Unsupervised Machine Learning: Clustering: k-means

Alipio Jorge

January 2021

Supervised vs. Unsupervised

- Instructor avaialable
  - **supervised**
- Instructor not available
  - **unsupervised**
- Instructor partially available
  - **semi-supervised**
- Instructor better than random but not very precise
  - **weakly supervised**
- Finding a representation using a surrogate task
  - **self-supervised**

Clustering

Cluster analysis

Partition a set of objects into groups

- Each group is a **cluster**
- Objects in the same cluster are **similar**
- Objecrs in different clusters are **dissimilar**

In the real world differences are not always easy to find

- Maximise intra-similarity
- Minimize inter-similarity

Why Clustering?

No labels available

- **Learn classes** without a supervisor
- Examples have **no class labels**

Applications

- Customer segmentation
- Divide patients in homogeneous groups
- Organize web results by content

Figure 1: ' '

Cluster analysis

- **Given**
  - a set of objects
  - a number k of desired groups
- **Obtain**
  - a mapping of each object into each group

Example

- $Objects = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $k = 3$
- $Groups = \{1, 1, 2, 1, 3, 2\}$

k-means clustering

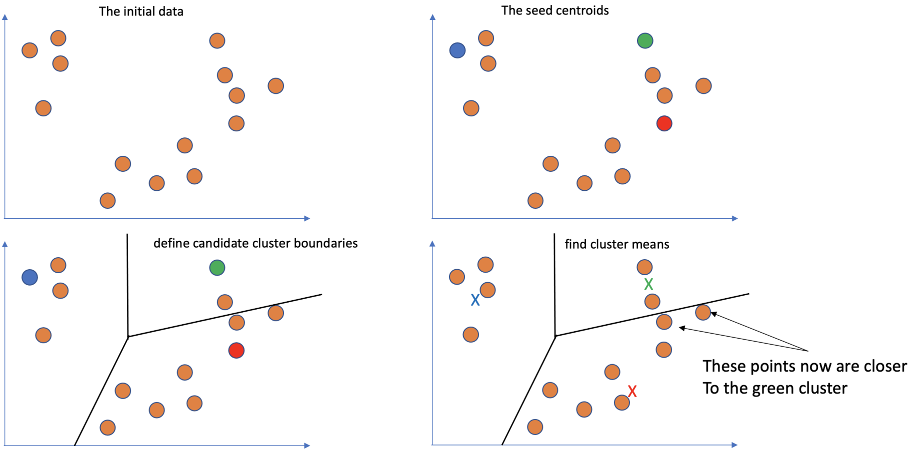
Setup

- Imagine the data points in a **multidimensional space**
  - The dimensions are the attributes
- Pick an appropriate **distance/similarity metric**
  - It should correspond to our **intuition** of the domain

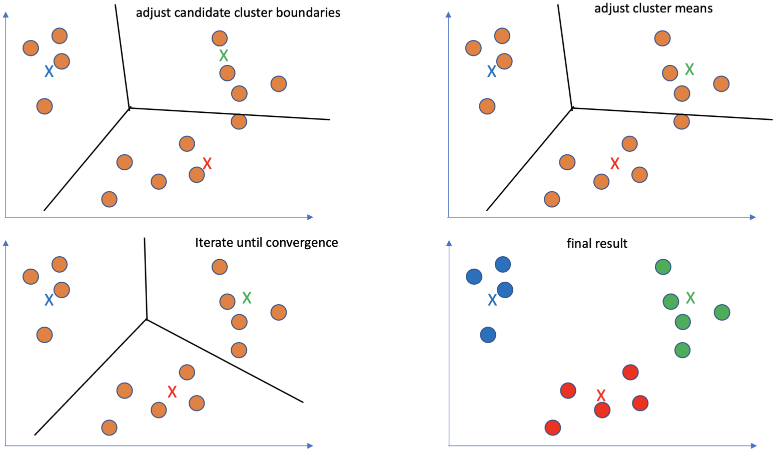
Strategy: partitioning method

- Obtain an **initial partitioning**
- Improve **iteratively**

k-means clustering



k-means clustering



k-means clustering

The algorithm

- **Input**
  - k: the number of clusters,
  - D: a data set containing n objects.
- **Output**
  - A set of k clusters.
- **Method**
  1. **arbitrarily** choose k objects from D as the initial cluster centers
  2. **repeat**
     - **(re)assign** each object to the cluster to which the object is the most similar the cluster center
     - **update** the cluster means, that is, calculate the mean value of the objects for each cluster
  3. **until** no change

k-means algorithm

Result of k-means

- k **disjoint** clusters $C_1, C_2, \ldots, C_k$
- $\bigcup_{i=1}^{k} C_i = D$, every point is in one cluster
- Each cluster C_i is characterized by a **centroid** $\mathbf{c}_i$
- A centroid is a vector but typically **not a true point**

$$\mathbf{c}_i = \mathrm{average}_j \mathbf{x}_j \ , \mathbf{x}_j \in C_i$$

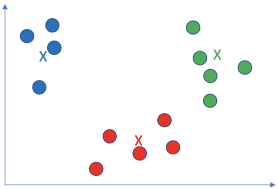


Figure 2: ' '

k-means clustering

Quality of a cluster

- we want to minimize **within-cluster variation**
- a measure of **error** (an objective function)
- the sum of the squared distances to the **centroid**

$$E = \sum_{i} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

The clustering problem

The complexity

- In general it is a **NP-hard problem** (https://mathworld.wolfram.com/NP-HardProblem.html)
- k-means is a **greedy approach**
- **Complexity of k-means** is $O(nkt)$
  - $t = iterations$
  - usually dominated by n (**in practice** $O(n)$)
  - very **efficient**

Convergence

- k-means **may not converge** to a global optimum (for a given k)
- results **depend** on the initial seed centers

Options in k-means

Initialization

- Random
- Heuristic
- User's choice

Calculating centroids

- Means
- Modes (for categorical values), a.k.a. k-modes
- Sample for scalability

Outliers

- Means can be affected by outliers
- **k-Medoids** is an alternative that uses **median**
  - and **absolute error** in the objective function

Evaluating the result of clustering

How can the results of clustering be **evaluated**?

- Is there a **cluster structure** in the data?
- Is the **number of clusters** adequate?
- **How good** are the clusters?

Preparing for clustering

Cluster structure

- Non uniform data
- Use **Hopkins statistic** to determine spatial randomness
  - $X \leftarrow$ sample m points from D
    - dx_i is the distance of each x_i to nearest neighbor in D
  - $Y \leftarrow$ generate m points uniformly
    - dy_i is the distance of each y_i to nearest neighbor in D
  - if H is close to 0.5 then D is not clusterable
    - $H > 0.5$ means good for clustering (some say $H > 0.75$)

$$H = \frac{\sum dy_i}{\sum dy_i + \sum dx_i}$$

Preparing for clustering

How many clusters?

- in general **not obvious**
- **elbow method**
  - **try** different values for k starting with 1 or around a reasonable number
  - **measure** within-cluster variance (or another quality measure)
    - it may be advisable to **average**
  - **plot** the curve for those values
  - **visually choose** the **turning** point of the curve

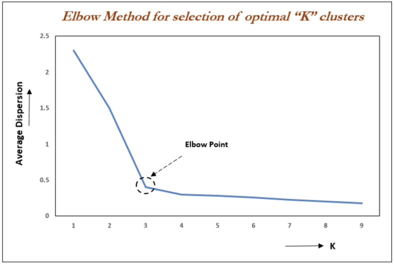


Figure 3: from "Statistics for Machine Learning" by Pratap Dangeti

Cluster Quality

Extrinsic Methods

- **Ground truth** is available
  - e.g., some cases, are labeled by experts
- **Completeness**: two cases with same label must be in same cluster
  - similar to **Recall**
- **Homogeneity**: all cases in one cluster should have same label
  - similar to ‘**Precision**
- Completeness and Homogeneity should be **balanced** (as in **F1**)
  - 1 cluster vs. n clusters
- e.g. **BCubed** recall and precision

Cluster Quality

Intrinsic Methods

- **NO ground truth**
  - typical scenario
- In general:
  - **compactness**
  - **separation**
- e.g. **silhouette coefficient**

Cluster Quality

Silhouette coefficient

- Is a measure and a **visualization** of cluster quality
- It helps to identify:
  - **compact** clusters
  - **well separated** clusters

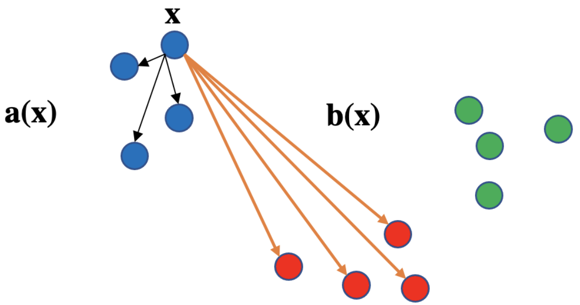
Silhouette Coefficient of one point $\mathbf{x}$

- $s(\mathbf{x})$ tends to 1 if the point is close to other points in same cluster **AND** very far from points in other clusters

Silhouette coefficient

How to calculate for a point $\mathbf{x}$

- calculate **average distances** of the point to each cluster
- $a(\mathbf{x})$ the distance within cluster
- $b(\mathbf{x})$ the distance to the nearest cluster
- $s(\mathbf{x}) = (b - a)/\max(b, a)$
- $-1 \leq s(\mathbf{x}) \leq 1$



Silhouette coefficient

visualization

- Plot **bars** for every point by cluster
- negative values stand out

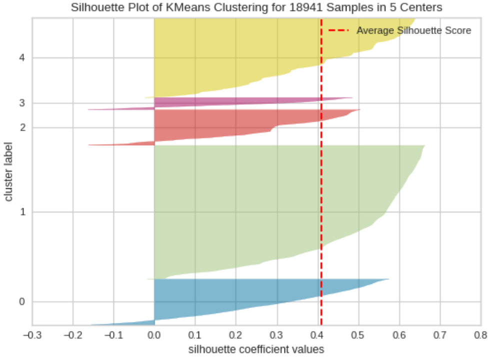
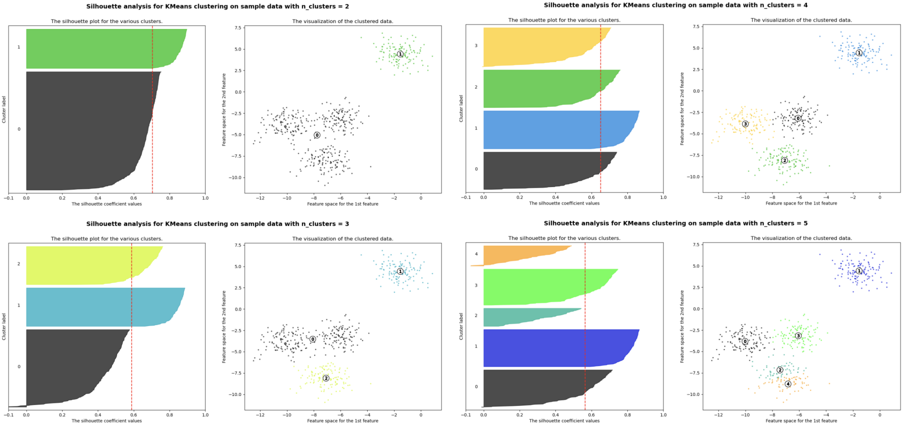


Figure 5: 5 cluster example from Yellowbrick

Silhouette 4 blobs Example

- From sklearn documentation "plot_kmeans_silhouette_analysis.html"



Other methods

Other than k-means

- **Hierarchical Clustering**
- **Density Based**
- etc.

References

- Books
  - Han, Kamber & Pei, Data Mining Concepts and Techniques, Morgan Kaufman.
- Scikit docs
  - https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html