Lecture 12: Clustering

Sarah Erfani and Karin Verspoor, CIS

Semest

THE UNIVERSITY OF MELBOURNE

Example clusters for the weather dataset

Outlook	Temperature	Humidity	Windy	Play
				no
				no
				yes
				yes
				yes
rainy	cool	n		
rainy	cool	n		o
:	:	:	:	:

A possible clustering of the weather dataset

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

Clustering

-
-
-
-
- Applications in pattern recognition, s
  diagnosis, ...

Clustering, basic contrasts

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

- Partial vs. complete
  - In some cases, we only want to cluster
- Heterogeneous vs. homogeneous
  - Clusters of widely different sizes, sh
- Incremental vs. batch clustering
  - Is the whole set of items clustered in one go?

Exclusive vs. overlapping clustering

Deterministic vs. probabilistic clustering

Instance	*Cluster*
5	2
6	2
7	4
⋮	⋮

Instance	Cluster 1	2	3	4

Hierarchical vs. partitioning clustering

Clustering, Desiderata

-
-
-
- Able to deal with noise and outliers
- Insensitive to order of input records

What is a good clustering?

Two clusters?

Four clusters?

Six clusters?

Types of Evaluation

Unsupervised.

- Measures the goodness of a clustering structure without respect to external information. Includes measures of cluster cohesion (compactness, tightness), and measures of cluster separation

- Measures the extent to which the clust
  a clustering algorithm matches som
  instance, *entropy* can measure h
  externally supplied class labels.

Relative.

- Compares different clusterings or clusters (using an unsupervised or supervised measure for the purpose of comparison).

Evaluating clusters mathematically

Most common measure is Sum of Squared Error (SSE) or *Scatter*

- For each point, the error is the distance to the nearest cluster
-
-
- Can show that the m_i that minim mean) of the cluster
- Given two clusters, we can choose the
- 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

Similarity / Proximity / Closeness

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

- Data points in Euclidean space
  - Euclidean distance
  -
  -
  -

b	1	0	1
c	1	1	0

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

- Documents
  - Cosine similarity
  - Jaccard measure
- Other measures
  - Correlation
  - Graph-based measures

k-means Clustering

Given k, the *k*-means algorithm is implemented in four steps:

1.
2.
3. st centroid
4. Recompute the centroid of each cluster
5. **until** the centroids don't change

- Exclusive, deterministic, partitioni

Example, Iterations

k-means Clustering – Details

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

ned

- Most of the convergence happens i
- Often the stopping condition is chan
  'Until relatively few points change cl
  (this way the stopping criterion will no
  or dimensionality)

Example, Impact of initial seeds

Example, Different outcomes

k-means, Pros and Cons

Strengths:

- relatively efficient:
  - $O(ndki)$, where n is no. instances, d is no. attributes, k is no. clusters, and i is no. iterations; normally $k, i \ll n$

- tends to converge to local minimum; s
  try multiple iterations with different s
- need to specify k in advance
- not able to handle non-convex clusters, or clusters of differing densities or sizes
- "mean" ill-defined for nominal or categorical attributes
- may not work well when the data contains outliers

Hierarchical Clustering

Bottom-up (= agglomerative) clustering

- Start with single-instance clusters
- At each step, join the two closest clusters (in terms of margin

- 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).

Agglomerative Clustering

4 e

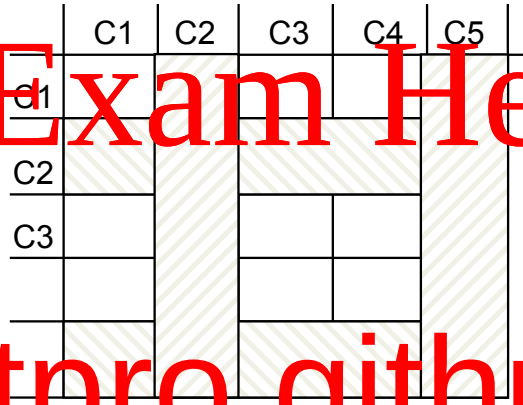
new cluster and the original clusters

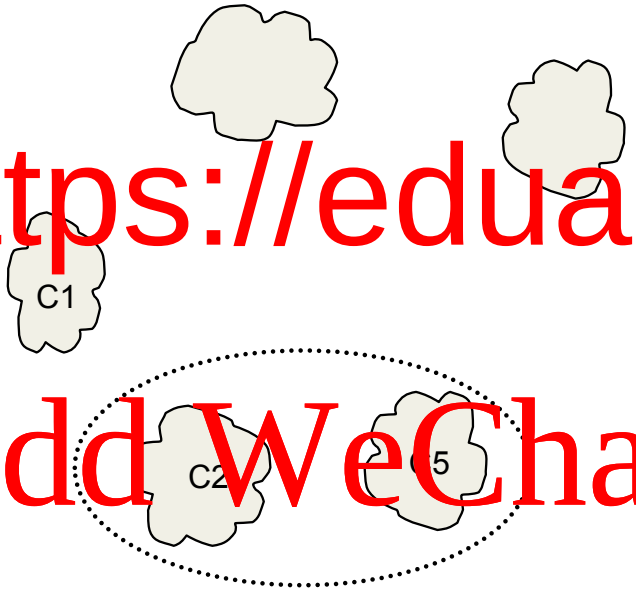
5 **until** Only one cluster remains

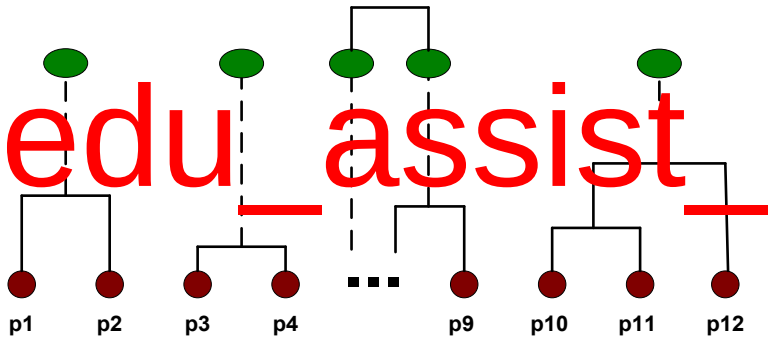
Example, Step 1

Example, Step 2

Example, Step 3







Graph-based measure of Proximity

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)
- Group Average: *Average* distance between all points (pairwise).

Agglomerative Clustering Example

	1	2	3	4	5
5	0.20	0.50	0.30	0.80	1.00

What are the two closest points?

Agglomerative Clustering Example

	1	2	3	4	5
1	1.00	0.90	0.10	0.65	0.20
2	0.90	1.00	0.70	0.60	0.50
3	0.10	0.70	1.00	0.40	0.30
4	0.65	0.60	0.40	1.00	0.80

Update (single link):

	1	2	3	4	5	6
6	—	—	0.70	0.65	0.50	1.00

Update (complete link):

	1	2	3	4	5	6
6	—	—	0.10	0.60	0.20	1.00

	1	2	3	4	5
1	1.00	0.90	0.10	0.65	0.20
2	0.90	1.00	0.70	0.60	0.50
3	0.10	0.70	1.00	0.40	0.30
4	0.65	0.60	0.40	1.00	0.80

Single link

Complete link

- Clustering is in the eyes of the beholder

- Algorithms for Clustering Data (198

`http://homepages.inf.ed.ac` `ering_`
`Jain_Dubes.pdf`

Summary

- What basic contrasts are there in different clustering methods?
- How does *k*-means operate, and what are its strengths and weaknesses?

Resources:

Tan, Steinbach, Kumar (2006) Introduction to 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`