lecture

October 1, 2025

Definition 1 (Clustering)**.** given a dataset $S = \{x_1, x_2, \ldots, x_n\}$, find "similar" points

$G(V, E, w)$

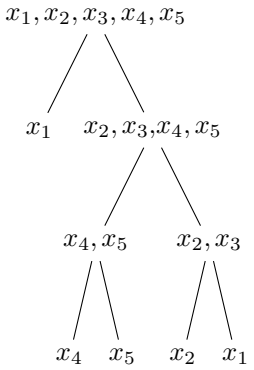
V is verticies, E is edges

w is the "weight"

"size of cut" talks about how much weight is removed

examples: clustering time series, K-center, median, mean

given an arbitrary metric space, minimize distance to centers w k (fixed) centers,



example cost function:

$$\sum_{i,j \in V} w_{ij} : \frac{\#\text{ of data points present when node i,j split}}{n} \tag{1}$$

linkage algos: single, average, complete

divisive (top-down), balanced, sparsest

example: $G(V, E, w \geq 0)$ $|V| = n$

find a binary tree that minimizes cost

$$\sum_{ij,\epsilon E} w_{ij} \, |\# \text{ leaves of } T_{ij}| \tag{2}$$

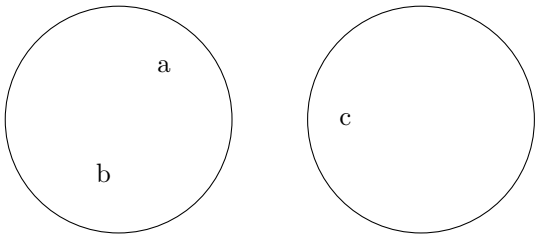
T_{ij} = is also called *lowest common ancestor* of i, j

tree reconstruction problem, use vector reconstruction

1 Convex Relaxations

assign a $\{0, 1\}$ variable for each pair of nodes, x_{ij}

want to represent that we want to look for tree with 0 1 variables, and also represent the objective with 01



$x_{a,b} = 1$ $x_{a,c} = 0$

$$\max_{\text{assigned } \vec{x}} \sum_{ij \epsilon E} w_{ij} x_{ij} \tag{3}$$

this system maximizes the "weight cut" in the original cluster finding problem