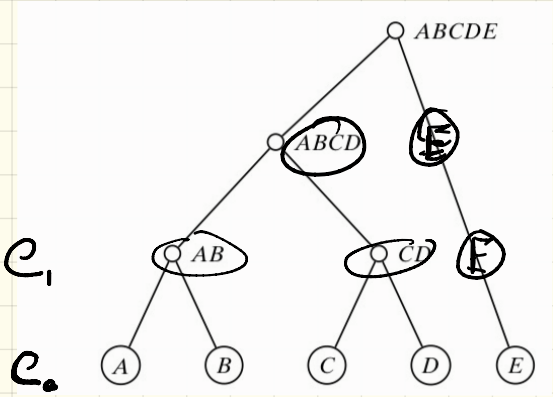
Hierarchical Clustering

Hierarchical Clustering problem:

Given n points in $\mathbb{R}^d$
create a seq
of **nested** partitions

Often! seq is viz as a **dendrogram**



Course grained
(one cluster w/ all pts)

trivial
Clustering

(fine grained clstering)
one point per cluster

2 types of algos for Hieral clustering

- divisive algos (top down)
- agglomerative (bottom up)

Benifits:

- Viz tool
- find a "meaningful" level
- given k find k-clusters

Recall

Given dataset $\mathbb{D} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ $\vec{x}_i \in \mathbb{R}^d$

Clustering is a partition of $\mathbb{D}$

Given a clustering $\mathcal{A} = \{A_1, A_2, \ldots, A_r\}$ is nested in clustering $\mathcal{B} = \{B_1, B_2, \ldots, B_s\}$ iff

$\forall A_i \in \mathcal{A} \; \exists B_j \in \mathcal{B} \ni A_i \subseteq B_j$

Hierarchical Clustering output

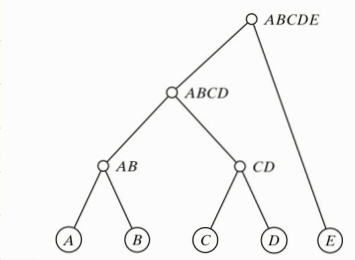
Seq of clusterings $\left(C_1 = \{\{\vec{x}_1\}, \{\vec{x}_2\}, \ldots, \{\vec{x}_n\}\}, C_2, \ldots C_n = \{\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}\} \right)$

w/ C_{i-1} is nested in C_i

$C_2, \ldots$ depends on algo & data

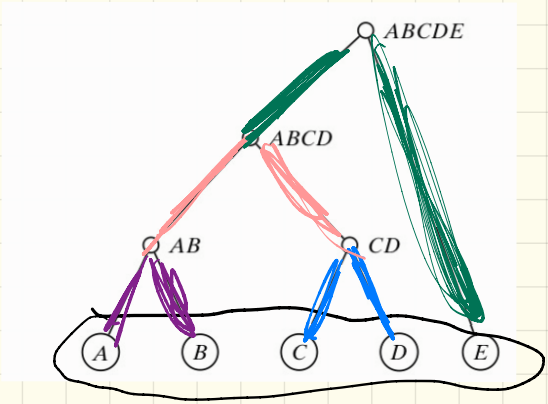
Viz w/ dendrogram

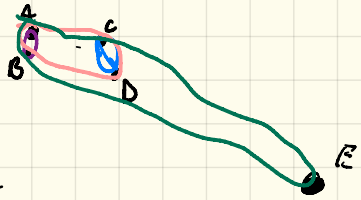
Clustering	Clusters
C_1	$\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$
C_2	$\{AB\}, \{C\}, \{D\}, \{E\}$
C_3	$\{AB\}, \{CD\}, \{E\}$
C_4	$\{ABCD\}, \{E\}$
C_5	$\{ABCDE\}$



Agglomarative Hierarchical Clustering

High level
- start w/ n clusts
  each w/ 1 point
- at each step
  merge 2 clusters
  (closest centers, closest pair, closest-furthest pair)
  based on "closest"
- repeat w/ 1 less cluster





Sanity check question
How many iterations
- start w/ n clusts
- each iter we remove 1 cluster
  ⇒ n-1 iteration

If we want k clusts, stop at step n-k

Algorithm 14.1: Agglomerative Hierarchical Clustering Algorithm

AGGLOMERATIVECLUSTERING($\mathbf{D}, k$):

$\mathcal{C} \leftarrow \{C_i = \{\mathbf{x}_i\} \mid \mathbf{x}_i \in \mathbf{D}\}$ // Each point in separate cluster
$\mathbf{\Delta} \leftarrow \{\|\mathbf{x}_i - \mathbf{x}_j\| : \mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}\}$ // Compute distance matrix
repeat
  Find the closest pair of clusters $C_i, C_j \in \mathcal{C}$
  $C_{ij} \leftarrow C_i \cup C_j$ // Merge the clusters
  $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$ // Update the clustering
  Update distance matrix $\mathbf{\Delta}$ to reflect new clustering
until $|\mathcal{C}| = k$

Distances between clusters

L_2 dist of $\vec{x}, \vec{y} \in \mathbb{R}^d$ $\quad \|\vec{x}-\vec{y}\| = \left(\sum_{i=1}^{d} (x_i - y_i)^2\right)^{1/2}$

Given clusters C_i & C_j

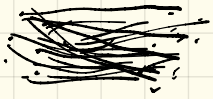
Single Link - closest pair

$$\delta(C_i, C_j) = \min \left\{ \|\vec{x}-\vec{y}\| \;\middle|\; \vec{x} \in C_i \text{ & } \vec{y} \in C_j \right\}$$

Mean distance - dist between centers

μ_i is center of C_i
μ_j is center of C_j

$$\delta(C_i, C_j) = \|\mu_i - \mu_j\|$$



Group avg - all pairs

$n_i = |C_i|$
$n_j = |C_j|$

$$\delta(C_i, C_j) = \frac{\sum_{\vec{x} \in C_i} \sum_{\vec{y} \in C_j} \|\vec{x}-\vec{y}\|}{n_i * n_j}$$

Min Variance (Ward's method)

- increas in the sum of squared errors when two clusters merged together

Recall Cluster C_i w/ center μ_i

$$SSE(C_i) = \sum_{\vec{x} \in C_i} \|\vec{x} - \mu_i\|^2$$

given clustering $\mathcal{C} = \{C_1, C_2, \dots, C_m\}$

$\exists$ μ_i is the center of C_i

$$SSE(\mathcal{C}) = \sum_{i=1}^{m} SSE(C_i) = \sum_{i=1}^{m} \sum_{\vec{x} \in C_i} \|\vec{x} - \mu_i\|^2$$

Wards method

$$S(C_i, C_j) = SSE(\{C_i \cup C_j\}) - SSE(C_i) - SSE(C_j)$$

after some alg

let $n_i = |C_i|$

$n_j = |C_j|$

μ_i: center of C_i

μ_j center of C_j

$$S(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \|\mu_i - \mu_j\|^2$$

weighted sq dist between individual cluster means and ½ of harmonic mean of cluster sizes

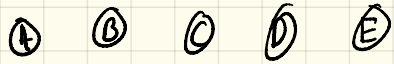
$$M_h(a,b) = \frac{2}{\frac{1}{a} + \frac{1}{b}} = \frac{2ab}{a+b}$$

Eg w/ single link

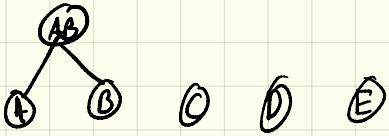
	B	C	D	E
A	1	3	2	4
B		3	2	3
C			1	3
D				5

↓ find smallest

	B	C	D	E
A	1	3	2	4
B		3	2	3
C			1	3
D				5

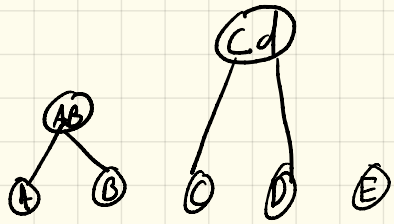


↓ update dist matrix

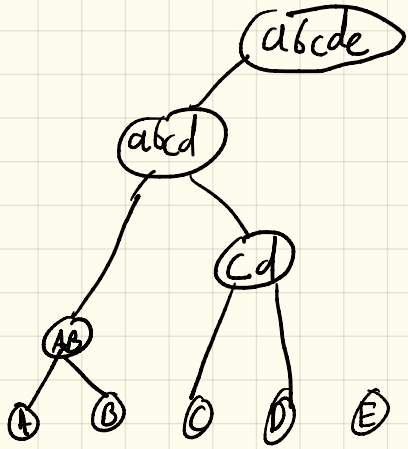


	C	D	E
AB	3	2	3
C		1	3
D			5

↓ updat dist matrix



	CD	E
AB	2	3
CD		3



Update C_i & $C_j \rightarrow C_{ij}$ update dist to all other C_r

Lance-Williams formulation

$\forall$ cluster C_r

$$s(C_{ij}, C_r) = \alpha_i s(C_i, C_r) + \alpha_j s(C_j, C_r) + \beta s(C_i, C_j) + \gamma \left| s(C_i, C_r) - s(C_j, C_r) \right|$$

$\alpha_i, \alpha_j, \beta, \gamma$ defined in table

Measure	α_i	α_j	β	γ
Single link	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Group average	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Mean distance	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i \cdot n_j}{(n_i+n_j)^2}$	0
Ward's measure	$\frac{n_i+n_r}{n_i+n_j+n_r}$	$\frac{n_j+n_r}{n_i+n_j+n_r}$	$\frac{-n_r}{n_i+n_j+n_r}$	0