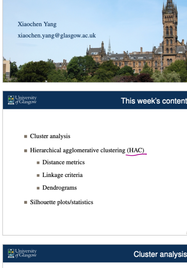
slide8

Friday, July 2, 2021 4:24 PM

TutorialSlide8



This week's content

- Cluster analysis
- Hierarchical agglomerative clustering (HAC)
  - Distance metrics
  - Linkage criteria
  - Dendrograms
- Silhouette plots/statistics

Cluster analysis

- An unsupervised learning technique (does not have class labels)
- Aim: Partition data into groups such as observations in the same group (**clusters**) are similar and observations in different groups (**clusters**) are dissimilar.
- Clustering algorithms
  - Hierarchical cluster analysis: HAC
  - Partitioning cluster analysis: *K*-means and *K*-medoids clustering
  - Parametric / model-based clustering: Gaussian mixture models (supplementary)

Hierarchical agglomerative clustering (HAC)

- All observations start by being their own clusters.
- Calculate the dissimilarity between all the current clusters.
- Merge the two clusters with the smallest dissimilarity into one new cluster.
- Iterate steps 2 and 3 until there is only one cluster.

Hierarchical agglomerative clustering

- All observations start by being their own clusters.
- Calculate the dissimilarity between all the current clusters.
- Merge the two clusters with the smallest dissimilarity into one new cluster.
- Iterate steps 2 and 3 until there is only one cluster.

Distance metrics, used for dissimilarity between observations:

- Euclidean distance, Manhattan distance, etc.

Linkage criteria, used for dissimilarities between clusters:

- complete linkage, single linkage, average linkage, etc.

Tutorial Q1

HAC

Perform hierarchical agglomerative clustering on the following data set using the Euclidean distance and complete linkage.

Solved

Obs.	X_1	X_2	X_3
1	3	4	2
2	2	2	3
3	4	4	1
4	1	4	4
5	2	1	4

Tutorial Q1

- All observations start by being their own clusters.
- Calculate the dissimilarity between all the current clusters.

Compute the Euclidean distance between each pair of observations:

Obs.	X_1	X_2	X_3
1	3	4	2
2	2	2	3
3	4	4	1
4	1	4	4
5	2	1	4

	1	2	3	4	5
1	0	-	-	-	-
2	2.45	0	-	-	-
3	1.41	3.66	0	-	-
4	2.83	2.45	4.24	0	-
5	3.74	1.41	4.69	3.16	0

Symmetric

$$d_E(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(3-2)^2 + (4-2)^2 + (2-3)^2} \approx 2.45.$$

Tutorial Q1: 1st merge

- Merge the two clusters with the smallest dissimilarity into one new cluster.
- Calculate the dissimilarity between all the current clusters.

	1	2	3	4	5
1	0	-	-	-	-
2	2.45	0	-	-	-
3	1.41	3.66	0	-	-
4	2.83	2.45	4.24	0	-
5	3.74	1.41	4.69	3.16	0

min

	1&3	2&5	4
1&3	0	-	-
2&5	4.69	0	-
4	4.24	3.16	0

Complete linkage: the maximum of the distances between all pairs of points with one point from cluster A and one point from cluster B

Tutorial Q1: 2nd and 3rd merge

Second merge:

	1&3	2&5&4
1&3	0	-
2&5&4	4.69	0

Third merge:

	1&3&2&5&4
1&3&2&5&4	0

Dendrogram

- 1st merge: merge 1&3 and 2&5 at height 1.41
- 2nd merge: merge 2&5&4 at height 3.16
- 3rd merge: merge all points at height 4.69

Cut the dendrogram

- Set the number of clusters

Figure: HAC with 2 clusters.

Cut the dendrogram

- Set the number of clusters
- Set a specific height

Figure: Dendrogram cut at height 2

Silhouette width

Silhouette width: How well is the cluster? which to choose?

$$s_i = (b_i - a_i) / \max(a_i, b_i),$$

- a_i: average distance between observation i and all other observations in the same cluster how compact the current cluster is?
- b_i: minimum average distance between observation i and observations in other clusters how far current pt to other clusters?
- $s_i \in [-1, 1]$, with a high value indicates that the observation is well matched to its own cluster and poorly matched to neighbouring clusters

Silhouette plots

Select the number of clusters

HAC in R

- `scale`: standardise the data
- `hclust(dist(Data), method=c("complete","single","centroid",...)`
- `as.dendrogram(hclust)`
- `cutree(hclust, h=NULL, k=NULL)` return the vector of cluster belongings

`dendextend` package

- `color.branches(dend, h=NULL, k=NULL)`

`cluster` package

- `silhouette(cluster,allocation,dist)`

`factoextra` package

- `fviz_nbclust(Data, FUN=hcut, method="silhouette")` select the # of clusters