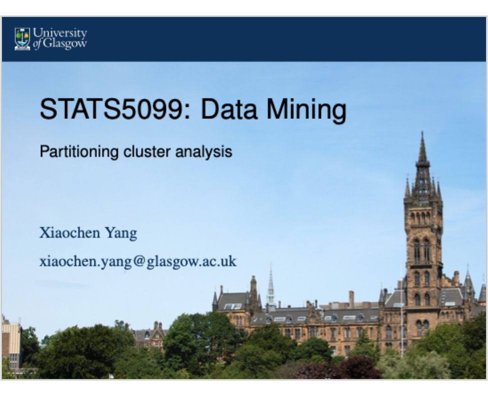
Slide9

Thursday, July 8, 2021 11:53 AM

TutorialSlide9



STATS5099: Data Mining

Partitioning cluster analysis

Xiaochen Yang
xiaochen.yang@glasgow.ac.uk

This week's content

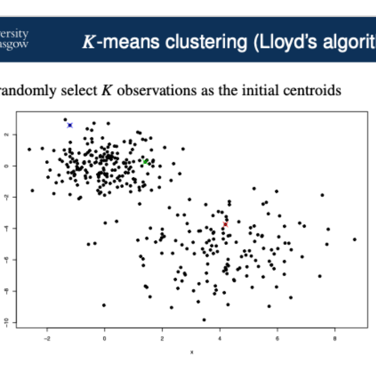
- K-means clustering
- K-medoids clustering

K-means clustering (Lloyd's algorithm)

1. randomly select K observations as the initial centroids (where each one represents a unique cluster);
2. assign each observation to its closest centroid;
3. compute new centroids as the average of all observations that are within a cluster;
4. re- assign each observation to its closest centroid;
5. repeat steps 3 and 4 until the observations are not reassigned or the maximum number of iterations is reached.

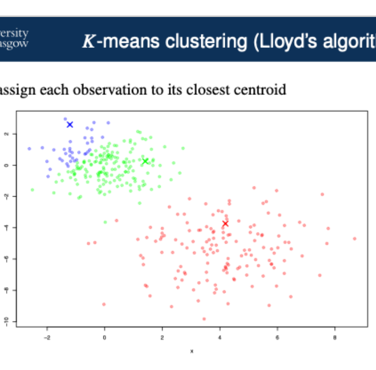
K-means clustering (Lloyd's algorithm)

1. randomly select K observations as the initial centroids



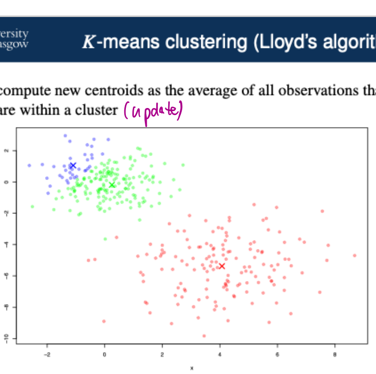
K-means clustering (Lloyd's algorithm)

2. assign each observation to its closest centroid



K-means clustering (Lloyd's algorithm)

2. compute new centroids as the average of all observations that are within a cluster (update)



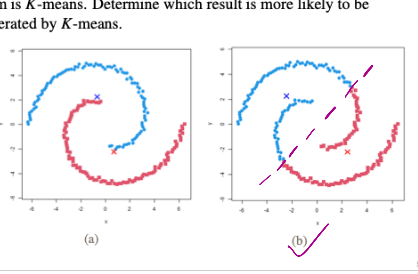
Question

Each dataset is clustered using two different methods, and one of them is K-means. Determine which result is more likely to be generated by K-means.

(a) (b) ✓ nearest centroid

Question

Each dataset is clustered using two different methods, and one of them is K-means. Determine which result is more likely to be generated by K-means.

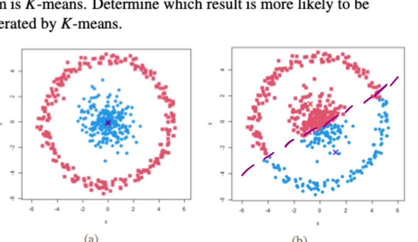


(a) (b) ✓

Question

Each dataset is clustered using two different methods, and one of them is K-means. Determine which result is more likely to be generated by K-means.

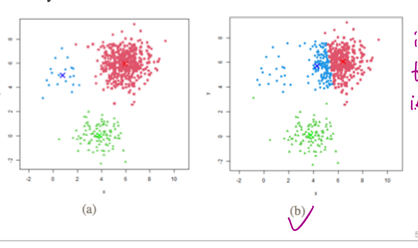
k-means
linear boundary



(a) (b)

Question

Each dataset is clustered using two different methods, and one of them is K-means. Determine which result is more likely to be generated by K-means.



(a) (b) ✓

if use two centers,
total variation is small
i.e. objective function is small

K-means clustering (theory)

K-means clustering attempts to find the assignment of observations to a fixed number of clusters K that minimises the **total within-cluster variation**:

objective function

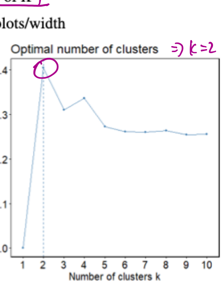
$$W(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} d_E(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_{C_k})^2$$

distance btw sample and mean of its cluster

$$= \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{x_i \in C_k} \sum_{x_j \in C_k} d_E(\boldsymbol{x}_i, \boldsymbol{x}_j)^2$$

K-means clustering (practice)

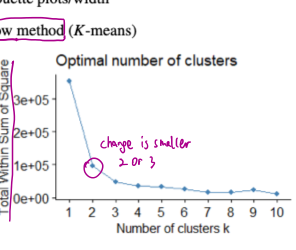
- decide the value of K
  - silhouette plots/width



K-means clustering (practice)

- decide the value of K
  - silhouette plots/width
  - Elbow method (K-means)

also PCA



K-means clustering (practice)

- decide the value of K
  - silhouette plots/width
  - Elbow method (K-means)
- sensitive to initial starting centres

K-means clustering (R)

k / repeat w/ 100 different starting time

```
km3 <- kmeans(data, centers=3, nstart=100)
             #data: 350 rows and 2 columns
km3

K-means clustering with 3 clusters of sizes 77, 203, 70

Cluster means:
         [,1]       [,2]
1 4.976916119 -4.3452691
2 0.004285656  0.0247829
3 3.126691886 -6.5371190

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 ...
[204] 1 1 3 3 1 3 3 3 1 1 3 1 3 3 1 1 1 3 3 1 3 3 1 1 1
 ...
[349] 1 3
```

# of pts w/in each cluster

which cluster each pt belongs

K-means clustering (R)

```
km3 <- kmeans(data, centers=3, nstart=100)
             #data: 350 rows and 2 columns
km3

Within cluster sum of squares by cluster:
[1] 262.8321+446.7470+274.4780
 (between_SS / total_SS =  81.1 %)
Available components:

[1] "cluster"      "centers"     "totss"     "withinss"
[5] "tot.withinss" "betweenss"   "size"      "iter"
[9] "ifault"
```

= total within SS

(total SS - total within SS) / before clustering (k=1)

need a large percentage

convergence of k-means

Pros and cons of K-means

Pros

- easy to understand and implement
- computationally faster than hierarchical agglomerative clustering in the case of a large number of variables
- an observation can change cluster when the centroids are recomputed

Cons

- can only handle numerical variables
- sensitive to outliers
- difficult to decide the optimal K
- assume that we deal with spherical clusters and that each cluster has roughly equal numbers of observations one cluster w/in another cluster

How to understand the drawbacks of K-means

K-medoids clustering

1. start with K randomly selected points for cluster medoids;
2. assign each observation to the cluster with the closest medoid;
3. for each of the K clusters:
   a. for each non-medoid point in the cluster k, make this the new cluster medoid;
   b. compute the cost of the configuration;
   c. choose the **point** with the **lowest cost** as the new cluster medoid;
4. repeat steps 2 and 3 until convergence.

K-medoids clustering

1. start with K randomly selected points for cluster medoids;
2. assign each observation to the cluster with the closest medoid;
3. for each of the K clusters:
   a. for each non-medoid point in the cluster k, make this the new cluster medoid;
   b. compute the cost of the configuration;
   c. choose the **point** with the **lowest cost** as the new cluster medoid;
4. repeat steps 2 and 3 until convergence.

k-means centroid cannot be in original data
k-medoids medoids should be in original data

$$\boldsymbol{x}_{i_k} = \arg\min_{\boldsymbol{x}_{i_k} \in C_k} \sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, \boldsymbol{x}_{i_k})$$

Quizzes

1. Considering the K-means algorithm, after current iteration, we have 3 centroids $(0), (2), (-1)$. Will points (3) and (4) be assigned to the same cluster in the next iteration? Y. Combine with 2.

2. Which of the following statements about the K-means algorithm are correct?
   (a) ✓ The K-means algorithm is sensitive to outliers. Euclidean dist.
   (b) ✗ For different initialisations, the K-means algorithm will definitely give the same clustering results.
   (c) ✓ The centroids in the K-means algorithm may not be any observed data points.
   (d) ✓ The K-means algorithm can detect non-convex clusters.

Quizzes

3. Which of the following statements about the K-medoids clustering algorithm are true?
   (a) ✓ K-medoids algorithm can detect spherical shaped clusters. use dissimilarity function as Euclidean dist.
   (b) ✓ The number of clusters must be specified in advance.
   (c) ✓ K-medoids is less sensitive to outliers than K-means. use different dissimilarity function
   (d) ✗ K-medoids is suitable for large volume of data. computation cost is higher
4. Which of the following statements are true?
   (a) ✓ Clustering analysis is unsupervised learning since it does not require labelled training data.
   (b) ✗ When clustering, we want to put two dissimilar data objects into the same cluster.
   (c) ✓ In order to perform cluster analysis, we need to have a similarity measure between data objects.
   (d) ✗ We must know the number of output clusters in advance for all clustering algorithms. does not need for hierch. algo. (last week)