Unsupervised Learning: Clustering 1

Ninghao Liu

University of Georgia

February 13, 2024

Some contents adopted from "Data Mining", Chapter 10, by Jiawei Han et al.

Overview

1. Basic Concepts
2. Partitioning methods
   - k-Means Clustering
   - Analysis of k-means Clustering

1. Basic Concepts

An application scenario:

- A company wants to devise targeted marketing strategies to enhance customer engagement and increase sales.
- Their customer base is diverse, encompassing various demographics, purchasing behaviors, and preferences.
- Applying a one-size-fits-all marketing strategy could lead to inefficient use of resources and missed opportunities to connect with specific customer segments.
- We may want to divide all customers into groups, and develop customized strategy for each group due to homophily.

Q: What kind of data mining techniques can help you to accomplish this task?

1. Basic Concepts

Classification: Training a model (parameterized by w) to fit the ground-truth label y given a data instance x.

$$\frac{1}{|\mathcal{D}_{train}|} \sum_{(x,y)\in\mathcal{D}_{train}} L(x, y, w) \tag{1}$$

Clustering: y is **unknown** in training.

- We need to *discover* these groupings.
- Useful when it is very costly or even infeasible to manually label the data.

1. Basic Concepts

Some high level ideas of **clustering**:

- It is the process of grouping a set of data objects into multiple subsets (i.e., **clusters**).
- Objects (i.e., instances) within a cluster have high similarity.
- Objects in a cluster are very dissimilar to those in other clusters.
- The definition of "dissimilarity" depends on which **distance metric** is used.

1. Basic Concepts

Types of clustering algorithms:

- Partitioning methods*.
- Hierarchical methods*.
- Density-based methods.
- Grid-based methods.

2. Partitioning methods

Definition: Given a data set $\mathcal{D}$ of N objects, and K, the number of clusters to form, clustering algorithms organize the objects into K partitions where $K \ll N$.

2.1 k-Means Clustering

Let the set of objects be $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$, k-means clustering aims to partition the N objects into K sets $S = \{S_1, S_2, ..., S_K\}$.

- Each cluster S_k has an centroid $\boldsymbol{\mu_k}$.
- Let $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denote the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

- Objects within a cluster are similar to one another.
- Objects in different clusters are very dissimilar.

2.1 k-Means Clustering

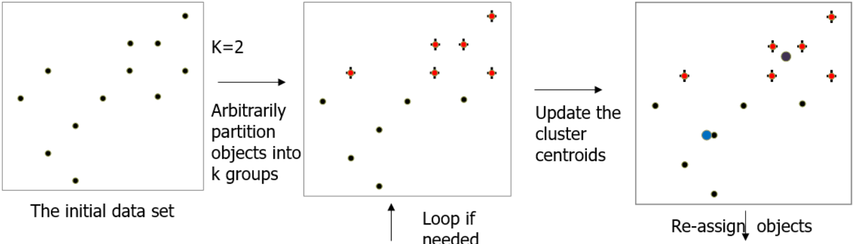
Algorithm:

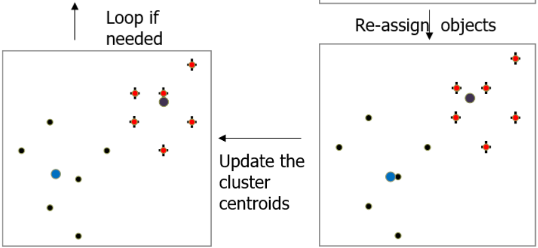
- Randomly assign objects to a cluster in $\{S_1, S_2, ..., S_K\}$.
- Initiate the K centroid vectors $\{\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, ..., \boldsymbol{\mu_K}\}$:

$$\boldsymbol{\mu}_k = \frac{1}{|S_k^{(t)}|} \sum_{\boldsymbol{x}_p \in S_k^{(t)}} \boldsymbol{x}_p. \tag{2}$$

- $t = 1$
- Alternate between the two steps.
  1. **Assignment:** Assign each objects to the cluster with the nearest centroid (i.e., $d(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$ is minimized).
  2. **Update:** Recalculate centroid vectors: $\boldsymbol{\mu}_k = \frac{1}{|S_k^{(t)}|} \sum_{\boldsymbol{x}_p \in S_k^{(t)}} \boldsymbol{x}_p$.
  3. $t = t + 1$.
- The algorithm converges when assignments no longer change.

2.1 *k*-Means Clustering





- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

2.2 *k*-Means Clustering Revisited

Let the set of objects be $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$, k-means clustering aims to partition the N observations into k sets $S = \{S_1, S_2, ..., S_K\}$.

- Each cluster S_k has an centroid $\boldsymbol{\mu_k}$.
- Let $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denote the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.
- Objects within a cluster are similar to one another.
- Objects in different clusters are very dissimilar.

But we haven't talked about any object-object distances.

- All of the distances involved in k-means are object-centroid ones.

2.2 k-Means Clustering Revisited

Equivalence between the two objectives:

- Given a set of observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$, k-means clustering divides the N observations into K sets $S = \{S_1, S_2, ..., S_K\}$.
- The objective in k-means is to find:

$$\underset{S}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in S_k} \|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2^2, \tag{3}$$

  where $\boldsymbol{\mu_k}$ is the mean of S_k.
- This is equivalent to:

$$\underset{S}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{2|S_k|} \sum_{\boldsymbol{x},\boldsymbol{y} \in S_k} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{4}$$

  i.e., minimizing the pairwise squared deviations of points in the same cluster.

k-means Clustering Revisited

Equivalence between the two objectives:

$$\operatorname*{argmin}_{S} \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in S_k} \|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2^2$$

$$\operatorname*{argmin}_{S} \sum_{k=1}^{K} \frac{1}{2|S_k|} \sum_{\boldsymbol{x}, \boldsymbol{y} \in S_k} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Analysis of k-means Clustering

Some additional details in k-means:

- How to evaluate clustering results?
- How to find the best K?
  - Validation.
  - We could use the Gap Statistic [1].
- Limitations of k-means?

[1] "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society. 2001.

Evaluation of Clustering: Purity

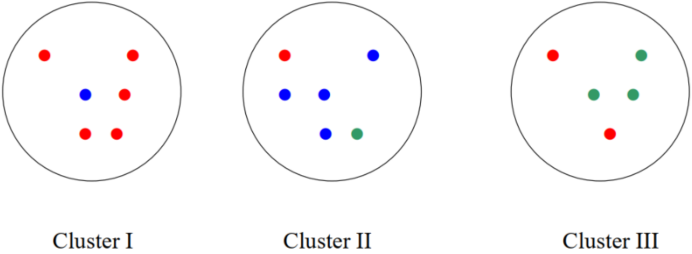
Assesses a clustering with respect to ground truth:

- Requires labels for **some** instances.
- Evaluate its ability to discover the latent classes in gold standard data.
- Suppose each cluster is S_1, S_2, ..., S_K, with n_1, n_2, ..., n_K members, respectively.
- A simple measure:

$$purity(S_k) = \frac{1}{n_k} \max_j (n_{kj}) \quad j \in \{1, 2, ..., K\} \tag{5}$$

  - The ratio between the dominant class and the cluster size.
  - Biased towards having many small clusters.

Evaluation of Clustering: Purity



Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

Evaluation of Clustering: Rand Index

Number of point pairs	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	20	24
Different classes in ground truth	20	72

Evaluation of Clustering: Rand Index

$$RI = \frac{A + D}{A + B + C + D}$$

Compare with standard Precision and Recall:

$$P = \frac{A}{A + B} \qquad R = \frac{A}{A + C}$$

People also define and use a cluster F-measure, which is probably a better measure.

Analysis of k-means Clustering

How to find the best K?

- Validation.
- What if we do not have ground-truths?
- Let W_K denote the loss function value when using K clusters

$$W_K = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in S_k} \|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2^2, \tag{6}$$

- The smaller W_K, the better?

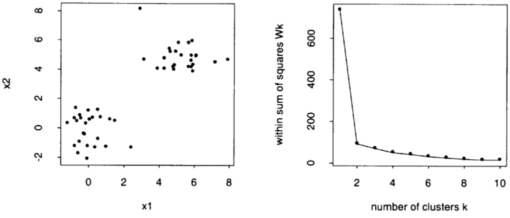
Analysis of k-means Clustering

How to find the best K?

- Let W_K denote the loss function value when using K clusters

$$W_K = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in S_k} \|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2^2, \tag{7}$$

- The smaller W_K, the better?



(a) (b)

Analysis of k-means Clustering

How to find the best K?

- Gap statistic.
  - Generate a synthetic dataset, under the uniform distribution.
  - Perform k-means on the synthetic dataset.
  - Compute the final loss value W'_K.
  - Do this for multiple rounds, and compute the average log(loss) value denoted as $E[\log(W'_K)]$.
  - Compute $E[\log(W'_K)] - \log(W_K)$

Analysis of k-means Clustering

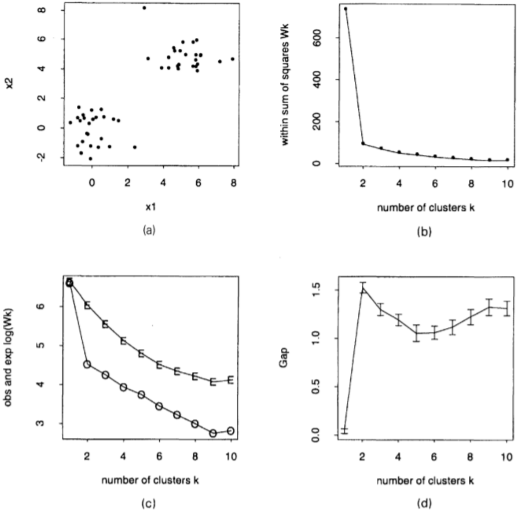


Fig. 1. Results for the two-cluster example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}^*_n\{\log(W_k)\}$ (E); (d) gap curve

Analysis of k-means Clustering

Vanilla k-means algorithm encourages "ball" clusters.