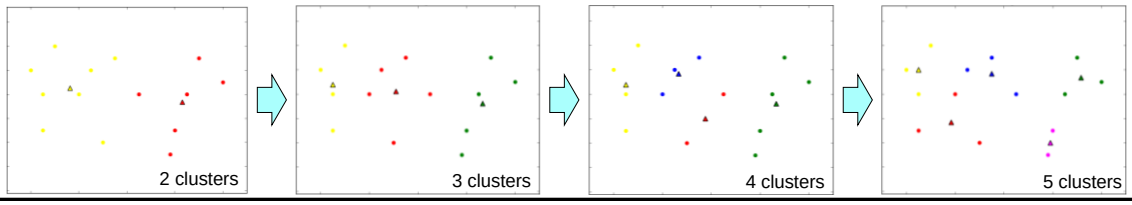


How many clusters?

- How many clusters are there in the data?
  - Class labels may suggest the value of K (e.g., digits 0…9)
  - Optimize distance V : for $K = 2, 3, \ldots$
    - Run K-Means, record distance
    - Problem: V minimized when $K = n$
      - What if we use a validation set?
    - Visually from scree plot
      - Point where 'mountain' ends, 'rubble' begins

$$V = \sum_{j} \sum_{x_i \to c_j} (c_j - x_i)^2$$





26

Conclusion

- Classification and Clustering appear similar
  - Classification: supervised categorization of observations
  - Clustering: unsupervised determination of (potential) categories
- k Nearest Neighbors simple and effective
  - But very costly to run effectively
- K-Means also rather simple and effective (and fast!)
  - Some unwanted results due to randomness of chosen centroids
  - Meaning one has to calculate several times for a data set → more costly

27