Workshop 8

COMP20008 Elements of Data Processing

Learning outcomes

By the end of today's class you should be able to:

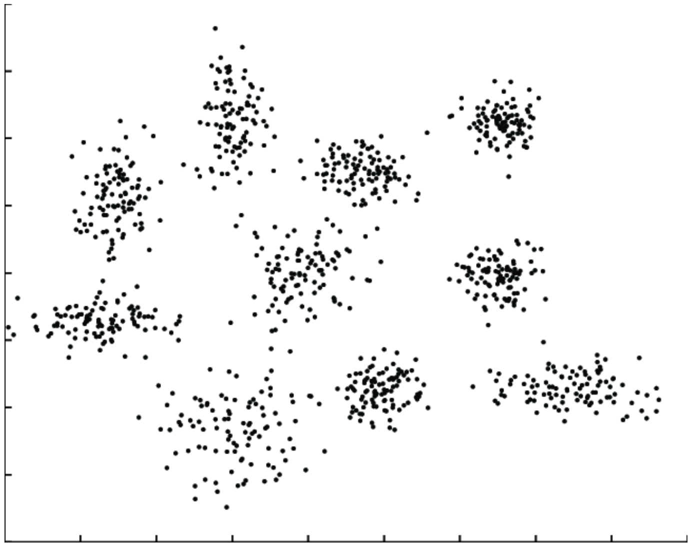
- explain k-means and hierarchical clustering
- interpret Visual Assessment for Clustering Tendency (VAT) plots

clustering

- perform linear regression in Python
- evaluate the quality of a regression model based on the residuals and coefficient of determination

linear regression

k-means clustering



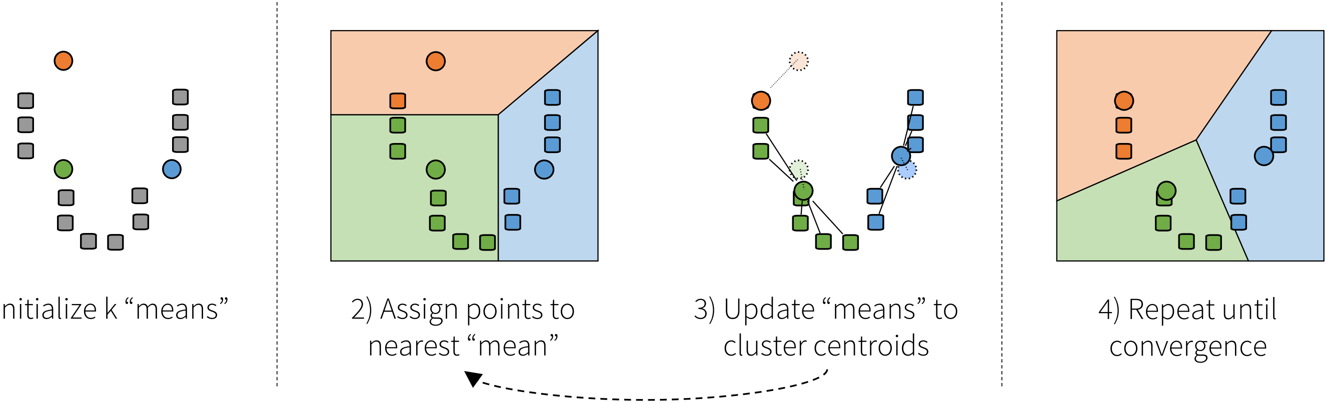
Clustering: dividing items into groups (clusters) according to similarity

k-means divides items into $\boldsymbol{k}$ clusters so as to *minimize* the total within-cluster variance

$$\sum_{i=1}^{k} \sum_{x \in c_i} \mathrm{dist}(x, \bar{c}_i)^2$$

Image source: Liang et al. (2012). Determining the number of clusters using information entropy for mixed data. Pattern Recognition. 45. 2251-2265.

k-means algorithm



1) Initialize k "means" | 2) Assign points to nearest "mean" | 3) Update "means" to cluster centroids | 4) Repeat until convergence

Image source: Wikipedia Commons

Q1: k-means in 1D

Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \ldots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

Q1: k-means in 1D

Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \ldots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

Existing centroids:

Update assignments:

Update centroids:

Q1: k-means in 1D

Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \ldots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

Iteration 1

Existing centroids: [1, 2]

Update assignments: [0, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Update centroids: [1, 6]

Q1: k-means in 1D

Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \ldots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

Iteration 2

Existing centroids:	[1, 6]
Update assignments:	[0, 0, 0, 1, 1, 1, 1, 1, 1, 1]
Update centroids:	[2, 7]

Q1: k-means in 1D

Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \ldots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

Iteration 3

Existing centroids:	[2, 7]
Update assignments:	[0, 0, 0, 0, 1, 1, 1, 1, 1, 1]
Update centroids:	[2.5, 7.5]

Q1: k-means in 1D

Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \ldots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

Iteration 4

Existing centroids:	[2.5, 7.5]
Update assignments:	[0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
Update centroids:	[3, 8]

Q1: k-means in 1D

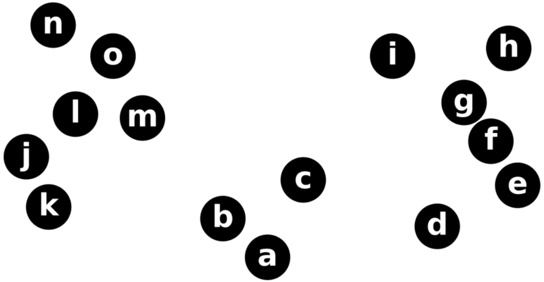
Consider the 1-dimensional data set with 10 data points $\{1, 2, 3, \ldots, 10\}$. Show the iterations of the k-means algorithm using Euclidean distance when $k = 2$, and the random seeds are initialized to $\{1, 2\}$.

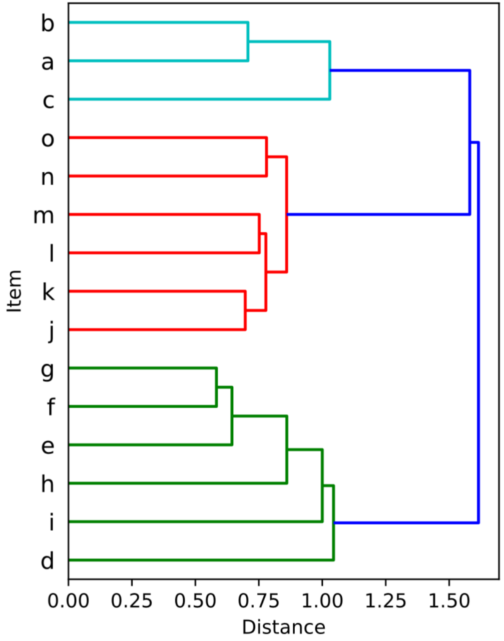
Iteration 4

Existing centroids:	[3, 8]
Update assignments:	[0, 0, 0, 0, 0, 1, 1, 1, 1, 1]
Update centroids:	[3, 8]

Agglomerative clustering

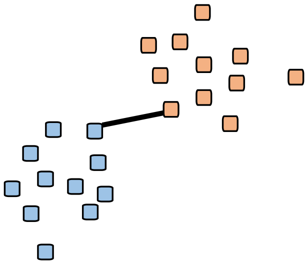
- “Bottom-up” hierarchical clustering
- Each item starts in its own cluster
- “Closest” clusters are merged moving up the hierarchy



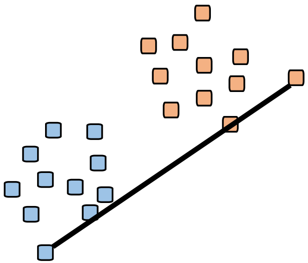


Agglomerative clustering

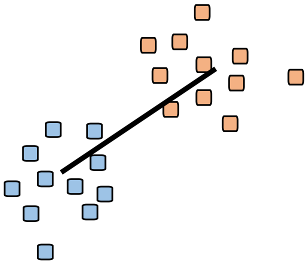
Need to choose a *linkage criterion*: a way of measuring the distance between pairs of clusters



Single-linkage



Complete-linkage



Centroid linkage

Q2: Single linkage clustering

Repeat Q1 using agglomerative hierarchical clustering and Euclidean distance, with single linkage (min) criterion.

Hint: Repeat the following until all clusters are merged

1. Compute the Euclidean distance between all pairs of clusters
2. Merge the two closest clusters

Q2: Single linkage clustering

Step 1: Compute the distance matrix

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0

Q2: Single linkage clustering

Step 2: Choose the closest clusters to merge

	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0

Q2: Single linkage clustering

Step 3: Update the distance matrix. Find the distance between the new cluster “11” = (1, 2) and the other clusters.

	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0

	11	3	4	5	6	7	8	9	10
11	0	1							
3	1	0	1	2	3	4	5	6	7
4		1	0	1	2	3	4	5	6
5		2	1	0	1	2	3	4	5
6		3	2	1	0	1	2	3	4
7		4	3	2	1	0	1	2	3
8		5	4	3	2	1	0	1	2
9		6	5	4	3	2	1	0	1
10		7	6	5	4	3	2	1	0

Q2: Single linkage clustering

Step 3: Update the distance matrix. Find the distance between the new cluster “11” = (1, 2) and the other clusters.

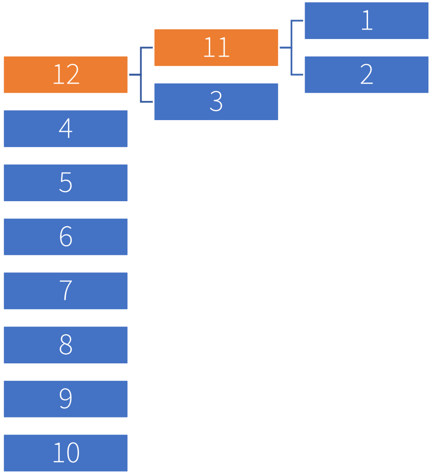
	1	2	3	4	5	6	7	8	9	10
1	0	1	2	3	4	5	6	7	8	9
2	1	0	1	2	3	4	5	6	7	8
3	2	1	0	1	2	3	4	5	6	7
4	3	2	1	0	1	2	3	4	5	6
5	4	3	2	1	0	1	2	3	4	5
6	5	4	3	2	1	0	1	2	3	4
7	6	5	4	3	2	1	0	1	2	3
8	7	6	5	4	3	2	1	0	1	2
9	8	7	6	5	4	3	2	1	0	1
10	9	8	7	6	5	4	3	2	1	0

	11	3	4	5	6	7	8	9	10
11	0	1	2	3	4	5	6	7	8
3	1	0	1	2	3	4	5	6	7
4	2	1	0	1	2	3	4	5	6
5	3	2	1	0	1	2	3	4	5
6	4	3	2	1	0	1	2	3	4
7	5	4	3	2	1	0	1	2	3
8	6	5	4	3	2	1	0	1	2
9	7	6	5	4	3	2	1	0	1
10	8	7	6	5	4	3	2	1	0

Q2: Single linkage clustering

Repeat step 2: Choose the closest clusters to merge

	11	3	4	5	6	7	8	9	10
11	0	1	2	3	4	5	6	7	8
3	1	0	1	2	3	4	5	6	7
4	2	1	0	1	2	3	4	5	6
5	3	2	1	0	1	2	3	4	5
6	4	3	2	1	0	1	2	3	4
7	5	4	3	2	1	0	1	2	3
8	6	5	4	3	2	1	0	1	2
9	7	6	5	4	3	2	1	0	1
10	8	7	6	5	4	3	2	1	0



Q2: Single linkage clustering

Repeat step 3: Update the distance matrix. Find the distance between the new cluster “12” = ((1, 2), 3) and the other clusters.

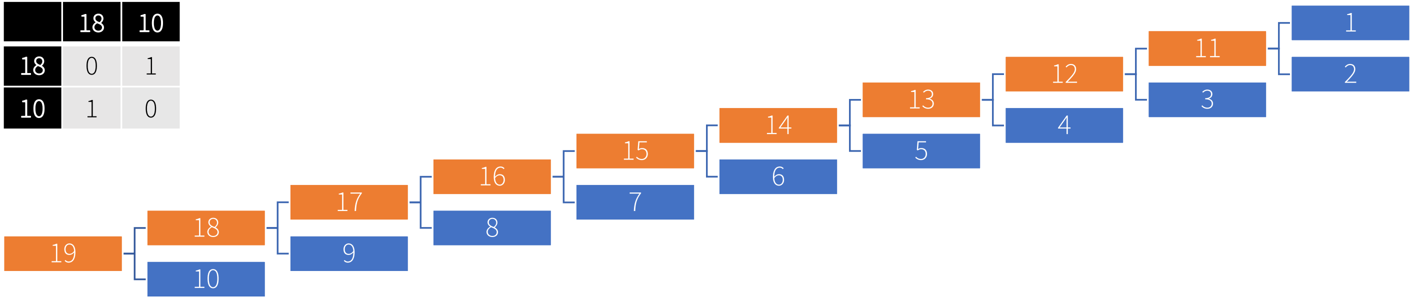
	11	3	4	5	6	7	8	9	10
11	0	1	2	3	4	5	6	7	8
3	1	0	1	2	3	4	5	6	7
4	2	1	0	1	2	3	4	5	6
5	3	2	1	0	1	2	3	4	5
6	4	3	2	1	0	1	2	3	4
7	5	4	3	2	1	0	1	2	3
8	6	5	4	3	2	1	0	1	2
9	7	6	5	4	3	2	1	0	1
10	8	7	6	5	4	3	2	1	0

	12	4	5	6	7	8	9	10
12	0	1	2	3	4	5	6	7
4	1	0	1	2	3	4	5	6
5	2	1	0	1	2	3	4	5
6	3	2	1	0	1	2	3	4
7	4	3	2	1	0	1	2	3
8	5	4	3	2	1	0	1	2
9	6	5	4	3	2	1	0	1
10	7	6	5	4	3	2	1	0

Q2: Single linkage clustering

Iterate… End up with cluster “18” = ((((((((1, 2), 3), 4), 5), 6), 7), 8), 9) and cluster 10

	18	10
18	0	1
10	1	0



Q3: K-means quality measure

Sum-squared error

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x \in c_i} \mathrm{dist}(x, \overline{c_i})^2$$

where $\boldsymbol{c_i}$ is the $\boldsymbol{i}$-th cluster, $\overline{\boldsymbol{c_i}}$ is the centroid for the $\boldsymbol{i}$-th cluster.

Which is better: high SSE or low SSE?

Q3: K-means quality measure

Sum-squared error

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in c_i} \text{dist}(x, \overline{c_i})^2$$

where $\boldsymbol{c_i}$ is the $\boldsymbol{i}$-th cluster, $\overline{\boldsymbol{c_i}}$ is the centroid for the $\boldsymbol{i}$-th cluster.

Which is better: high SSE or low SSE?

Low SSE

Q3: K-means quality measure

As the number of clusters $\boldsymbol{k}$ increases, would you expect the SSE to increase or decrease?

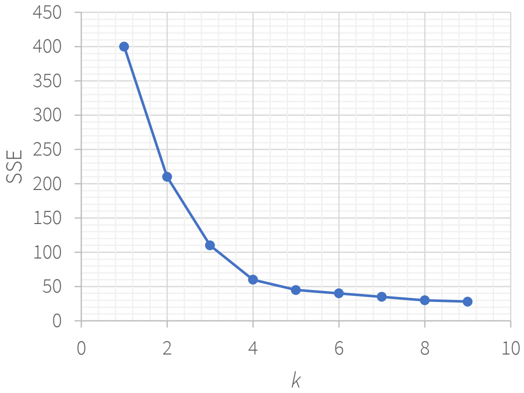
Q3: K-means quality measure

As the number of clusters $\boldsymbol{k}$ increases, would you expect the SSE to increase or decrease?

Expect a *decrease*

There's a method for choosing $\boldsymbol{k}$ based on the SSE:

- Plot SSE versus $\boldsymbol{k}$
- Set optimal $\boldsymbol{k}$ at the elbow
- Negligible improvement in fit beyond the elbow



Q3: K-means quality measure

Suggest another method to evaluate the quality of clustering.

Q3: K-means quality measure

Suggest another method to evaluate the quality of clustering.

- Visualisation (e.g. plot silhouette scores grouped by cluster)
- Sum of inter-cluster distances (want it to be large)
- Dunn index: ratio of smallest inter-cluster distance to largest cluster diameter (want it to be large)
- Silhouette score: for each item find the distance to the closest cluster and subtract the average intra-cluster distance, then normalise (larger score is better)