Greedy clustering methods

Prof. Ramin Zabih

http://cs100r.cs.cornell.edu

Cornell University
Computer Science

Administrivia

- Guest lecture on Tuesday
  - Prof. Charlie van Loan, on ellipse fitting
- Prelim 3: Thursday 11/29 (last lecture)
  - Review session the day before
- A6 is due Friday 11/30 (LDOC)
- Final projects due Friday 12/7
- Course evals!

  http://www.engineering.cornell.edu/courseeval/

Clustering

- Generally speaking, the goal is to maximize similarity within a cluster, and to minimize similarity between clusters
  - We will assume that the space and the number of clusters are known
  - Usually (but not always) a safe assumption
    - Sometimes you only know distances between pairs of points
      - May not be embeddable in a space
  - Often assume a known number of clusters

Example

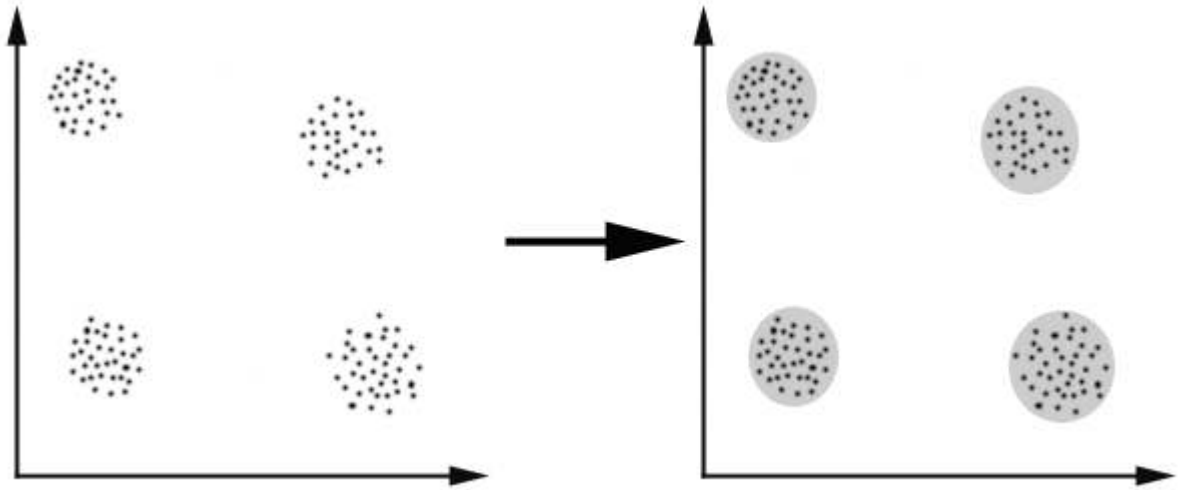

Figure from Johan Everts

Applications of clustering

- Economics or politics
  - Finding similar-minded or similar behaving groups of people (market segmentation)
  - Find stocks that behave similarly
- Spatial clustering
  - Earthquake centers cluster along faults
  - Find epidemics (famous example: cholera)
- Classify documents for web search
  - Automatic directory construction (like Yahoo!)
  - Find more documents "like this one"

Clustering algorithms

- There are many types of methods
- How is a cluster is represented?
  - Is a cluster represented by a data point, or by a point in the middle of the cluster?
  - This is similar to an argument we saw before…
- An interesting class of methods uses graph partitioning
  - Edge weights are distances
- There are many classes of algorithms

Greedy methods

- Many CS problems can be solved by repeatedly doing whatever seems best at the moment
  - I.e., without needing a long-term plan
- These are called greedy algorithms
- Example: hill climbing for convex function minimization
- Example: sorting by swapping out-of-order pairs

K-center clustering

- Find K cluster centers that minimize the maximum distance between any point and its nearest center
  - We want the worst point in the worst cluster to still be good (i.e., close to its center)
- This is a nice optimization problem
  - Given a clustering, we can describe how good it is, so we have a figure of merit
- But to find the optimal clustering we have to try every one!

A greedy method

- Pick a random point to start with, this is your first cluster center
- Find the farthest point from the cluster center, this is a new cluster center
- Find the farthest point from any cluster center and add it

Example

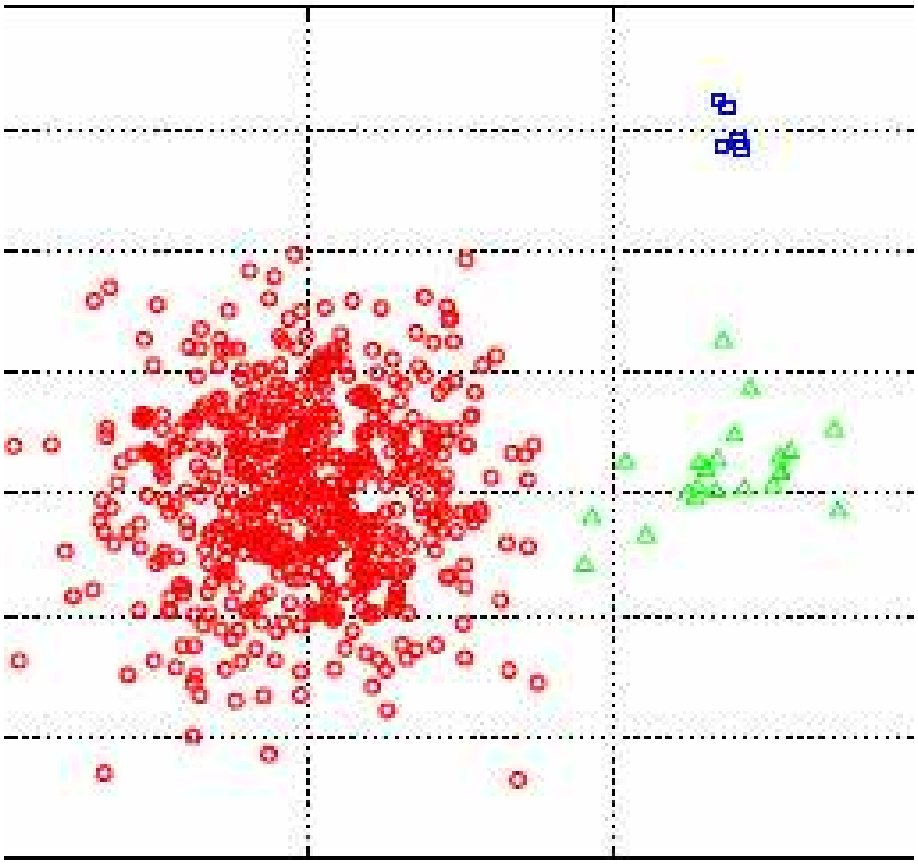

Figure from Jia Li

An amazing property

- This algorithm gives you a figure of merit that is no worse than twice the optimum
- Such results are very difficult to achieve, and the subject of much research
  - Mostly in CS681, a bit in CS482
  - You can't find the optimum, yet you can prove something about it!