ENGR 102
Review / Exercises

Dr. Mehmet Ercan Nergiz

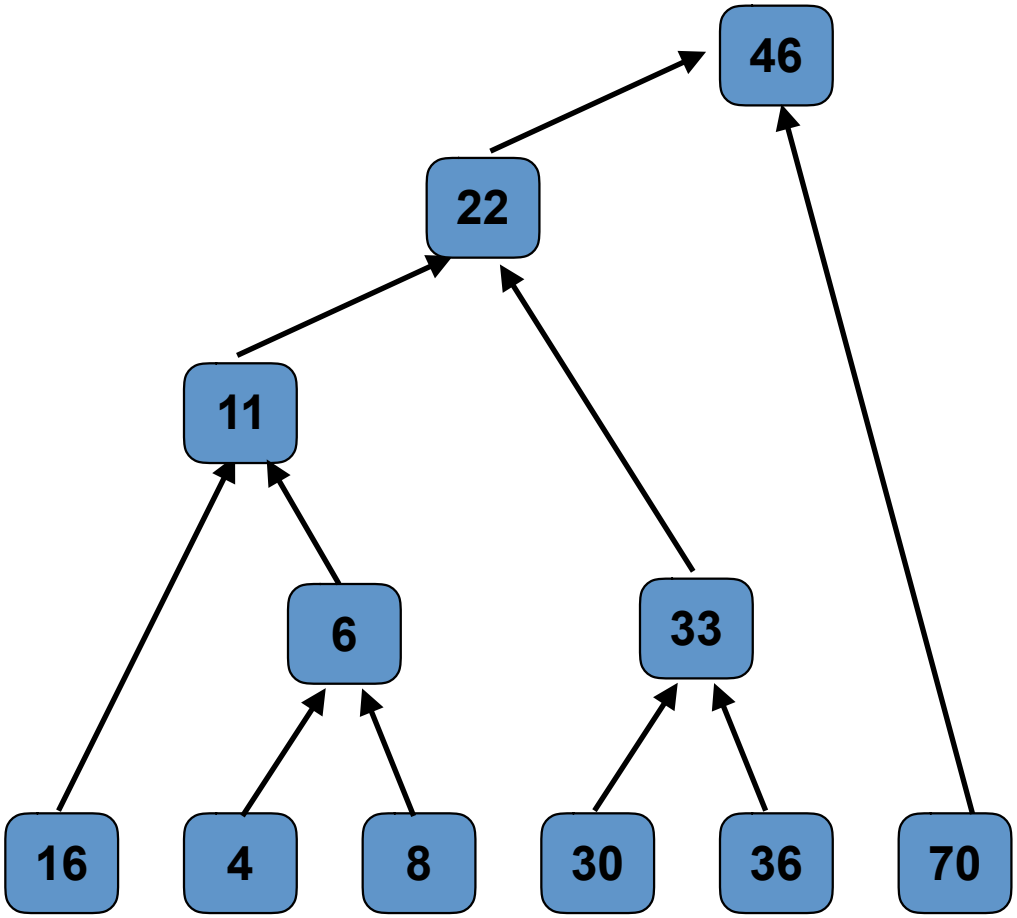
Hierarchical Clustering

- Cluster the following set of numbers assuming
  - hierarchical clustering algorithm
  - Euclidean distance function
  - and a parent cluster vector is calculated by averaging child cluster vectors.
- Draw the corresponding dendogram.

4,8,30,16,34,70

Hierarchical Clustering

4,8,30,16,36,70



Exercise

- Write a function that
  - inputs two clusters
  - connects them into a new parent cluster
  - and returns the parent cluster.
- Write a function that
  - inputs a list of numbers
  - clusters the numbers as above
  - and returns a single cluster representing an in-memory dendogram.

```
class Cluster:
  def __init__(self, left, right, v):
      self.left = left
      self.right = right
      self.v = v
```

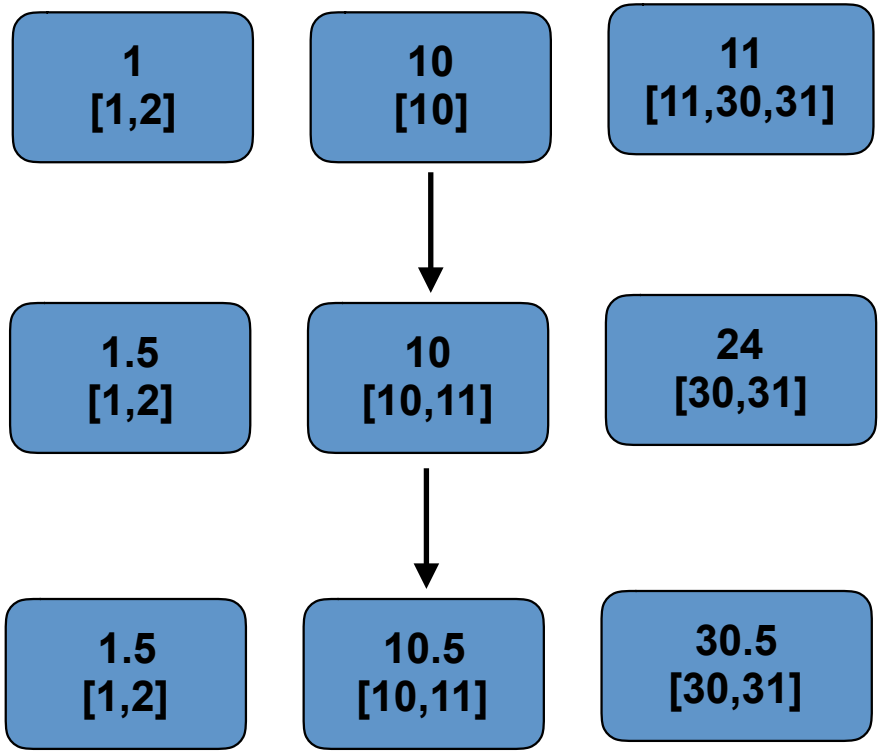
Note: left and right are references to child clusters. v is the value representing the cluster (and any cluster under it.)

k-means Clustering

- We start with *k* clusters, each with a cluster representative (value).
- We distribute each data point *p* to the cluster with representative closest to *p*.
- After distribution, we update each representative as the average of all points within it.
- Start all over again, re-apply distribution multiple times with new representatives.

3-means Clustering

1,10,11,30,31,2




```
class Cluster:
  def __init__(self, rep):
      self.rep = float(rep)
      self.points = []
```

Note: rep is a float and represents cluster. points is the list of points inside the cluster.

Beautiful Soup

- Write functions that inputs an html text and
  - returns the number of tags containing some string data.
  - returns the number of tags containing a target word.
  - returns the number of times a word appears inside of a tag <a ….>

Searching, Ranking

- Write a function that
  - inputs a list of text (containing space separated words) and a target word
  - returns the set of indices for text containing the target word.
  - E.g. [“aa bb cc”, “aa bb bb”, “dd aa”, “bb bb”], target: “bb” should return [0,1,3]

- Extend the above function such that the set of indices are sorted w.r.t. the following criteria:
  - the number of times the word appears in text (more is better)
  - the number of words in text (fewer is better)
  - E.g., return [3,1,0]

Searching, Ranking

- Write a function that
  - inputs a list of football game scores and a target team
  - returns all the game scores of the team.
  - E.g. [“FB 6 - GS 0”, “FB 1 - BJK ”, “BJK 0 - GS 1”], target: “FB”
    should return [“FB 6 - GS 0”, “FB 1 - BJK ”]

- Extend the above function such that the set of indices are sorted w.r.t. the criteria used in the project:

Classification

- Given a set of labeled training points *TP* and unlabeled point *p*, a k-NN classifier predicts the label of *p* as follows:
  - calculate the set of points *CP* (*CP*⊆*TP*) of size *k*, that is closest to *p* in *TP*.
    - e.g., closeness can be defined in terms of euclidian distance.
  - return the most frequent label in *CP*.
- E.g., *TP*=[(0,0,"BAD"),(1,1,"BAD"),(2,2,"BAD") (10,10,"GOOD"),(11,11,"GOOD")], *p*=(9,9,?), k=3
  - *CP*=[(2,2,"BAD")(10,10,"GOOD"),(11,11,"GOOD")]
  - the most frequent label is GOOD.
  - return GOOD.