Intro to Clustering

Machine Learning

Clustering

- Clustering is an Unsupervised Learning Technique
- A Cluster: collection of objects that are **similar**
- Objective is to group similar data points into a group
    - Segmenting customers into similar groups
    - Automatically organizing similar files/emails into folders
- Simplifies data by reducing many data points into a few clusters

- Some specific applications
  - Image processing : used to cluster of pixels representing objects in each frame. The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame. Successive frames are examined to identify any changes to the clusters. These newly identified clusters may indicate unauthorized access to a facility.
  - Medical : Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters under various health conditions
  - Customer segmentation : Cluster customers on basis of frequency of purchase, recency of purchase, value of purchase and look for common attributes among high value customers. Target all potential customers who have similar attributes

Distance

- Do define “similarity” you need a measure of distance
- Examples of common distance measures
  - Manhattan Distance
  - Eucledian Distance
  - Chebyshev Distance

More Distance Measures

- Mahalanobis distance

- Jaccard distance

Importance of Scaling

greatlearning
Learning for Life

	Cust_ID	Name	Avg_Mthly_Spend	No_Of_Visits	Apparel_Items	FnV_Items	Staples_Items
1	1	A	10000	2	1	1	0
2	2	B	7000	3	0	10	9
3	3	C	7000	7	1	3	4
4	4	D	6500	5	1	1	4
5	5	E	6000	6	0	12	3
6	6	F	4000	3	0	1	8
7	7	G	2500	5	0	11	2
8	8	H	2500	3	0	1	1
9	9	I	2000	2	0	2	2
10	10	J	1000	4	0	1	7

- Most distance measures are highly influenced by the scale of each variable.
- Variables with large scales have a much greater influence over the distance.
- Hence all measurements are converted to the same scale. For example z-scores.

Distance measures

- Choice of distance measures play a key role in cluster analysis.
- Knowledge of the distribution of data (gaussian or otherwise) will help
- Are the various attributes independent or influence each other
- Are their outliers in the data on the various dimensions
- Though Euclidian distance is the most commonly used distance metric, it has three main features that should be kept in view
  - It is highly scale dependent. Changing the units of one variable can have a huge influence on the results. Hence standardizing the dimensions is a good practice
  - It completely ignores the relationship between measurements
  - It is sensitive to outliers. If the data has outliers that cannot be handled or removed, use of Manhattan distance is preferred

Types of Clustering

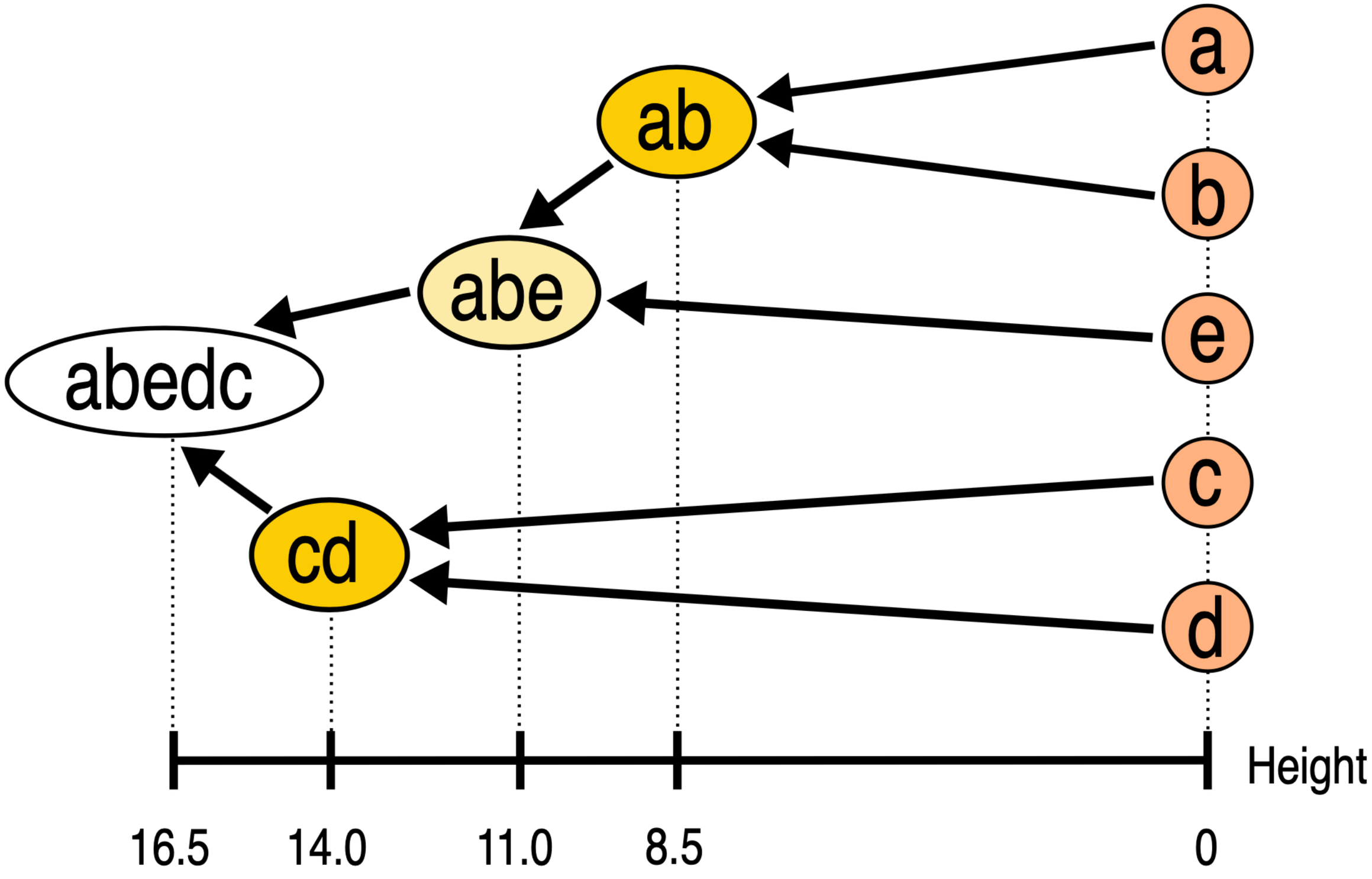
1. Connectivity based clustering (Hierarchical clustering): based on the idea that related objects are closer to each other. Can we then create a hierarchy of clusters/groups.

    - Useful when you want flexibility in how many clusters you ultimately want. For example, imagine grouping items on an online marketplace like Etsy or Amazon.

    - In terms of outputs from the algorithm, in addition to cluster assignments you also build a nice tree **(dendrogram)** that tells you about the hierarchies between the clusters. You can then pick the number of clusters you want from this tree.

    - In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis.

    - Algorithms can be agglomerative (start with 1 object and aggregate them into clusters) or divisive (start with complete data and divide into partitions).



Types of Clustering

greatlearning
Learning for Life

2. Centroid based clustering (Eg. K- Means clustering): The objective is to find K clusters/groups. The way these groups are defined is by creating a centroid for each group. The centroids are like the heart of the cluster, they “capture” the points closest to them and add them to the cluster.

    - Large K produces smaller groups and a small K produces larger groups
    - K-Means uses Eucledian distances and is the most popular
    - Other variants like K-medians and K-mediods use other distance measures