Lecture 12: Clustering

COMP90049
Knowledge Technologies

Hasti Samadi & Sarah Erfani & Karin Verspoor, CIS

Semester 2, 2019

THE UNIVERSITY OF MELBOURNE
POSTERA CRESCAM LAUDE

Dataset with no label

Outlook	Temperature	Humidity	Windy
sunny	hot	high	FALSE
sunny	hot	high	TRUE
overcast	hot	high	FALSE
rainy	mild	high	FALSE
rainy	cool	normal	FALSE
rainy	cool	normal	TRUE
⋮	⋮	⋮	⋮

Dataset with no label

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

Outlook	Temperature	Humidity	Windy
sunny	hot	high	FALSE
sunny	hot	high	TRUE
overcast	hot	high	FALSE
rainy	mild	high	FALSE
rainy	cool	normal	FALSE
rainy	cool	normal	TRUE
⋮	⋮	⋮	⋮

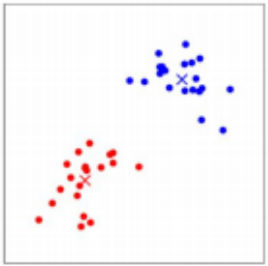
- What can we do with a data set without labels?

Clustering

- We can find groups of data in our dataset which are **similar (close)** to one another -- what we call **clusters**.

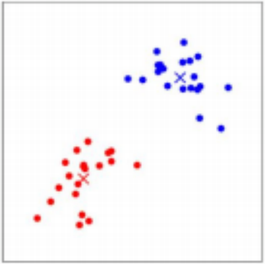
Clustering

- We can find groups of data in our dataset which are **similar (close)** to one another -- what we call **clusters**.



Clustering

- We can find groups of data in our dataset which are **similar (close)** to one another -- what we call **clusters**.



- In clustering it is also important that the objects (instances) in each cluster are **different from** (or unrelated to) the objects in other groups.

Clustering

- Clustering is an ***unsupervised learner***
- Applications in:
  - Pattern recognition
  - Spatial data analysis
  - Medical diagnosis
  - Marketing…

A possible clustering of the weather dataset

Outlook	Temperature	Humidity	Windy	Cluster
sunny	hot	high	FALSE	?
sunny	hot	high	TRUE	?
overcast	hot	high	FALSE	?
rainy	mild	high	FALSE	?
rainy	cool	normal	FALSE	?
rainy	cool	normal	TRUE	?
overcast	cool	normal	TRUE	?
sunny	mild	high	FALSE	?
sunny	cool	normal	FALSE	?
rainy	mild	normal	FALSE	?
sunny	mild	normal	TRUE	?
overcast	mild	normal	TRUE	?
overcast	hot	high	FALSE	?
rainy	mild	high	TRUE	?

A possible clustering of the weather dataset

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

Outlook	Temperature	Humidity	Windy	Cluster
sunny	**hot**	**high**	FALSE	0
sunny	**hot**	**high**	TRUE	0
overcast	**hot**	**high**	FALSE	0
rainy	**mild**	high	FALSE	1
rainy	**cool**	**normal**	FALSE	1
rainy	**cool**	**normal**	TRUE	1
overcast	**cool**	**normal**	TRUE	1
sunny	mild	**high**	FALSE	0
sunny	**cool**	**normal**	FALSE	1
rainy	**mild**	**normal**	FALSE	1
sunny	**mild**	**normal**	TRUE	1
overcast	**mild**	**normal**	TRUE	1
overcast	**hot**	**high**	FALSE	0
rainy	**mild**	high	TRUE	1

Clustering over the weather dataset (cf. outputs)

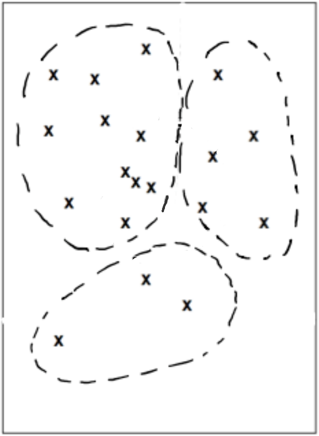
Outlook	Temperature	Humidity	Windy	Cluster	Play
sunny	hot	high	FALSE	0	no ✓
sunny	hot	high	TRUE	0	no ✓
overcast	hot	high	FALSE	0	yes ✗
rainy	mild	high	FALSE	1	yes ✓
rainy	cool	normal	FALSE	1	yes ✓
rainy	cool	normal	TRUE	1	no ✗
overcast	cool	normal	TRUE	1	yes ✓
sunny	mild	high	FALSE	0	no ✓
sunny	cool	normal	FALSE	1	yes ✓
rainy	mild	normal	FALSE	1	yes ✓
sunny	mild	normal	TRUE	1	yes ✓
overcast	mild	normal	TRUE	1	yes ✓
overcast	hot	high	FALSE	0	yes ✗
rainy	mild	high	TRUE	1	no ✗

Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?

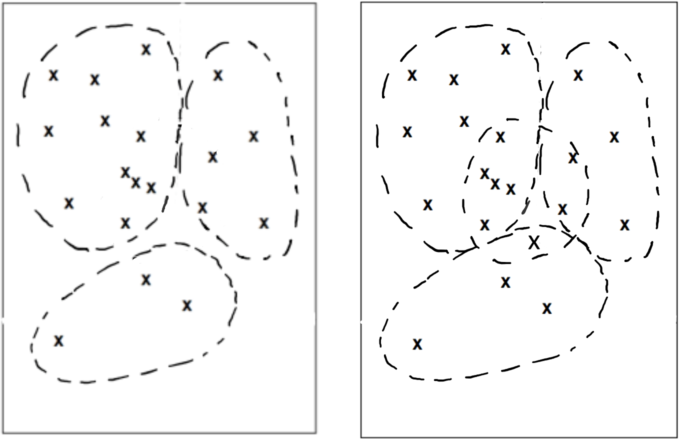
Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?



Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?



Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?
- Deterministic vs. probabilistic clustering (Hard vs. soft clustering)
  - Can an item be partially or weakly in a cluster?

Instance	*Cluster*
1	2
2	3
3	2
4	1
5	2
6	2
7	4
:	:

	Cluster			
Instance	1	2	3	4
1	0.01	**0.87**	0.12	0.00
2	0.05	0.25	**0.67**	0.03
3	0.00	**0.98**	0.02	0.00
4	**0.45**	0.39	0.08	0.08
5	0.01	**0.99**	0.00	0.00
6	0.07	**0.75**	0.08	0.10
7	0.23	0.10	0.20	**0.47**
:	:			

Clustering, basic contrasts

Clustering

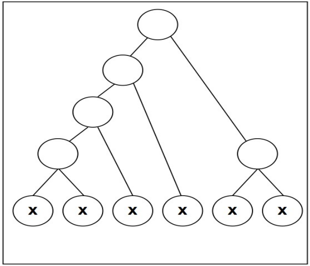
COMP90049 Knowledge Technologies

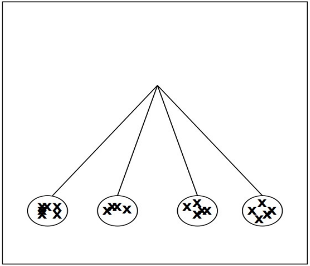
Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?
- Deterministic vs. probabilistic clustering (Hard vs. soft clustering)
  - Can an item be partially or weakly in a cluster?
- Hierarchical vs. partitioning clustering
  - Do the clusters have subset relationships between them? e.g. nested in a tree?





Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?
- Deterministic vs. probabilistic clustering (Hard vs. soft clustering)
  - Can an item be partially or weakly in a cluster?
- Hierarchical vs. partitioning clustering
  - Do the clusters have subset relationships between them? e.g. nested in a tree?
- Partial vs. complete
  - In some cases, we only want to cluster some of the data

Clustering, basic contrasts

- Partial vs. complete
  - In some cases, we only want to cluster some of the data

Clustering, basic contrasts

Clustering

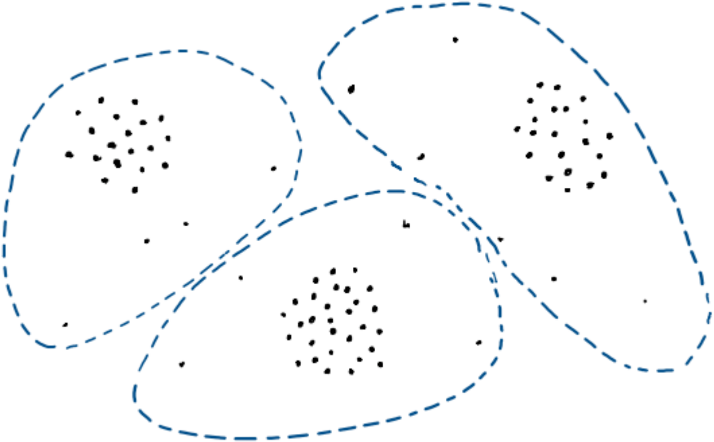
COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

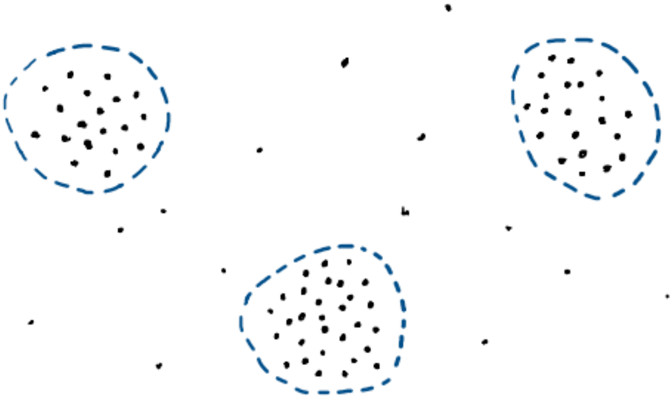
Summary
Resources

- Partial vs. complete
  - In some cases, we only want to cluster some of the data



Clustering, basic contrasts

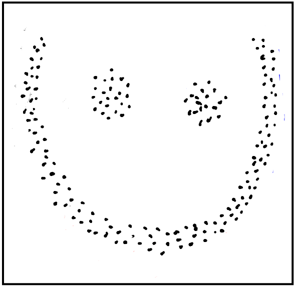
- Partial vs. complete
  - In some cases, we only want to cluster some of the data

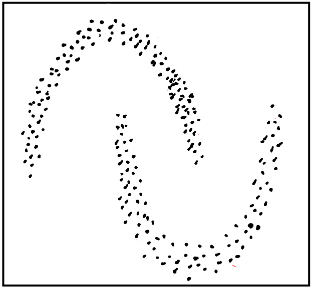


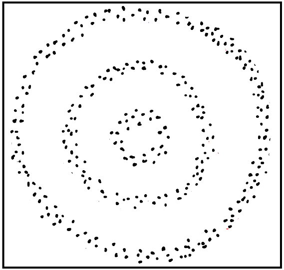
Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?
- Deterministic vs. probabilistic clustering (Hard vs. soft clustering)
  - Can an item be partially or weakly in a cluster?
- Hierarchical vs. partitioning clustering
  - Do the clusters have subset relationships between them? e.g. nested in a tree?
- Partial vs. complete
  - In some cases, we only want to cluster some of the data
- Heterogenous vs. homogenous
  - Clusters of widely different sizes, shapes, and densities

Heterogenous vs. homogenous







Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?
- Deterministic vs. probabilistic clustering (Hard vs. soft clustering)
  - Can an item be partially or weakly in a cluster?
- Hierarchical vs. partitioning clustering
  - Do the clusters have subset relationships between them? e.g. nested in a tree?
- Partial vs. complete
  - In some cases, we only want to cluster some of the data
- Heterogenous vs. homogenous
  - Clusters of widely different sizes, shapes, and densities
- Incremental vs. batch clustering
  - Is the whole set of items clustered in one go?

What is a good clustering?

Two clusters?

Four clusters?

Six clusters?

What number of clusters is better?

2 Clusters

4 Clusters

6 clusters

18 clusters

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Clustering Evaluation

Clustering evaluation measures come in two basic types:

- **Unsupervised:** Measures the goodness of a clustering structure without respect to external information.
  - How cohesive are individual clusters?
  - How separate is one cluster from other clusters?
- **Supervised:** how well do cluster labels match externally supplied class labels?

Unsupervised Evaluation

A "good" cluster analysis should have one or both of:

- **high cluster cohesion**, i.e. instances in a given cluster should be closely related to each other

$$cohesion(C_i) = \frac{1}{\sum_{x,y \in C_i} Distance(x,y)}$$

- **high cluster separation**, i.e. instances in different clusters should be distinct from each other

$$separation(C_i, C_j) = \sum_{x \in C_i, y \in C_{j \neq i}} Distance(x,y)$$

Unsupervised Evaluation

- Most common measure evaluating the quality of clusters (esp. for k-means) is **Sum of Squared Error (SSE)** or *Scatter*

Sum of Squared Error (SSE) :

- For each point, the error is the distance to the nearest cluster
- To get SSE, we square these errors and sum them.

$$\sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x)$$

- x is a data point in cluster C_i and m_i is the representative point for cluster C_i (Centroid)
- Can show that the m_i that minimises SSE corresponds to the center (mean) of the cluster
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase k , the number of clusters
- However, a good clustering with smaller k can have a lower SSE than a poor clustering with higher k

Two clusters?

Six clusters?

Supervised Evaluation

- **Supervised** evaluation of cluster "validity" measures the degree to which predicted class labels match the actual class labels, e.g. based on the distribution of actual class labels within each cluster.

Clustering over the weather dataset (cf. outputs)

Outlook	Temperature	Humidity	Windy	Cluster	Play
sunny	**hot**	**high**	FALSE	0	no
sunny	**hot**	**high**	TRUE	0	no
overcast	**hot**	**high**	FALSE	0	yes
rainy	**mild**	high	FALSE	1	yes
rainy	**cool**	**normal**	FALSE	1	yes
rainy	**cool**	**normal**	TRUE	1	no
overcast	**cool**	**normal**	TRUE	1	yes
sunny	mild	**high**	FALSE	0	no
sunny	**cool**	**normal**	FALSE	1	yes
rainy	**mild**	**normal**	FALSE	1	yes
sunny	**mild**	**normal**	TRUE	1	yes
overcast	**mild**	**normal**	TRUE	1	yes
overcast	**hot**	**high**	FALSE	0	yes
rainy	**mild**	high	TRUE	1	no

Distance / Similarity / Proximity

Based our definition:

- Clustering is finding groups of data in our dataset which are **similar or close** to one another and **different or separated** from other clusters

- A key component of any clustering algorithm is a measurement of the distance between any points.

Measuring the distance

- **Data points** (in Euclidean space)
  - Euclidean (L2) distance
  - Manhattan (L1) distance

Measuring the distance

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

- **Data points** (in Euclidean space)
    - Euclidean (L2) distance
    - Manhattan (L1) distance
- **Discrete values**
    - Hamming distance (discrepancy between the bit strings)

Sunny	011
Overcast	101
Rainy	110

For two bit strings, the number of positions at which the corresponding symbols are different

Measuring the distance

- **Data points** (in Euclidean space)
  - Euclidean (L2) distance
  - Manhattan (L1) distance
- **Discrete values**
  - Hamming distance (discrepancy between the bit strings)

Sunny	011
Overcast	101
Rainy	110

For two bit strings, the number of positions at which the corresponding symbols are different

011
101

Measuring the distance

- **Data points** (in Euclidean space)
  - Euclidean (L2) distance
  - Manhattan (L1) distance
- **Discrete values**
  - Hamming distance (discrepancy between the bit strings)

Sunny	011
Overcast	101
Rainy	110

For two bit strings, the number of positions at which the corresponding symbols are different

011
101

Hamming Distance = 2

Measuring the distance

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

- **Data points** (in Euclidean space)
  - Euclidean (L2) distance
  - Manhattan (L1) distance
- **Discrete values**
  - Hamming distance (discrepancy between the bit strings)
- **Documents**
  - Cosine similarity
  - Jaccard measure

Measuring the distance

- **Data points** (in Euclidean space)
  - Euclidean (L2) distance
  - Manhattan (L1) distance
- **Discrete values**
  - Hamming distance (discrepancy between the bit strings)
- **Documents**
  - Cosine similarity
  - Jaccard measure
- **Other measures**
  - Correlation
  - Graph-based measures

k-means Clustering

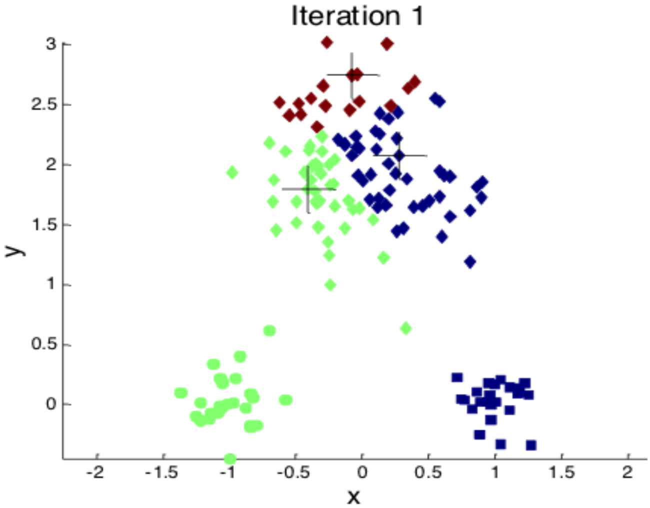
- **K-Means** is one of the most popular "clustering" algorithms.
- Group the dataset into **K** clusters, with the use of **K centroids**

k-means Clustering – Algorithm

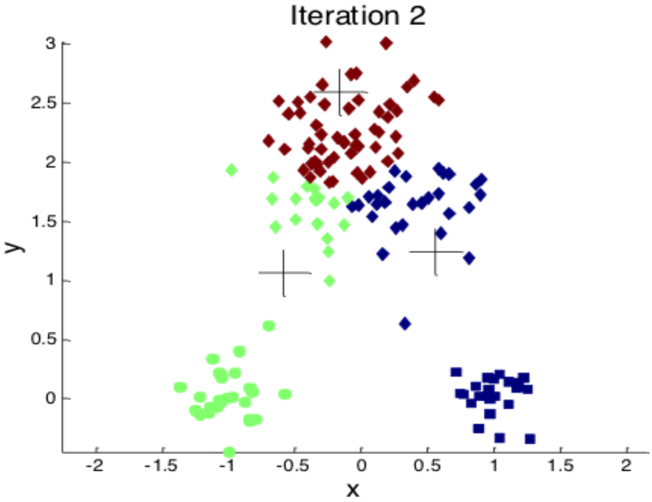
Given **k**, the k-means algorithm is implemented as follows:

1. Select K points at random to act as ***seed*** clusters $(\mu_1, \ldots, \mu_k)$, the initial ***centroids***
2. ***repeat***
3. Assign each instance to the cluster with ***nearest*** centroid
4. Recompute the centroids of each clusters (the centroid is the center, i.e. *mean* point of the cluster)
5. ***Until*** the centroids don't change

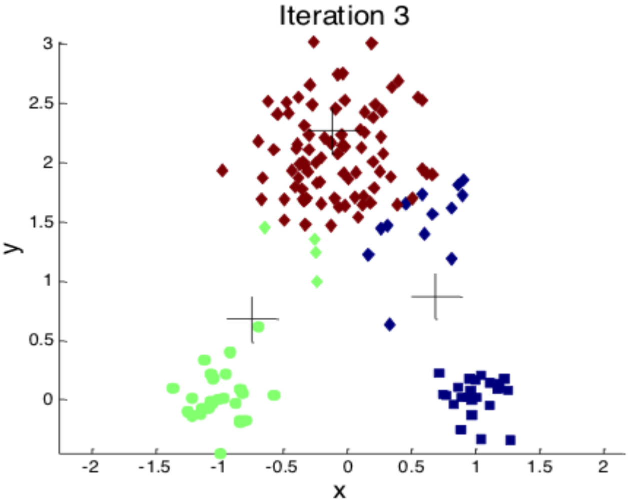
Example, Iterations



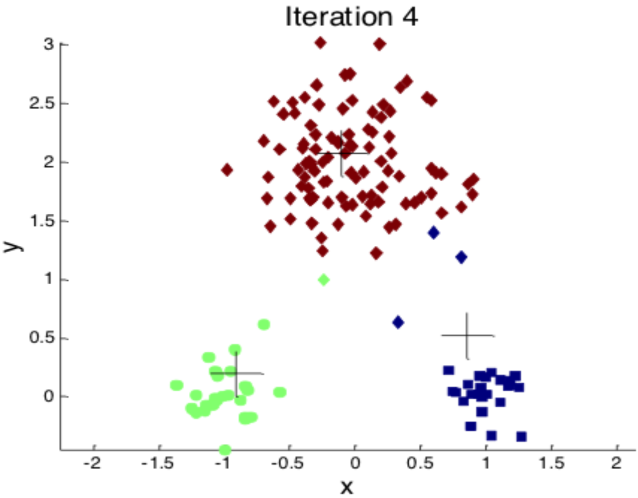
Example, Iterations



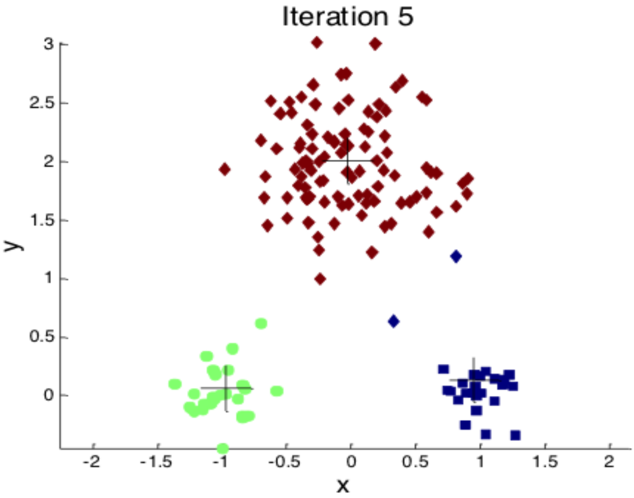
Example, Iterations



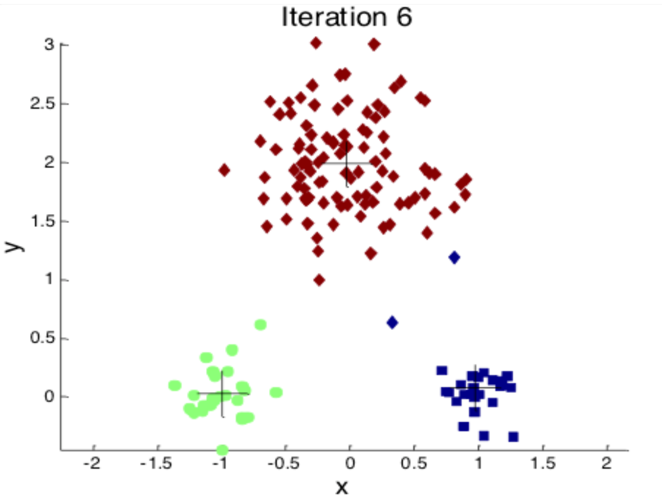
Example, Iterations



Example, Iterations



Example, Iterations



k-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.

k-means Clustering – Details

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.

k-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Nearest' is based on proximity/similarity/distance metric.

k-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Nearest' is based on proximity/similarity/distance metric.
- K-means will converge for common similarity measures mentioned above.
  - Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to 'Until relatively few points change clusters'

k-means, Pros and Cons

- Exclusive, deterministic, partitioning, batch clustering method

k-means, Pros and Cons

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

- Exclusive, deterministic, partitioning, batch clustering method

Strengths:

- relatively efficient:

  $O(ndki)$, where

  - n is number of *instances*,
  - d is number of *attributes*,
  - k is number of *clusters*,
  - i is number of *iterations*; normally $k, i << n$

- Unfortunately we know the value of i in advance

k-means, Pros and Cons

Weaknesses:

- Need to specify *k* in advance

k-means, Pros and Cons

Weaknesses:

- Need to specify k in advance
- “Mean” is ill-defined for nominal or categorical attributes

k-means, Pros and Cons

Weaknesses:

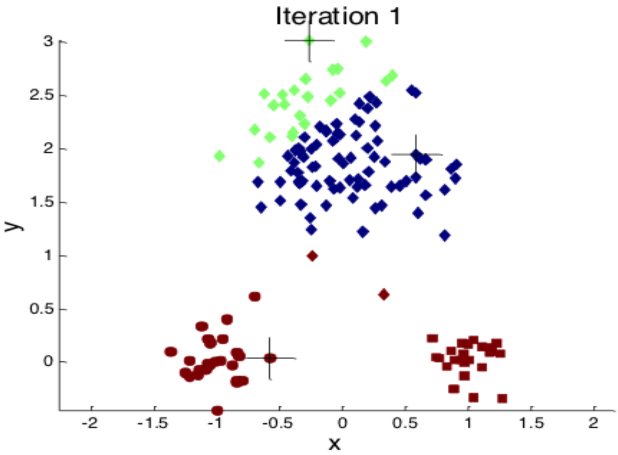
- Need to specify *k* in advance
- "Mean" is ill-defined for nominal or categorical attributes
- May not work well when the data contains outliers

k-means, Pros and Cons

Weaknesses:

- Need to specify *k* in advance
- “Mean” is ill-defined for nominal or categorical attributes
- May not work well when the data contains outliers
- Tends to converge to local minimum; sensitive to seed instances

Example, Impact of initial seeds



Clustering

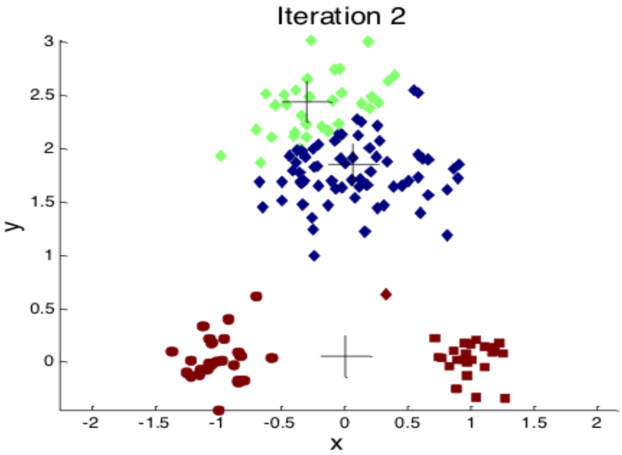
COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

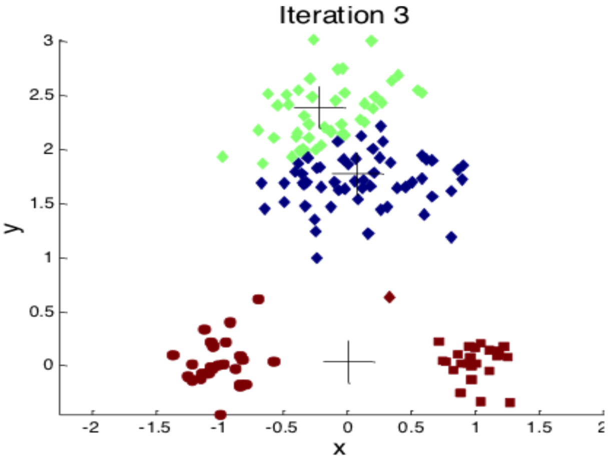
Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

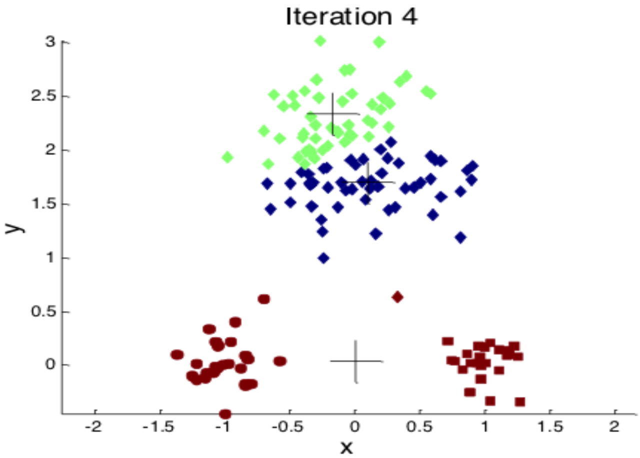
Example, Impact of initial seeds



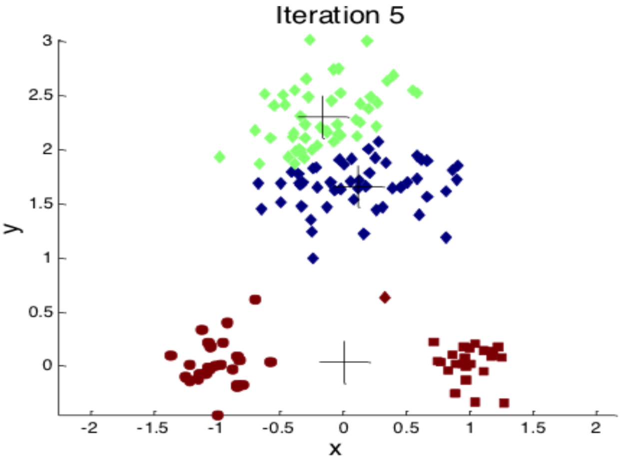
Example, Impact of initial seeds



Example, Impact of initial seeds



Example, Impact of initial seeds

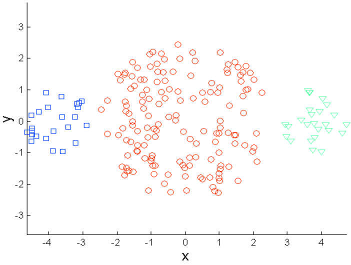


Other limitations *of k*-means

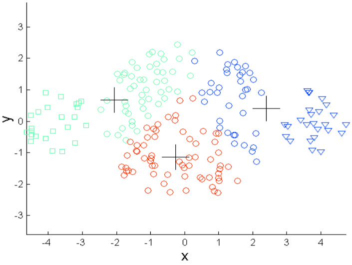
- K-means has problems when clusters are different in
  - Sizes
  - Densities

Other limitations *of k*-means

Clusters with different sizes:



Intended clusters



K-means (3 Clusters)

Other limitations of *k*-means

Clustering

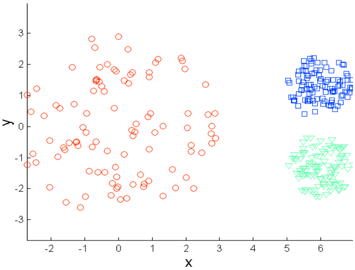
COMP90049
Knowledge
Technologies

Clustering
- Definition
- Types
- Evaluation

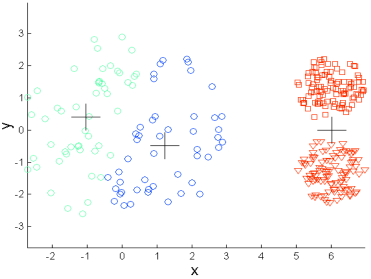
Methods
- Similarity
- k-Means
- k-Ms limitation
- Hierarchal

Summary
Resources

Clusters with different Densities:



Intended clusters



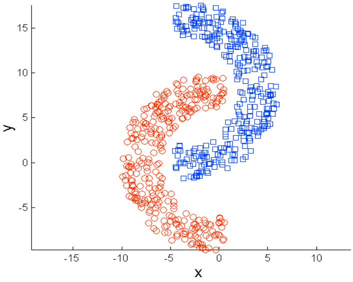
K-means (3 Clusters)

Other limitations *of k*-means

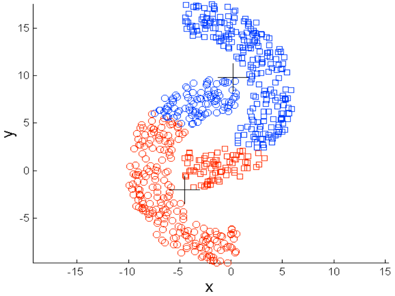
- K-means has problems when clusters are different in
  - Sizes
  - Densities
- K-means does not work well on non-globular shapes

Other limitations *of k*-means

Clusters with non-globular Shapes :

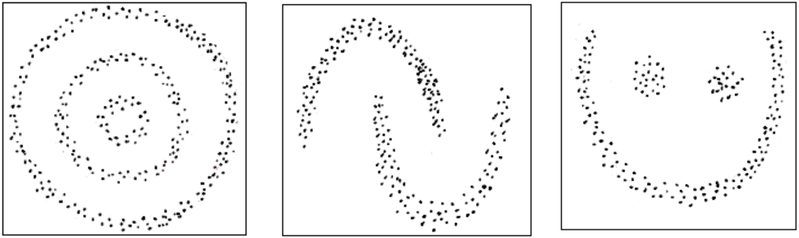


Intended clusters

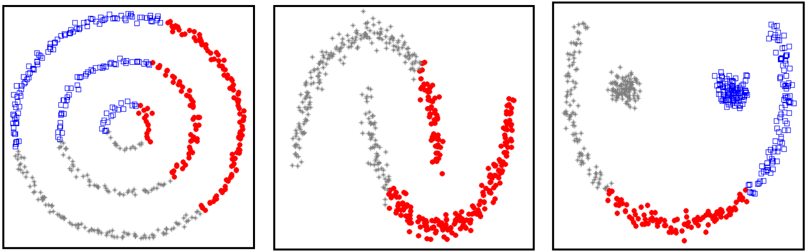


K-means (2 Clusters)

Other limitations *of k*-means



Other limitations *of k*-means



Other limitations *of k*-means

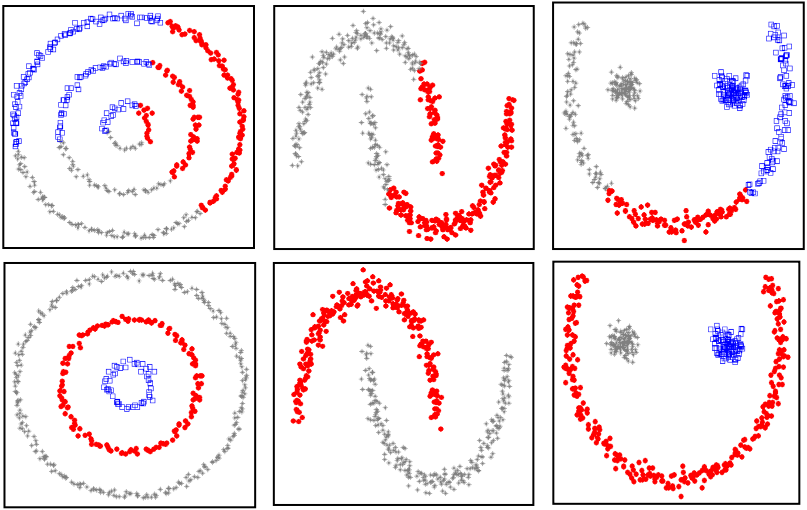
Clustering

COMP90049
Knowledge
Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources



Hierarchical Clustering

Bottom-up (= agglomerative) clustering

- Start with single-instance clusters
- At each step, join the two closest clusters (in terms of margin between clusters, distance between mean, ...)

Hierarchical Clustering

Bottom-up (= agglomerative) clustering

- Start with single-instance clusters
- At each step, join the two closest clusters (in terms of margin between clusters, distance between mean, ...)

Top-down (= divisive) clustering

- Start with one universal cluster
- Find two partitioning clusters
- Proceed recursively on each subset
- Can be very fast

Bottom-up (= agglomerative) clustering

- Start with single-instance clusters
- At each step, join the two closest clusters (in terms of margin between clusters, distance between mean, ...)

Top-down (= divisive) clustering

- Start with one universal cluster
- Find two partitioning clusters
- Proceed recursively on each subset
- Can be very fast

➢ In contrast to *k* -means clustering, hierarchical clustering only requires a measure of similarity between *groups* of data points (no seeds, no *k* value).

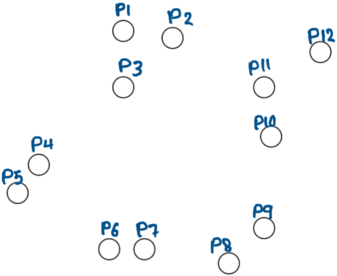
Compute the proximity matrix, if necessary.

repeat

- Merge the closest two clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

until Only one cluster remains

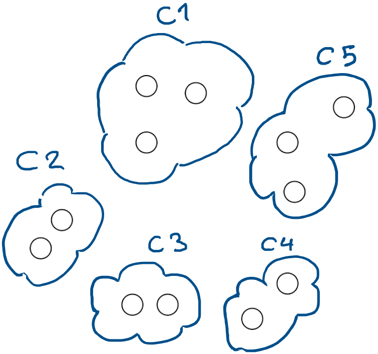
Example, Step 1



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

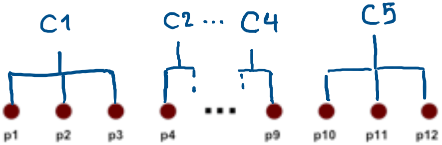
Proximity Matrix

Example, Step 2

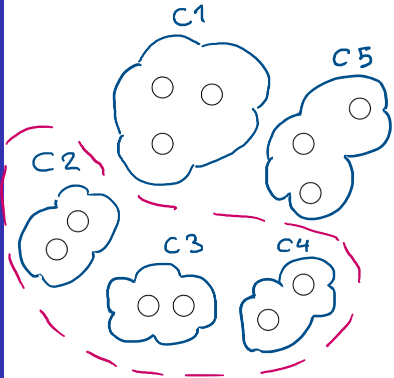


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

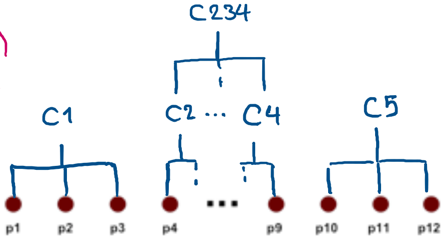
Proximity Matrix



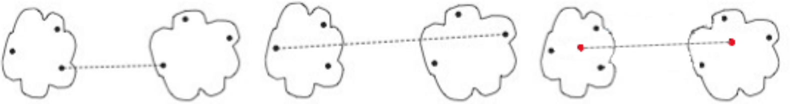
Example, Step 3



	c1	c234	c5
c1			
c234			
c5			



Graph-based measure of Proximity



(a) MIN (single link.) (b) MAX (complete link.) (c) Centroid

Updating the proximity matrix:

- **Single Link:** *Minimum* distance between any two points in the two clusters. (most similar members)
- **Complete Link:** *Maximum* distance between any two points in the two clusters. (most dissimilar members)
- **Centroid:** Distance between the centroids of each cluster

Agglomerative Clustering Example

	1	2	3	4	5
1	-	0.90	0.10	0.75	0.20
2	0.90	-	0.80	0.60	0.50
3	0.10	0.80	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

Based on this similarity matrix, what are the two closest points?

Agglomerative Clustering Example

	1	2	3	4	5
1	-	0.90	0.10	0.75	0.20
2	0.90	-	0.80	0.60	0.50
3	0.10	0.80	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

Based on this similarity matrix, what are the two closest points?

Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Update Single Link: update the approximate matrix based on the most similar members

	1	2	3	4	5
1	-	0.90	0.10	0.75	0.20
2	0.90	-	0.80	0.60	0.50
3	0.10	0.80	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

	12	3	4	5
12				
3				
4				
5				

Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Update Single Link: update the approximate matrix based on the most similar members

	1	2	3	4	5
1	-	0.90	**0.10**	0.75	0.20
2	0.90	-	**0.80**	0.60	0.50
3	**0.10**	**0.80**	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

	12	3	4	5
12				
3				
4				
5				

Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Update Single Link: update the approximate matrix based on the most similar members

	1	2	3	4	5
1	-	0.90	**0.10**	0.75	0.20
2	0.90	-	**0.80**	0.60	0.50
3	**0.10**	**0.80**	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

	12	3	4	5
12	-	**0.80**		
3	**0.80**	-		
4			-	
5				-

Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Update Single Link: update the approximate matrix based on the most similar members

	1	2	3	4	5
1	-	0.90	0.10	**0.75**	0.20
2	0.90	-	0.80	**0.60**	0.50
3	0.10	0.80	-	0.40	0.30
4	**0.75**	**0.60**	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

	12	3	4	5
12	-	0.80	**0.75**	
3	0.80	-	0.40	
4	**0.75**	0.40	-	
5				-

Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Update Single Link: update the approximate matrix based on the most similar members

	1	2	3	4	5
1	-	0.90	0.10	0.75	**0.20**
2	0.90	-	0.80	0.60	**0.50**
3	0.10	0.80	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	**0.20**	**0.50**	0.30	0.70	-

	12	3	4	5
12	-	0.80	0.75	**0.50**
3	0.80	-	0.40	0.30
4	0.75	0.40	-	0.70
5	**0.50**	0.30	0.70	-

Agglomerative Clustering Example

Merge points 12 & 3 into new cluster: 123

Using Single Link

	12	3	4	5
12	-	**0.80**	0.75	0.50
3	**0.80**	-	0.40	0.30
4	0.75	0.40	-	0.70
5	0.50	0.30	0.70	-

Agglomerative Clustering Example

Merge points 12 & 3 into new cluster: 123

Using Single Link

	12	3	4	5
12	-	0.80	0.75	0.50
3	0.80	-	0.40	0.30
4	0.75	0.40	-	0.70
5	0.50	0.30	0.70	-

	123	4	5
123	-	**0.75**	**0.50**
4	**0.75**	-	0.70
5	**0.50**	0.70	-

Agglomerative Clustering Example

Merge points 123 & 4 into new cluster: 12345

Using Single Link

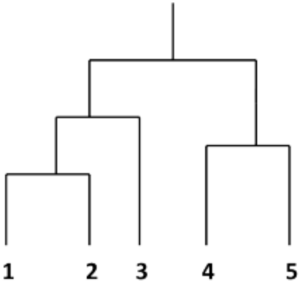
	1234	5
1234	-	0.70
5	0.70	-

Agglomerative Clustering Example

Merge points 123 & 4 into new cluster: 12345

Using Single Link

	1234	5
1234	-	0.70
5	0.70	-



Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Using Complete Link: update the approximate matrix based on the most dissimilar members

	1	2	3	4	5
1	-	0.90	0.10	0.75	0.20
2	0.90	-	0.80	0.60	0.50
3	0.10	0.80	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

	12	3	4	5
12				
3				
4				
5				

Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Using Complete Link: update the approximate matrix based on the most dissimilar members

	1	2	3	4	5
1	-	0.90	**0.10**	0.75	0.20
2	0.90	-	**0.80**	0.60	0.50
3	**0.10**	**0.80**	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

	12	3	4	5
12	-	**0.10**		
3	**0.10**	-		
4			-	
5				-

Agglomerative Clustering Example

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

Merge points 1 & 2 into a new cluster: 12

Using Complete Link: update the approximate matrix based on the most dissimilar members

	1	2	3	4	5
1	-	0.90	0.10	**0.75**	0.20
2	0.90	-	0.80	**0.60**	0.50
3	0.10	0.80	-	0.40	0.30
4	**0.75**	**0.60**	0.40	-	0.70
5	0.20	0.50	0.30	0.70	-

	12	3	4	5
12	-	0.10	**0.60**	
3	0.10	-	0.40	
4	**0.60**	0.40	-	
5				-

Agglomerative Clustering Example

Merge points 1 & 2 into a new cluster: 12

Using Complete Link: update the approximate matrix based on the most dissimilar members

	1	2	3	4	5
1	-	0.90	0.10	0.75	**0.20**
2	0.90	-	0.80	0.60	**0.50**
3	0.10	0.80	-	0.40	0.30
4	0.75	0.60	0.40	-	0.70
5	**0.20**	**0.50**	0.30	0.70	-

	12	3	4	5
12	-	0.10	0.60	**0.20**
3	0.10	-	0.40	0.30
4	0.60	0.40	-	0.70
5	**0.20**	0.30	0.70	-

Agglomerative Clustering Example

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

Merge points 4 & 5 into new cluster: 45

Using Complete Link

	12	3	4	5
12	-	0.10	0.60	0.20
3	0.10	-	0.40	0.30
4	0.60	0.40	-	**0.70**
5	0.20	0.30	**0.70**	-

Agglomerative Clustering Example

Clustering

COMP90049 Knowledge Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

Merge points 4 & 5 into new cluster: 45

Using Complete Link

	12	3	4	5
12	-	0.10	0.60	0.20
3	0.10	-	0.40	0.30
4	0.60	0.40	-	0.70
5	0.20	0.30	0.70	-

	12	3	45
12			
3			
45			

Agglomerative Clustering Example

Merge points 4 & 5 into new cluster: 45

Using Complete Link

	12	3	4	5
12	-	0.10	0.60	0.20
3	0.10	-	**0.40**	**0.30**
4	0.60	**0.40**	-	0.70
5	0.20	**0.30**	0.70	-

	12	3	45
12	-	0.10	0.20
3	0.10	-	**0.30**
45	0.20	**0.30**	-

Agglomerative Clustering Example

Merge points 45 & 3 into new cluster: 345

Using Complete Link

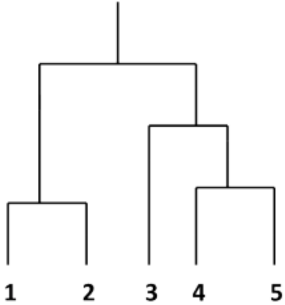
	12	345
12	-	0.10
345	0.10	-

Agglomerative Clustering Example

Merge points 45 & 3 into new cluster: 345

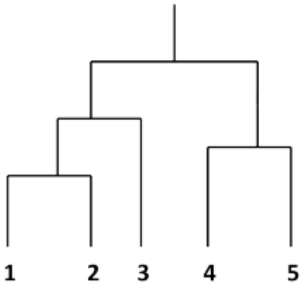
Using Complete Link

	12	345
12	-	0.10
345	0.10	-

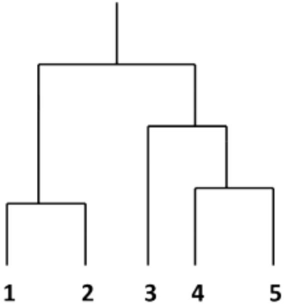


Agglomerative Clustering Example

Single Link



Complete Link



Clustering vs Classification

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

Clustering vs Classification

Clustering

COMP90049
Knowledge
Technologies

Clustering
Definition
Types
Evaluation

Methods
Similarity
k-Means
k-Ms limitation
Hierarchal

Summary
Resources

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

Summary

- What basic contrasts are there in different clustering methods?
- How does *k* -means operate, and what are its strengths and weaknesses?
- How to evaluate clusters
- What are some challenges we face when clustering data?

Resources

Tan, Steinbach, Kumar (2006) Introduction to Data Mining. Chapter 8, Cluster Analysis
http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Jain, Dubes (1988) Algorithms for Clustering Data.
http://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf

MID-SEMESTER EXAM REVIEW

	Q1	Q2	Q3	Q4	Q5	Q6
Average	1.41	1.70	0.68	1.28	1.25	0.68
Median	1.5	2	0.75	1.5	1.5	1
SD	0.52	0.55	0.30	0.85	0.69	0.41