Preliminary survey

1. Experience in programming:

1. Never, 2. Less than one year, 3. Over one year

2. Experience in data science:

1. Never, 2. Less than one year, 3. Over one year

3. Fill in the blanks with suitable words or phrases from the following options:

DO NOT refer to any materials.

Please select "I don't know" if you are not sure about the answer.

Clustering is the task of _________.

A. reducing the number of random variables under consideration by obtaining a set of principal variables

B. finding the most frequent and relevant patterns in large datasets.

C. grouping a set of data samples in such a way that samples in the same group are more similar to each other than to those in other groups

D. identifying to which of a set of categories a new observation belongs

E. I don't know.

Clustering belongs to __________ learning because data samples have no labels.

A. supervised

B. unsupervised

C. semi-supervised

D. self-organized

E. I don't know.

Many clustering algorithms require to __________ in advance.

A. set the number of clusters

B. annotate data samples so that algorithms can learn to classify information

C. reduce the number of dimensions in the dataset

D. make polynomial features of given degrees

E. I don't know.

__________ is an index that estimates the optimal number of clusters.

A. F-measure

B. Jaccard index

C. Hamming distance

D. Gap statistic

E. I don’t know.

If __________ than that of the other attributes, the effect of the attribute is ignored.

A. the scale of an attribute is much larger

B. the scale of an attribute is much smaller

C. I don’t know.

Therefore, __________ is commonly used to normalize the range of attributes of data.

A. standardization

B. feature selection

C. dimensionality reduction

D. average pooling

E. I don’t know.

__________ is commonly used to encode categorical features into numerical features before clustering.

A. Neighbor embedding

B. Feature embeddings

C. Target encoding

D. One-hot encoding

E. I don’t know.

__________ clustering is one of the most commonly used clustering algorithms.

A. K-means

B. Nearest neighbor

C. t-SNE

D. Logistic

E. I don’t know.

If data has too many features, clustering algorithms severely degrade their performance due to __________.

A. no free lunch theorem

B. ugly duckling theorem

C. curse of dimensionality

D. Laplace’s demon

E. I don’t know.

To address the curse of dimensionality, __________ is one of the effective options.

A. feature selection

B. normalization

C. k-means++

D. multiple imputation

E. I don’t know.

K-means clustering algorithm assumes that __________.

A. the data has hierarchical structure

B. clusters are spherical and of equal size

C. all attributes are scaled individually

D. the data has no outliers

E. I don’t know.

Hierarchical clustering algorithms are broadly divided into __________ approaches.

A. linear and non-linear

B. stochastic and deterministic

C. agglomerative and divisive

D. batch and online

E. I don’t know.

The results of hierarchical clustering can visualize by using a __________.

A. directed graph

B. disjunctive graph

C. dendrogram

D. histogram

E. I don’t know.

Proximity along the horizontal axis of the dendrogram __________.

A. represents the similarity of two observations

B. doesn't represent the similarity of two observations

C. I don’t know.

Here is a result of clustering data analysis. Read carefully the report below.

Analysis report

Results

Algorithms	Estimated number of clusters			
	Gap statistic	Silhouette score	Davies-Bouldin score	Calinski and Harabasz score
K-means	3	2	2	2
Hierarchical Clustering	3	3	3	3

According to the majority rule, the estimated number of clusters is **3**.

Clustering

- Clustering is the task of **grouping a set of data samples in such a way that samples in the same group are more similar to each other** than to those in other groups.
- In general, data samples have no labels so that we **need to interpret the results of clustering**.
  This type of problems are called unsupervised learning.
- Many clustering algorithms **require to set the number of clusters in advance**.
  However, the optimal number of clusters is unknown.
  Therefore, our system conducts the cluster analysis with a range of candidates of the number of clusters.
- A wide variety of indices have been proposed to find the optimal number of clusters.
  This system **estimates the optimal number of clusters by majority vote** of these indices.
- The indices that MALSS uses are as follows: Gap statistic, Silhouette score, Davies-Bouldin score, Calinski and Harabasz score.

Data summary

Number of rows	150
Number of columns	4 (numerical: 4, categorical: 0)

- Clustering algorithms are affected by the difference of the scale of each attributes.
  If the scale of an attribute is much smaller than that of the other attributes (e.g. the attribute has one digit and the other attributes have five digits.), **the effect of the attribute is ignored**.
  Therefore, **this system standardizes the data** to a mean is zero and standard deviation is one by default.
  If scaling is not needed, set *standardize* parameter to *True*.
- Note that some features (e.g. country code) may have to be handled as the categorical feature even though they look like numerical features.

In such case, you need to encode categorical features into numerical features by yourself before setting data to this system.

One-hot encoding is commonly used.

- **Clustering algorithms work poorly as the number of features increases**, which is known as **curse of dimensionality**.

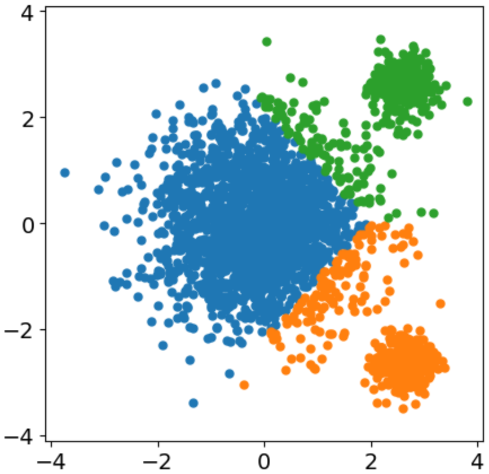
  The one-hot encoding sometimes causes the problem.

  Feature selection or **dimensionality reduction** may help in such case.

K-means clustering

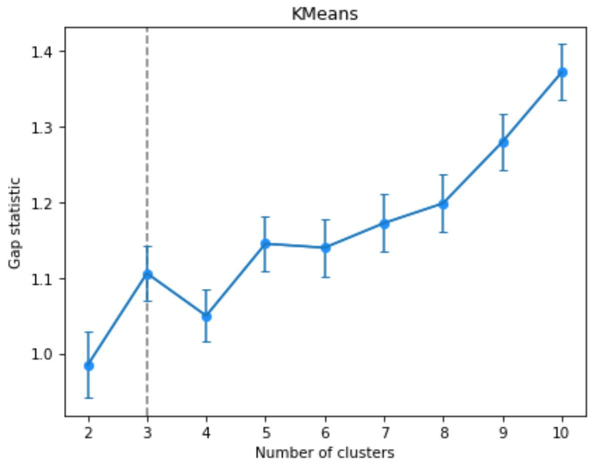
- K-means clustering is one of the most commonly used clustering algorithm.
- Note that **k-means algorithm assumes that clusters are spherical and of equal size**.

  If clusters are not spherical or not of equal size, k-means algorithm may produce undesirable results (see the figure below).

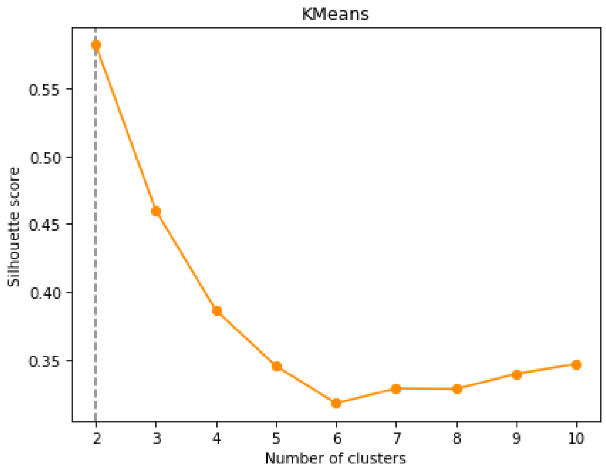


k-means algorithm doesn't work well when clusters are not of equal size.

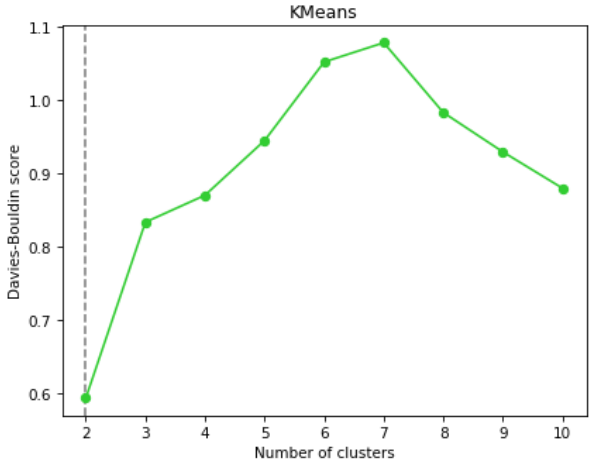
Results of estimating the number of clusters



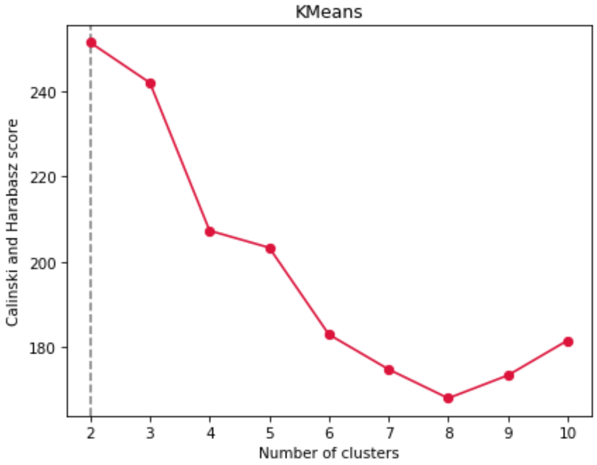
Gap statistics



Silhouette score



Davies-Bouldin score



Calinski and Harabasz score

Hierarchical clustering

- Hierarchical clustering is one of the most commonly used clustering algorithm as well as k-means algorithm.
- Hierarchical clustering algorithms are broadly divided into two groups.

  One is **agglomerative (bottom-up) approach** that each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

  The other is **divisive (top-down) approach** that all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

  This system supports agglomerative clustering methods.
- Hierarchical clustering algorithms have an advantage that they can visualize the results of clustering as a cluster tree called **dendrogram**.

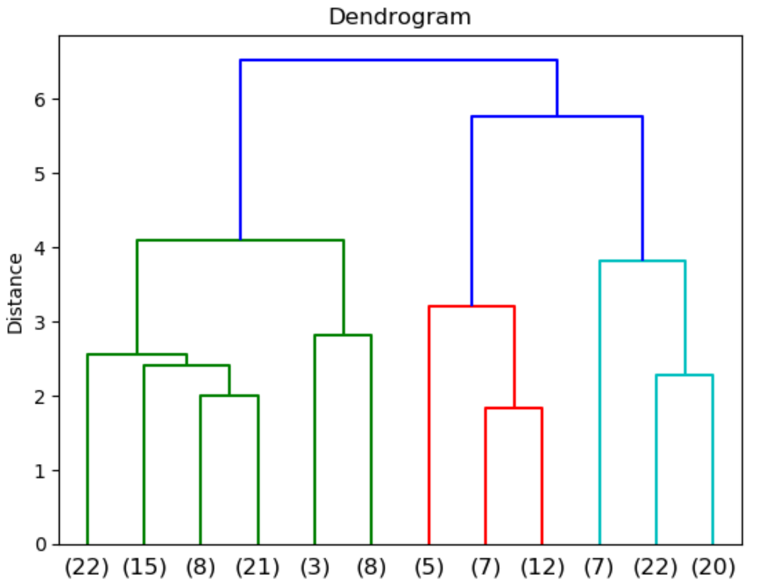
  Note the following points when referring the dendrogram:

  - **Proximity along the horizontal axis of the dendrogram doesn't represent the similarity of two observations**.

    We need to see the location on the vertical axis where branches containing those two observations first are fused.
  - **The assumption that an arbitrary data has hierarchical structure might be unrealistic** though the results of hierarchical clustering always have hierarchical structures.
  - Dendrogram strongly depends on the type of linkage, which defines the dissimilarity between two groups of observations.

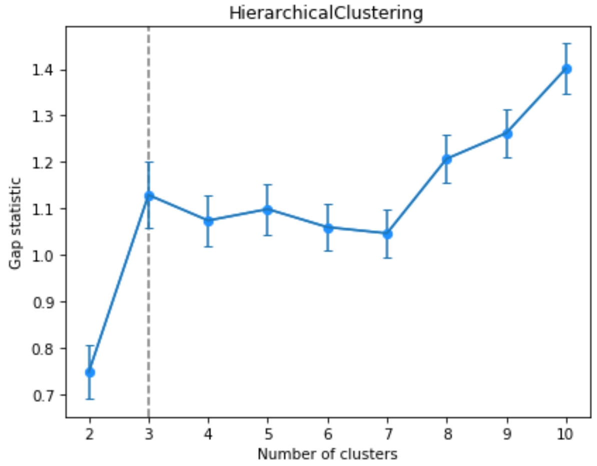
    This system adopts a *complete linkage* method.

Dendrogram

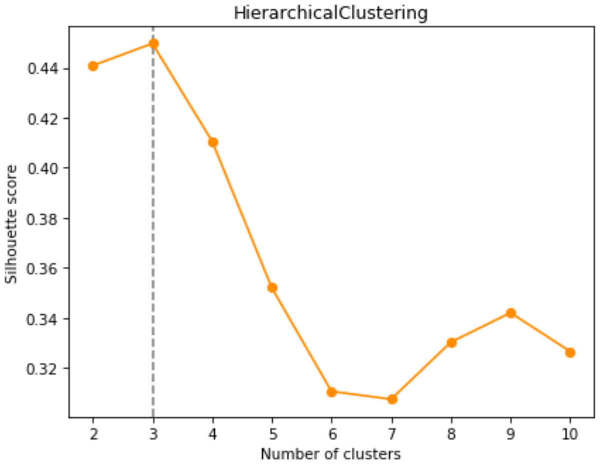


A number with brackets represents the number of the samples in the node and a number without brackets represents the index of the sample in the data.

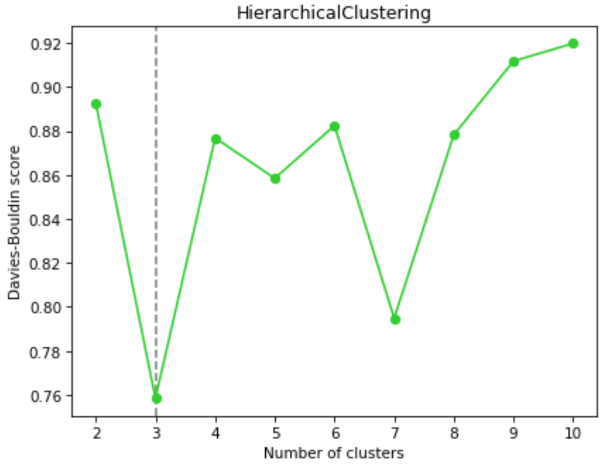
Results of estimating the number of clusters



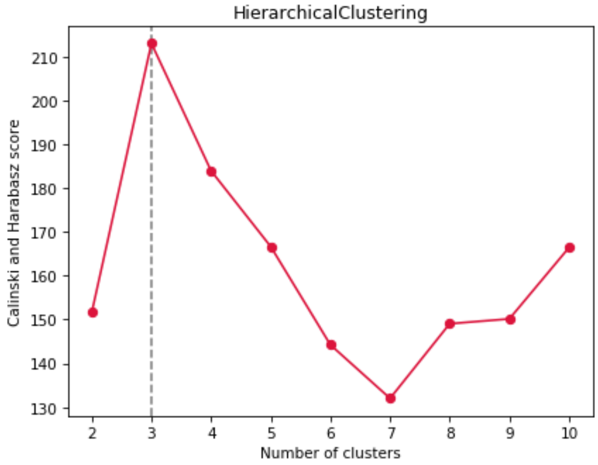
Gap statistics



Silhouette score



Davies-Bouldin score



Calinski and Harabasz score

Thank you for completing our survey.

Please input your Worker ID below, and enter COMPL3T3 as your completion code in MTurk.

Worker ID:_______________