Clustering: HAC
Hierarchical Agglomerative Clustering (HAC)

- Assumes a *distance function* for determining the similarity of two instances.
  - One can also assume a similarity function, and reverse some of the operations (e.g., minimum distance -> maximum similarity) in the algorithm to make it work for similarities

- Starts with each object in a separate cluster and then repeatedly joins the two closest clusters until only one cluster is left
HAC Example
HAC Algorithm

Start with all objects in their own cluster.
Repeat until there is only one cluster:
   Among the current clusters, determine the two clusters, $c_i$ and $c_j$, that are closest
   Replace $c_i$ and $c_j$ with a single cluster $c_i \cup c_j$

**Problem**: we assume a distance/similarity function that computes distance/similarity between examples, but here we also need to compute distance between clusters.

How?
Distance Between Clusters

- Assume a distance function that determines the distance of two objects: $D(x, x')$.
- There are multiple ways to define a cluster distance function:
  - Single Link: distance of two closest members of clusters
  - Complete Link: distance of two furthest members of clusters
  - Average Link: average distance
Single Link Agglomerative Clustering

- Use minimum distance of all pairs:
  \[ D(C_i, C_j) = \min_{x \in c_i, x' \in c_j} D(x, x') \]
Single-link’s chaining effect

- Single link is famous for its chaining effect
- It can gradually adds more and more examples to the “chain”
- Create long straggling clusters
Complete Link

- Maximum distance of all pairs:
  \[ D(C_i, C_j) = \max_{x \in C_i, x' \in C_j} D(x, x') \]

- Makes “tight,” spherical clusters
Complete link is outlier sensitive

**Figure 17.4** Outliers in complete-link clustering. The five documents have the x-coordinates $1 + 2\varepsilon, 4, 5 + 2\varepsilon, 6$ and $7 - \varepsilon$. Complete-link clustering creates the two clusters shown as ellipses. The most intuitive two-cluster clustering is $\{\{d_1\}, \{d_2, d_3, d_4, d_5\}\}$, but in complete-link clustering, the outlier $d_1$ splits $\{d_2, d_3, d_4, d_5\}$ as shown.
Updating the Cluster Distances after merging is a piece of

- After merging $c_i$ and $c_j$, the distance of the resulting cluster to any other cluster, $c_k$, can be computed by:
  
  - Single Link:
    \[
    D((c_i \cup c_j), c_k) = \min(D(c_i, c_k), D(c_j, c_k))
    \]
  
  - Complete Link:
    \[
    D((c_i \cup c_j), c_k) = \max(D(c_i, c_k), D(c_j, c_k))
    \]

- This is constant time given previous distances
Average Link

• Basic idea: average similarity between members of the two clusters

• Averaged across all ordered pairs in the merged cluster

\[ D(c_i, c_j) = \frac{1}{|c_i \cup c_j|(|c_i \cup c_j| - 1)} \sum_{x \in (c_i \cup c_j)} \sum_{x' \in (c_i \cup c_j); x' \neq x} D(x, x') \]

• Compared to single link and complete link:
  – Computationally more expensive – naively it can be \( O(n^2) \) to compute the new distance between a pair of clusters
  – If we use cosine similarity, we can compute the new similarity in constant time
  – Achieves a compromise between single and complete link
HAC creates a Dendrogram

- Dendrogram draws the tree such that the height of a tree branch = the distance between the two merged clusters at that particular step.
- The distances are always monotonically increasing.
How many clusters

- Cutting the dendrogram at a specific similarity level gives us a simple flat clustering
- Different cutting places lead different numbers of clusters
- One option is to cut the dendrogram where the gap between two successive combination distances is largest
Single Link
Complete Link

K=2

K=3

K=4
Average Link

![Diagram of Average Link]
K-means
Comments on HAC

- HAC is a convenient tool that often provides interesting views of a dataset
- Primarily HAC can be viewed as an intuitively appealing clustering procedure for data analysis/exploration
- We can create clusterings of different granularity by stopping at different levels of the dendrogram
- HAC often used together with visualization of the dendrogram to decide how many clusters exist in the data
- Different linkage methods (single, complete and average) often lead to different solutions