CIS 490 Machine Learning

Lecture 15

Instructor: (Julia) Hua Fang

1

Last Time

- Unsupervised Learning: **Clustering**
  - K-means: you are expected to know
    - ✓ Distance measure
    - ✓ Objective function
    - ✓ Algorithm, how to choose K clusters and application areas
    - ✓ K-means issues and remedies
  - Run K-means in R Studio
- Final Project Proposal

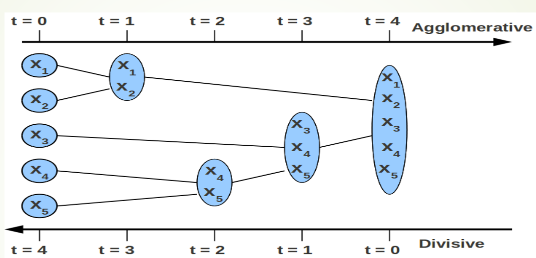
2

Unsupervised Learning: **Clustering**

- Unsupervised Learning: **Clustering**
  - Hierarchical clustering:
    - Types of hierarchical clustering
    - Similarity measures for hierarchical clustering
    - **Dendrogram**: How to interpret and derive Dendrogram.
    - Hierarchical clustering Algorithm
  - Run K-means in R Studio
- Written Homework 2

3

Hierarchical Clustering: Two types



- Agglomerative (bottom-up) Cluster
- Divisive (top-down) Clustering

4

Hierarchical Clustering: Two types

- Agglomerative (bottom-up) Cluster
  - Start with each example in its own singleton cluster
  - At each time-step, greedily merge 2 most similar clusters
  - Stop when there is a single cluster of all examples, else go to 2
- Divisive (top-down) Clustering
  - Start with all examples in the same cluster
  - At each time-step, remove the "outsiders" from the least cohesive cluster Stop when each example is in its own singleton cluster, else go to 2

Agglomerative is more popular and simpler than divisive (but less accurate)

5

Hierarchical Clustering: (Dis)similarity measures

6

Hierarchical Clustering: (Dis)similarity between clusters

How to compute the dissimilarity between two clusters R and S?

- Min-link or single-link: results in chaining (clusters can get very large)

$$d(R,S) = \min_{\mathbf{x}_R \in R, \mathbf{x}_S \in S} d(\mathbf{x}_R, \mathbf{x}_S)$$

The minimum distance between data points of each cluster

- Max-link or complete-link: results in small, round shaped clusters

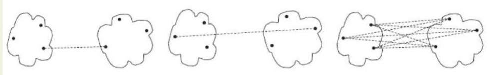
$$d(R,S) = \max_{\mathbf{x}_R \in R, \mathbf{x}_S \in S} d(\mathbf{x}_R, \mathbf{x}_S)$$

The maximum distance between data points of each cluster

- Average-link: compromise between single and complete linkage

$$d(R,S) = \frac{1}{|R||S|} \sum_{\mathbf{x}_R \in R, \mathbf{x}_S \in S} d(\mathbf{x}_R, \mathbf{x}_S)$$

The mean distance between data points of each cluster



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

- Ward's method (Ward): minimum variance criterion; minimizes the total within-cluster variance

7

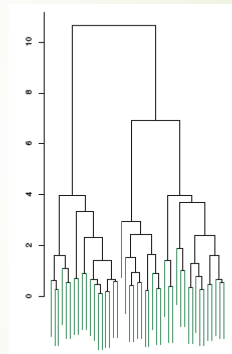
Dendrogram

(key component in hierarchical clustering)

8

Dendrogram

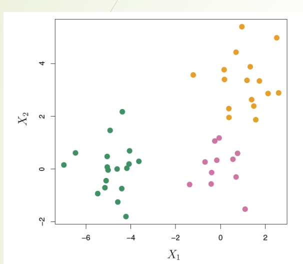
- A tree-like visual representation of the observations, called a *Dendrogram*, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to *n*.

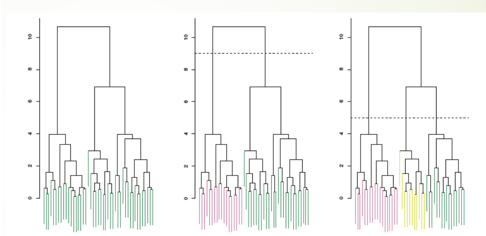


9

How to interpret a Dendrogram

- simulated data: 45 observations from a three-cluster model; treat these cluster labels in distinct colors as unknown





Left Center Right

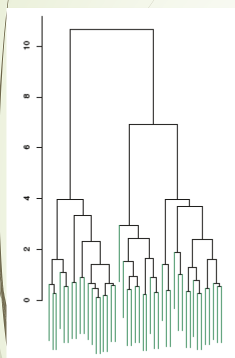
Center: cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.

Right: now cut at a height of 5. This cut results in three distinct clusters, shown in different colors.

Note: the colors are simply used for display purposes in this figure.

10

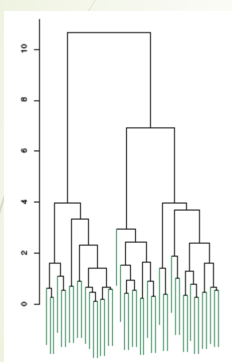
How to interpret a Dendrogram



- Each leaf of the dendrogram (called "singleton"): one of the 45 observations.
- Move up the tree, some leaves begin to fuse into branches: observations similar to each other.
- As move higher up the tree, branches themselves fuse, either with leaves or other branches.

11

How to interpret a Dendrogram



What implications can you expect?

Implications

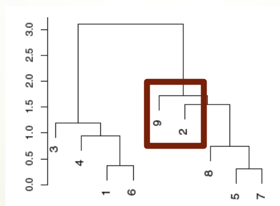
- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other.
- The height of this fusion, as measured on the vertical axis, indicates how different the two observations are.

12

Illustrative Example: In-class exercise

Questions:

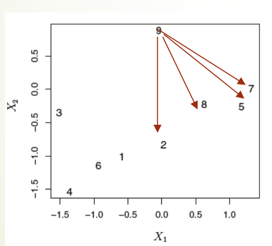
- Observations 5 and 7 are quite similar to each other, as are observations 1 and 6? Yes
- Observations 9 and 2 are quite similar to each other on the basis that they are located near each other on the dendrogram? No

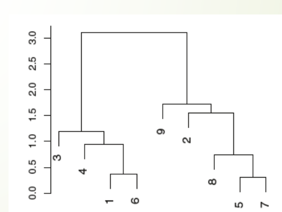


13

Illustrative Example: In-class exercise

observation 9 is no more similar to observation 2 than it is to observations 8, 5, and 7





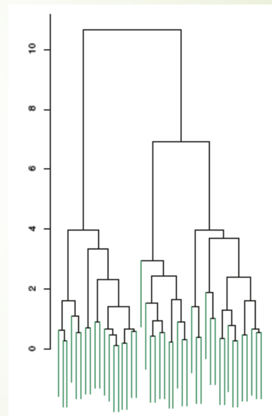
14

Property of Dendrogram

- All agglomerative possess a monotonicity property:

  ie. the higher the level of merger, the more different/dissimilar between merged clusters.

  - **the height of each node** is proportional to the value of the intergroup dissimilarity between its two daughters



15

Notes for Interpreting Dendrogram

- Different hierarchical methods, as well as small changes in the data, can lead to quite different dendrograms.
- Hierarchical methods **impose** hierarchical structure whether or not such structure actually exists in the data.
- Do not draw conclusions about the similarity of two observations based on their proximity along the horizontal axis.
  - draw conclusions about the similarity of two observations based on the location on the **vertical axis** where branches containing those two observations first are fused.
  - Groups that merge at high values are candidates for natural clusters.

16

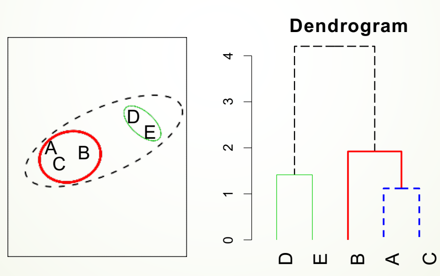
Hierarchical Clustering: Algorithm

17

Hierarchical Clustering: Algorithm

The approach in words:

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.



39 / 52

18

Algorithm: Bottom-up Hierarchical Clustering

Simplified version :

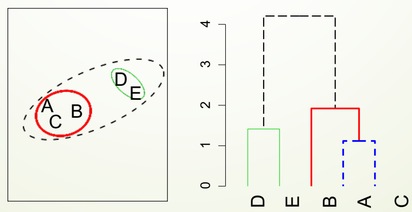
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.



19

Algorithm: Example of Complete Linkage (Max-link) Clustering

- The table is an example of a distance matrix.
  - Only the lower triangle is shown, because a distance matrix will be symmetric and the upper triangle can be filled in by reflection.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	$\boxed{2}$	8	0

Note: these are IDs

20

Example of Complete Linkage (Max-link) Clustering

Start clustering: The smallest distance is between 3 and 5, so they get linked up or merged first into a the cluster '35'.

- Eg., d(1,3)= 3 and d(1,5)=11. using complete linkage clustering, pick D(1,"35")=11. This gives us the new distance matrix.

Then: the items with the smallest distance get clustered next. This will be 2 and 4.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	$\boxed{2}$	8	0

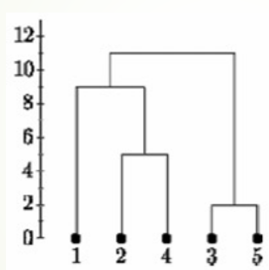
	35	1	2	4
$\boxed{35}$	0			
1	11	0		
2	10	9	0	
4	9	6	5	0

Note: 35 means Case 3 and Case 5, not a number!

21

Example of Complete Linkage (Max-link) Clustering

- Continuing in this way, after 6 steps, everything is clustered. This is summarized in Dentrogram.

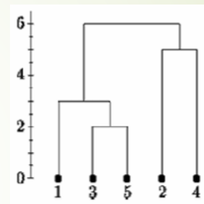


- the y-axis shows the cluster height: the distance between the objects at the time they were clustered.
- Different visualizations use different measures of cluster height.

22

Example of Single linkage (Min-link) Clustering

- single linkage dendrogram for the same distance matrix.
  - It starts with cluster "35" but the distance between "35" and each item is now the minimum of d(x,3) and d(x,5). So c(1,"35")=3.



	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	$\boxed{2}$	8	0

Note: these are IDs

	35	1	2	4
$\boxed{35}$	0			
1	3	0		
2	10	9	0	
4	9	6	5	0

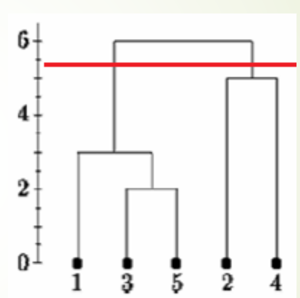
Note: 35 means Case 3 and Case 5, not a number!

23

Example of Single linkage (Min-link) Clustering

- cut the single linkage tree at the point shown below:

  two clusters.

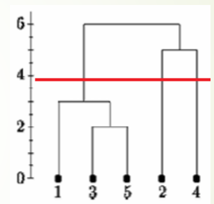


24

Example of Single linkage (Min-link) Clustering

- How about cutting the tree lower?

one cluster and two singletons.



25

Hierarchical clustering: How to choose number of clusters?

26

Brief intro to Gap statistic

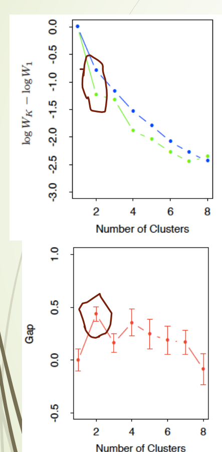
The recently proposed Gap statistic (Tibshirani et al., 2001b) compares the curve $\log W_K$ (W_K: the within cluster dissimilarity)

- Gap statistic estimates the optimal number of clusters to be the place where the gap between the two curves (expected vs. observed) is largest.

Tibshirani, R., Walther, G. and Hastie, T. (2001b). Estimating the number of clusters in a dataset via the gap statistic, Journal of the Royal Statistical Society, Series B. 32 (2): 411–423.

27

Brief intro to Gap statistic



The graph shows the result of the Gap statistic applied to simulated data.

- The top panel shows $\log W_K$ for $k = 1, 2, \ldots, 8$ clusters (green curve) and the expected value of $\log W_K$ over 20 simulations from uniform data (blue curve).
- The bottom panel shows the gap curve, which is the expected curve minus the observed curve. Shown also are error bars of half-width

$$s'_K = s_K\sqrt{1+1/20},$$

where s_K is the standard deviation of $\log W_K$ over the 20 simulations.
- The Gap curve is maximized at $K = 2$ clusters.

28

Suppl.: Other criteria

- Generally use visual criteria, e.g. silhouette plots. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

  e.g. The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

  - If most objects have a high value, then the clustering configuration is appropriate.
  - If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

29

Suppl.: Other criteria

- Other numerical criteria: cophenetic correlation (a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points)

  (Others: Dunn's validity index, Hubert's gamma, G2/G3 coefficient, adjusted Rand index, etc.)

Note: Compare to other clustering methods gives an idea of the stability of the cluster solution

30

31

R: Running Hierarchical clustering in R studio

R: Run Hierarchical clustering in Rstudio using USArrest data

Let's go through the instruction file "R_Hierarchical_S22.docx" posted with LS15 slides at myCourses.

31

Group Homework 2

- Due April 13th : 75 Points
- Posted in the "Homework (HW)" folder at myCourses.

Summary of your group meet time and duration need to be included in your HW on the last page

In person or Zoom:

Group meet time and duration (e.g., 5pm-7pm, Feb 1st):

Average time in communication and discussion regarding assigned group work (via email or other social media, e.g. What's app.):

Participants (Print and sign your names):

Contribution report need to be included in your HW:

If your team members contribute equally to this project, please make this statement "Each member contributes equally" on your last page, so that each of you will receive the same score.

If your team members do not contribute equally to this project, please note your team members' names, and mark the percentage of effort each member makes (e.g., Sukumar: 80% then if your group receives a project score of 30, then this member with 80% effort will only get 24).

Participants: Print and sign your names

32