Foundations of AI

10. Machine Learning Revisted

Unsupervised Learning

Wolfram Burgard, Andreas Karwath, Bernhard Nebel, and Martin Riedmiller

Clustering (1)

- Common technique for statistical data analysis (machine learning, data mining, pattern recognition, …)
- Classification of a data set into subsets (clusters)
- Ideally, data in each subset have a similar characteristics (proximity according to distance function)

Clustering (2)

- Needed: distance (similarity / dissimilarity) function, e.g., Euclidian distance
- Clustering quality
  - Inter-clusters distance maximized
  - Intra-clusters distance minimized
- The quality depends on
  - Clustering algorithm
  - Distance function
  - The application (data)

Types of Clustering

- Hierarchical Clustering
  - Agglomerative Clustering (buttom up)
  - Divisive Clustering (top-down)

- Partitional Clustering
  - K-Means Clustering (hard & soft)
  - Gaussian Mixture Models (EM-based)

K-Means Clustering

- Partitions the data into *k* clusters (k is to be specified by the user)
- Find *k* reference vectors $\boldsymbol{m}_j$, $j = 1,\ldots,k$ which best explain the data $\boldsymbol{X}$
- Assign data vectors to nearest (most similar) reference $\boldsymbol{m}_i$

$$\left\|\mathbf{x}^t - \mathbf{m}_i\right\| = \min_j \left\|\mathbf{x}^t - \mathbf{m}_j\right\|$$

r-dimensional data vector in a real-valued space

reference vector (center of cluster = mean)

Reconstruction Error
(K-Means as Compression Alg.)

- The total reconstruction error is defined as

$$E\left(\{\mathbf{m}_i\}_{i=1}^{k} \middle| \mathsf{X}\right) = \sum_t \sum_i b_i^t \left\|\mathbf{x}^t - \mathbf{m}_i\right\|^2$$

with

$$b_i^t = \begin{cases} 1 & \text{if } \left\|\mathbf{x}^t - \mathbf{m}_i\right\| = \min_j \left\|\mathbf{x}^t - \mathbf{m}_j\right\| \\ 0 & \text{otherwise} \end{cases}$$

- Find reference vectors which minimize the error
- Taking its derivative with respect to m_i and setting it to 0 leads to

$$\mathbf{m}_i = \frac{\sum_t b_i^t \mathbf{x}^t}{\sum_t b_i^t}$$

K-Means Algorithm

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$

Repeat

For all $\boldsymbol{x}^t \in \mathcal{X}$

$$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

For all $\boldsymbol{m}_i, i = 1, \ldots, k$

$$\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$$

Until $\boldsymbol{m}_i$ converge

Recompute the cluster centers $\boldsymbol{m}_i$ using current cluster membership

Assign each $\boldsymbol{x}^t$ to the closest cluster

K-Means Example

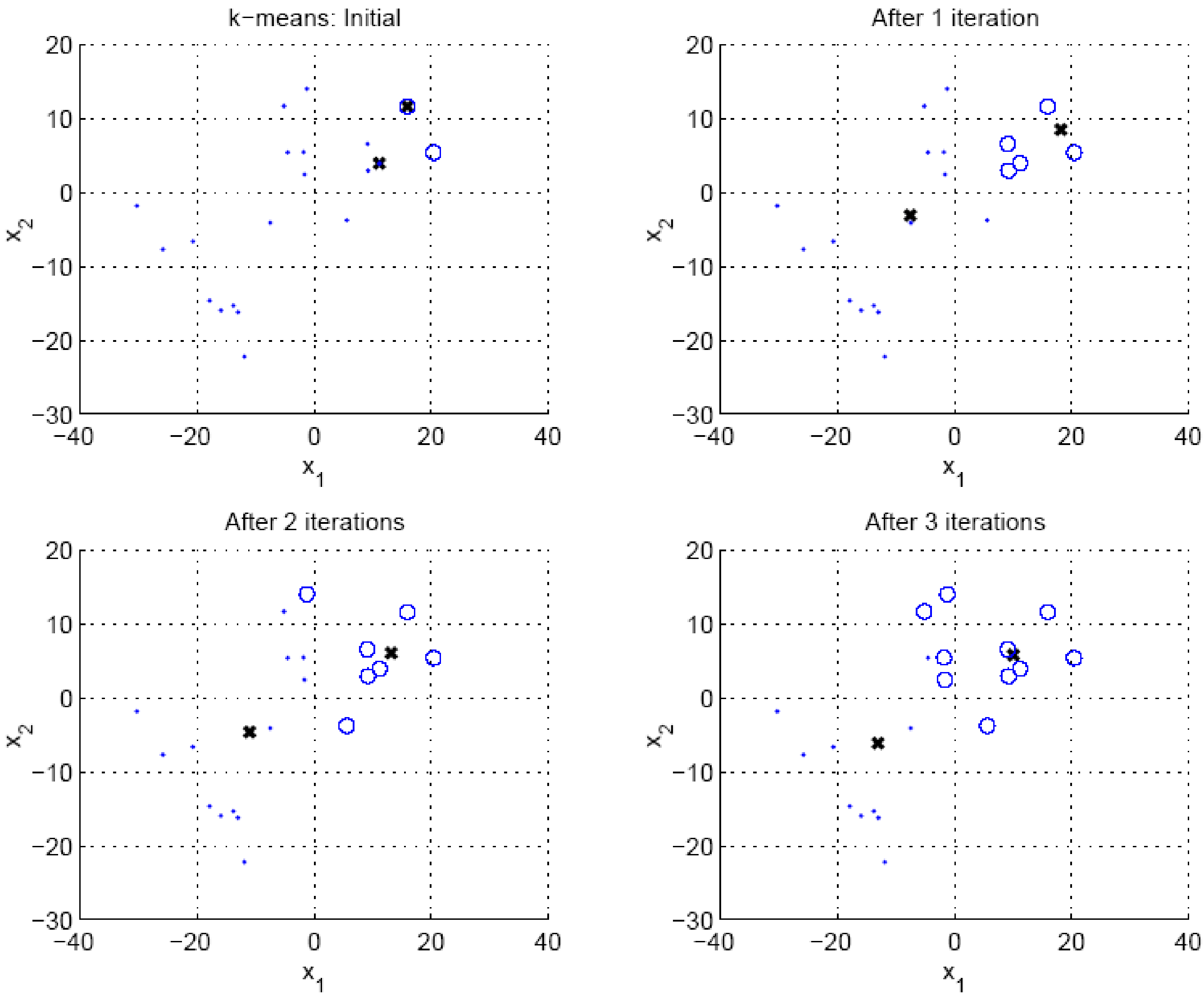


Image source: Alpaydin, Introduction to Machine Learning

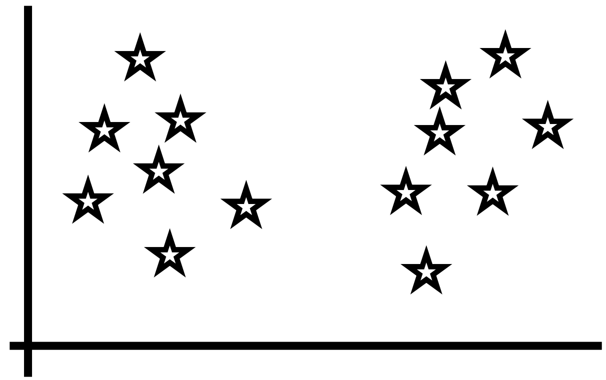
Strength of K-Means

- Easy to understand and to implement
- Efficient O(nkt)
  n = #iterations, k = #clusters, t = #data points
- Converges to a local optimum (global optimum is hard to find)
- Most popular clustering algorithm

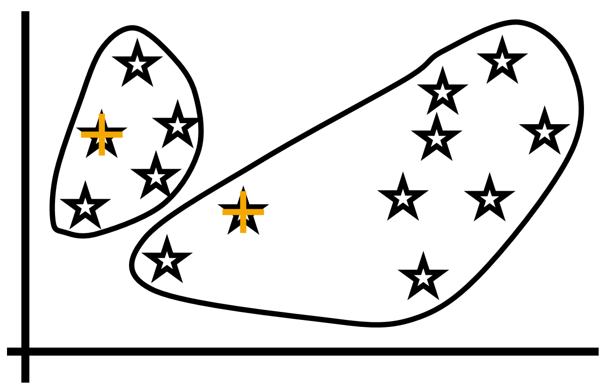
Weaknesses of K-Means

- User needs to specify #clusters (k)
- Sensitive to initialization (strategy: use different seeds)
- Sensitive to outliers since all data points contribute equally to the mean (strategy: try to eliminate outliers)

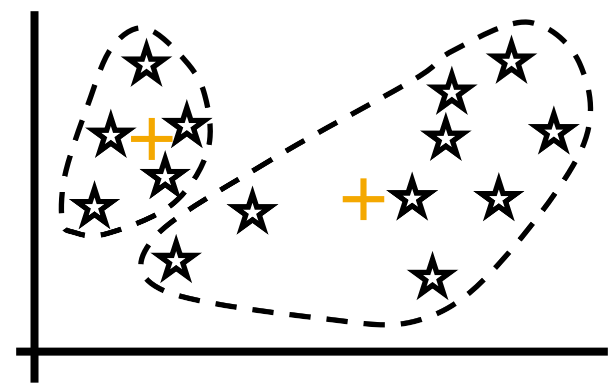
An example



(A). Random selection of *k* centers

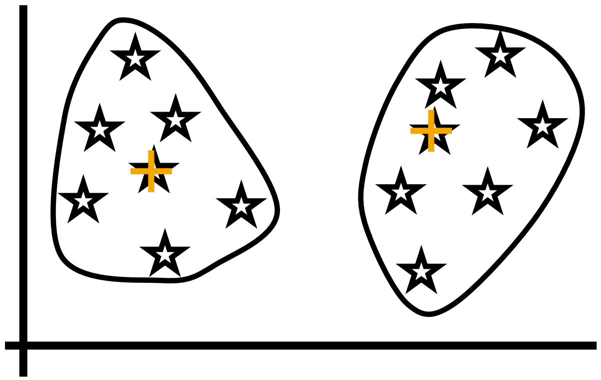


***Iteration* 1: (B). Cluster assignment**

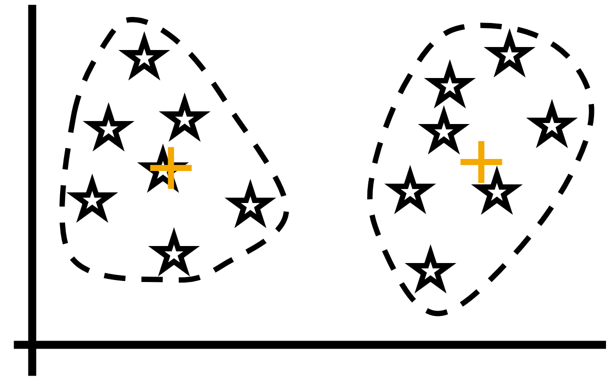


(C). Re-compute centroids

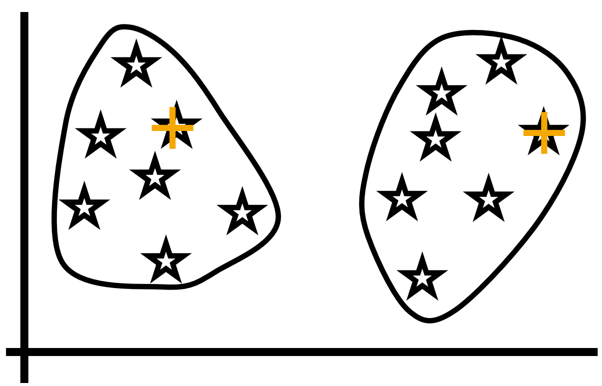
An example (cont ...)



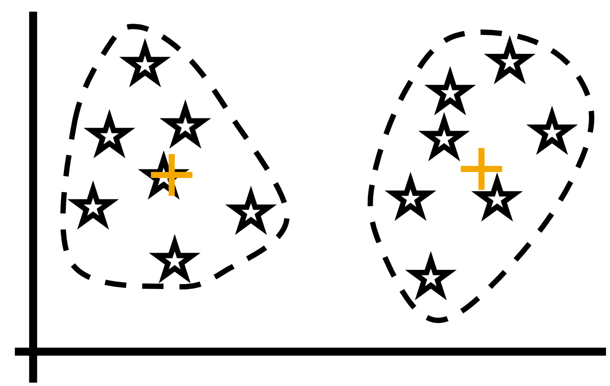
Iteration 2: (D). Cluster assignment



(E). Re-Compute centeroids

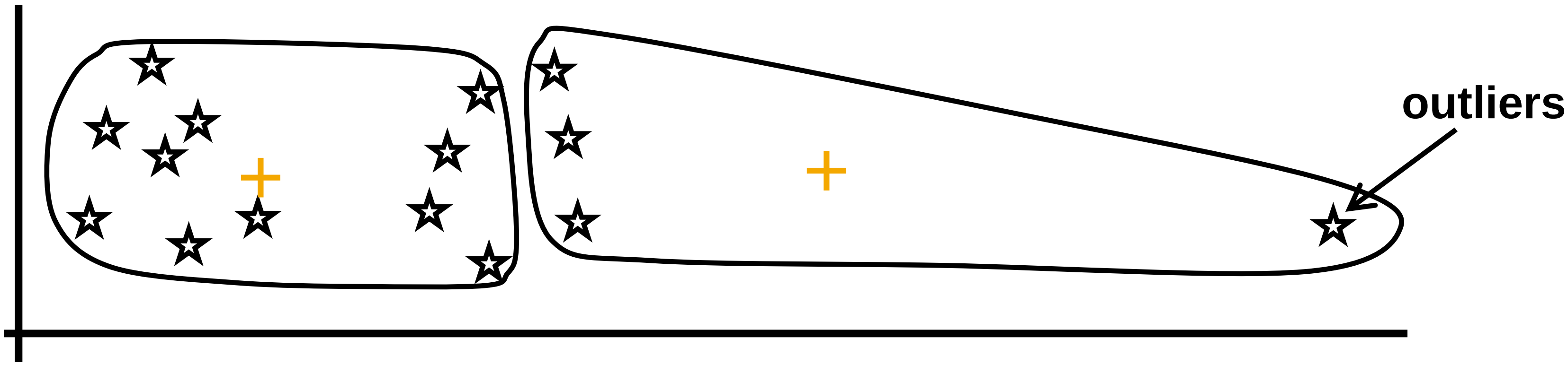


Iteration 3: (F). Cluster assignment

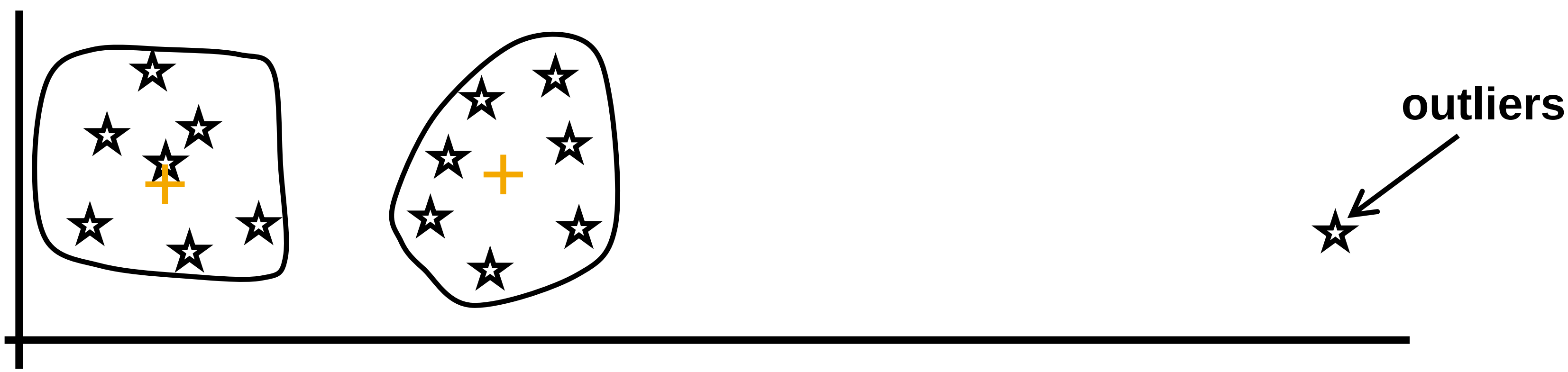


(G). Re-Compute centeroids

Weaknesses of k-means: Problems with outliers



(A): Undesirable clusters



(B): Ideal clusters

Soft Assignments

- So far, each data point was assigned to exactly one cluster
- A variant called soft k-means allows for making fuzzy assignments
- Data points are assigned to clusters with certain probabilities

Soft K-Means Clustering

- Each data point is given a soft assignment to all means

$$c_{tk} = \frac{\exp(-\beta\,||x^t - m_k||^2)}{\sum_i \exp(-\beta\,||x^t - m_i||^2)}, \quad \sum_k c_{tk} = 1$$

- β is a "stiffness" parameter and plays a crucial role
- Means are updated $m_k = \frac{\sum_t c_{tk} x^t}{\sum_t c_{tk}}$

- Repeat assignment and update step until assignments do not change anymore

Soft K-Means Clustering

- Points between clusters get assigned to both of them
- Points near the cluster boundaries play a partial role in several clusters
- Additional parameter β
- Clusters with varying shapes can be treated in a probabilistic framework (mixtures of Gaussians)

After Clustering

- Allows knowledge extraction through

  number of clusters (if adaptive),

  cluster parameters, i.e., center, range of features.

Clustering as Preprocessing

- Estimated group labels h_j (soft) or b_j (hard) may be seen as the dimensions of a new k dimensional space, where we can then learn our discriminant or regressor.
- Local representation (only one b_j is 1, all others are 0; only few h_j are nonzero) vs distributed representation

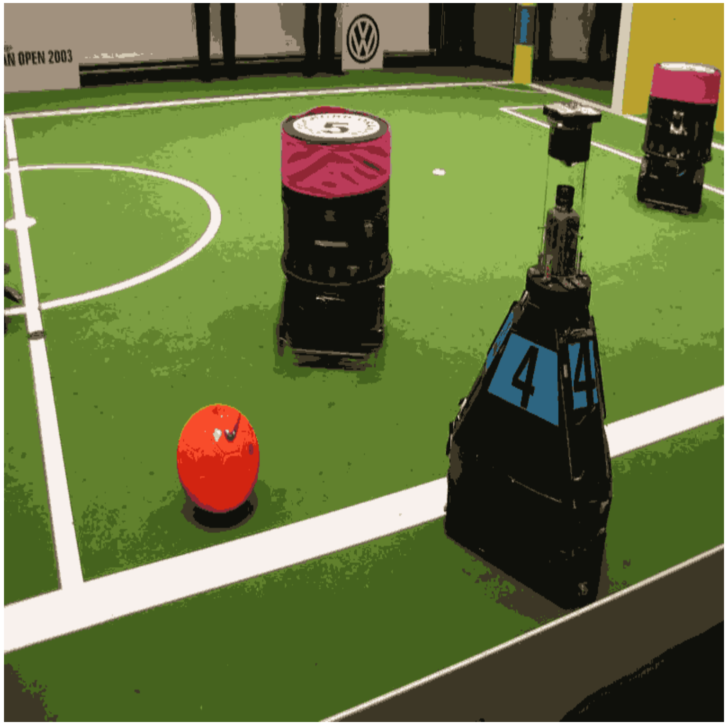
Examples of Clustering

Original

16 Colors

Examples of Clustering

Original



16 Colors

Summary

- K-Means is the most popular clustering algorithm
- It is efficient and easy to implement
- Converges to a local optimum
- A variant of hard k-means exists allowing soft assignments
- Soft k-means corresponds to the EM algorithm which is a general optimization procedure