Lecture 23: Clustering
Clustering

Can we find $k$ students to hold candy bowls so that each 6.046 student is within $r$ seats from a candy bowl?
Widely used tool in:

- Websites, search results, click stream data, customer behavior patterns, community discovery in social networks, photographs, music, movies, genes and proteins, word behaviors, biological species, .... (lot's more!)
What is clustering?

• The scenario:
  – Given many objects

• The goal:
  – Divide into groups so that
    • Objects in same group are close
    • Objects in different groups are farther apart

• But:
  – What are the objects?
  – What is “close” and “far”? (notion of distance?)
What is the goal?

• How would you cluster this?

• Types:
  – Purple – finds connected objects (computer vision)
  – Green – all elements in same group are close
Agglomerative clustering

• Try to maximize spacing between clusters
  – Want “really close” points in same cluster
  – Clusters can be large!

• Algorithm idea:
  – “grow” a graph on the nodes such that connected components correspond to clusters.
  – Bring nearby points into same cluster quickly.

• Seems like a good idea to consider points in order of distance!
Agglomerative clustering details

• See notes!
Min-Radius clustering

• Clusters correspond to “balls” in $\mathbb{R}^n$
Min-radius clustering details

• See notes!
Set Cover

- Input: $k, S_1, S_2, \ldots, S_n$ such that $S_i \subseteq P$
- Output: Is there $S_{i_1}, \ldots, S_{i_k}$ such that $\bigcup_{j=1}^{k} S_{i_j} = P$?

But Set cover is NP-complete!

- That’s ok, this reduction is in the wrong direction!
- But, min radius clustering is also NP-complete!
- Well, good news -- this reduction is in the right direction for something!
O(log n) -Approximation algorithm for set-cover

• If there is a set-cover of size $k$
• Outputs set-cover of size at most $k \log n$
More on approximation algorithms for radius clustering (see notes)

• Get best radius, pretty good number of clusters
• Get best number of clusters, pretty good radius
Other clustering measures:

- Diameter
- Average distance to center
- Not requiring point in P to be center
- Not requiring all points to be clustered
- Graph based distances – min/sparse cuts
- And on and on....