Lesson 10 Cluster Analysis

Lusine Zilfimian

May 04 (Monday), 2020

Contents

- Quiz

Contents

- Quiz
- Intro to CA

Contents

- Quiz
- Intro to CA
- Different Types of Clusterings

Contents

- Quiz
- Intro to CA
- Different Types of Clusterings
- Different Types of Clusters

Contents

- Quiz
- Intro to CA
- Different Types of Clusterings
- Different Types of Clusters
- K-Means

Contents

- Quiz
- Intro to CA
- Different Types of Clusterings
- Different Types of Clusters
- K-Means
- Agglomerative Hierarchical Clustering

Last Lecture ReCap

- Why the DT is greedy algorithm?

Last Lecture ReCap

- Why the DT is greedy algorithm?
- How to select the test condition to split?

Last Lecture ReCap

- Why the DT is greedy algorithm?
- How to select the test condition to split?
- What is the worst case of distibution after split?

Idea of CA

- Cluster analysis divides data into groups (clusters) that are **meaningful**, **useful**, or both

Idea of CA

- Cluster analysis divides data into groups (clusters) that are **meaningful**, **useful**, or both
- Classification vs Cluster Analysis

Idea of CA

- Cluster analysis divides data into groups (clusters) that are **meaningful**, **useful**, or both
- Classification vs Cluster Analysis
- Emaples

What is the Cluster Analysis?

- The goal is that the objects within a group be similar to one another and different from the objects in other groups.

What is the Cluster Analysis?

- The goal is that the objects within a group be similar to one another and different from the objects in other groups.
- There are different ways of dividing the data into cluster.

Different Types of Clusterings

- Hierarchical versus Partitional

Different Types of Clusterings

- Hierarchical versus Partitional
- Exclusive versus Overlapping versus Fuzzy

Different Types of Clusterings

- Hierarchical versus Partitional
- Exclusive versus Overlapping versus Fuzzy
- Complete versus Partial

Different Types of Clusters

- Well-Separated

Different Types of Clusters

- Well-Separated
- Center-based

Different Types of Clusters

- Well-Separated
- Center-based
- Density-based clusters

Different Types of Clusters

- Well-Separated
- Center-based
- Density-based clusters
- Conceptual clusters

K-means

- Select K points as initial centroids

K-means

- Select K points as initial centroids
- Each point is then assigned to the closest centroid

K-means

- Select K points as initial centroids
- Each point is then assigned to the closest centroid
- Recompute the centroid of each cluster

K-means

- Select K points as initial centroids
- Each point is then assigned to the closest centroid
- Recompute the centroid of each cluster
- Until no point changes clusters

K-means

- Select K points as initial centroids
- Each point is then assigned to the closest centroid
- Recompute the centroid of each cluster
- Until no point changes clusters
- Or other condition

Sensitivity to Initial points

- Randomly selected initial centroids may be poor

Sensitivity to Initial points

- Randomly selected initial centroids may be poor
- Example

How to choose k

- One effective approach is to take a sample of points and cluster them using a hierarchical clustering technique

How to choose k

- One effective approach is to take a sample of points and cluster them using a hierarchical clustering technique
- This approach often works well, but is practical only if the sample is relatively small, a few hundred to a few thousand

How to choose k

- One effective approach is to take a sample of points and cluster them using a hierarchical clustering technique
- This approach often works well, but is practical only if the sample is relatively small, a few hundred to a few thousand
- Select the point that is farthest from any of the initial centroids already selected

Bisecting K-means

- Split the set of all points into two clusters, select one of these clusters to split, and so on

Bisecting K-means

- Split the set of all points into two clusters, select one of these clusters to split, and so on
- It can be use to have hierarchical clustering.

K-means and Different Types of Clusters

- Diferent sizes

K-means and Different Types of Clusters

- Diferent sizes
- Different densities

K-means and Different Types of Clusters

- Diferent sizes
- Different densities
- non-spherical shapes

Goodness of clustering structure

- Unsupervised

Goodness of clustering structure

- Unsupervised
- Supervised

Goodness of clustering structure

- Unsupervised
- Supervised
- By adding new features

Goodness of clustering structure

- Unsupervised
- Supervised
- By adding new features
- Different means

Goodness of clustering structure

- Unsupervised
- Supervised
- By adding new features
- Different means
- By using different algorithms

Disadvantages

- Sensitive to outliers

One potential disadvantage of K-means clustering is that it requires us to pre-specify the number of clusters K.

Disadvantages

- Sensitive to outliers
- Sensitive to initial points

One potential disadvantage of K-means clustering is that it requires us to pre-specify the number of clusters K.

Disadvantages

- Sensitive to outliers
- Sensitive to initial points
- Categorical data

One potential disadvantage of K-means clustering is that it requires us to pre-specify the number of clusters K.

Disadvantages

- Sensitive to outliers
- Sensitive to initial points
- Categorical data
- Required k

One potential disadvantage of K-means clustering is that it requires us to pre-specify the number of clusters K.

Agglomerative Hierarchical Clustering

- Starting with individual points as clusters

Agglomerative Hierarchical Clustering

- Starting with individual points as clusters
- Merge the two closest clusters until only one cluster remains

Agglomerative Hierarchical Clustering

- Starting with individual points as clusters
- Merge the two closest clusters until only one cluster remains
- Two groups of distance measures:

Agglomerative Hierarchical Clustering

- Starting with individual points as clusters
- Merge the two closest clusters until only one cluster remains
- Two groups of distance measures:
- Distance between records

Agglomerative Hierarchical Clustering

- Starting with individual points as clusters
- Merge the two closest clusters until only one cluster remains
- Two groups of distance measures:
- Distance between records
- Distance between clusters

Proximity between Clusters

- Single link

Proximity between Clusters

- Single link
- Complete link

Proximity between Clusters

- Single link
- Complete link
- Group average

Proximity between Clusters

- Single link
- Complete link
- Group average
- Centroid

Proximity between Clusters

- Single link
- Complete link
- Group average
- Centroid
- Medoid