k-Center Problems

Joey Durham

Graphs, Combinatorics and Convex Optimization Reading Group

Summer 2008

Outline

- General problem definition
- Several specific examples
  - k-Center, k-Means, k-Mediod
- Approximation methods
- Other methods
  - Lloyd algorithm
  - Annealing
- Summary of properties

General k-Center Problem

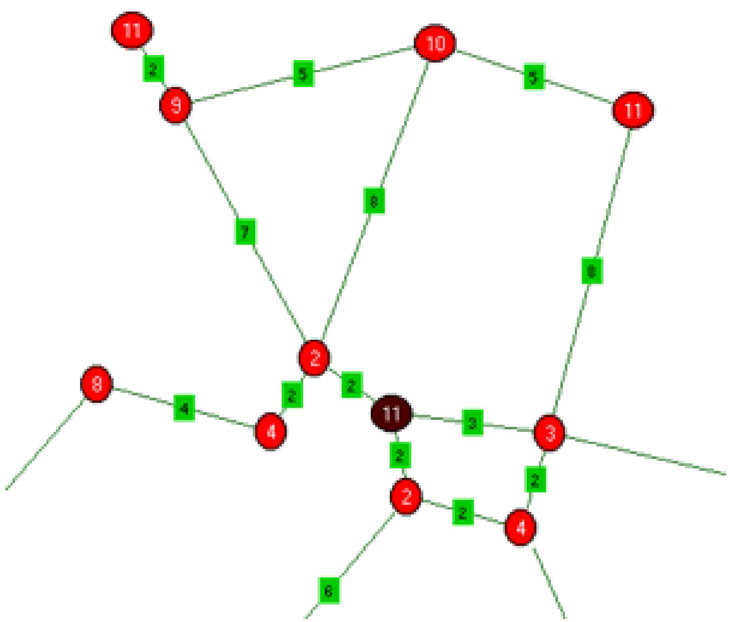


Image from www.graph-magics.com

- Given:
  - *n* in points in a vector space or a complete graph
  - Distance function satisfying the triangle inequality
- Find *k* “centroids” to minimize some measure of cluster size
- NP-hard

Applications

- Data clustering
- Statistical analysis
- Deployment
- Task allocation
- Image classification
- Facility location

Image from www.spatialanalysisonline.com

Variations on k-Center

- Centroids
  - Member of data set
  - Any point in vector space
- Cluster measures
  - Maximum distance => minimize worst case
  - Sum of distances => minimize expected distance
  - Sum of square distances => minimize variance
- Vertex weights
- Added centroid cost
  - Facility location problem

k-Means Clustering

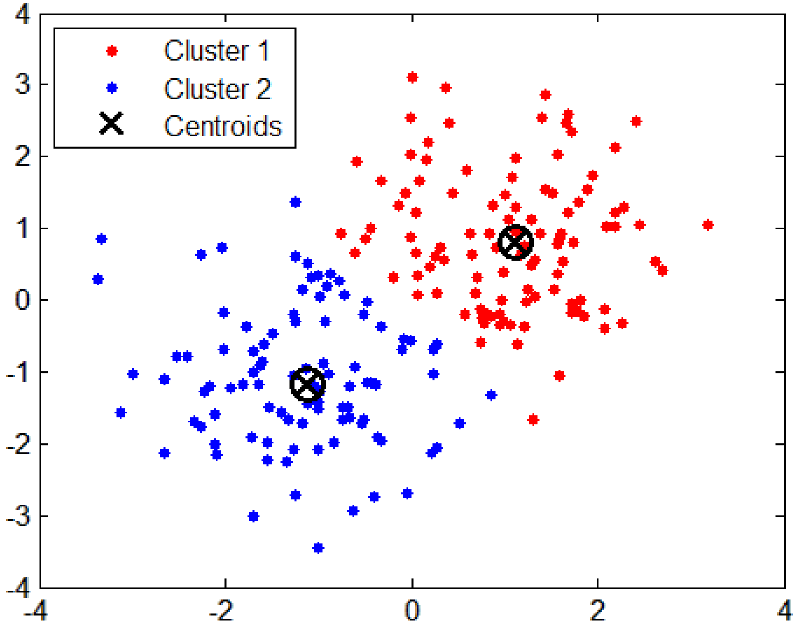
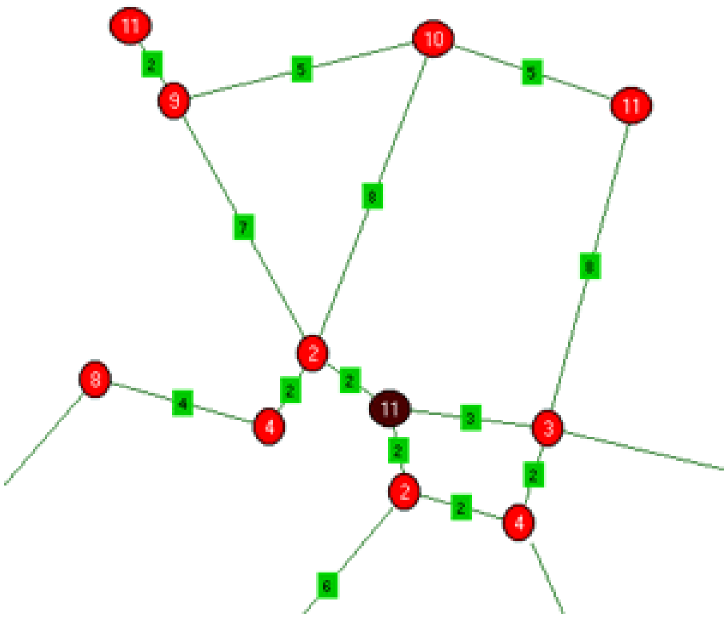


Image from www.mathworks.com

- Vector space, Euclidean distances
- Minimize intra-cluster variance
- Centroids NOT in data set
  - k-medoids: centroids in set
- The most famous: 21,000+ hits on Google Scholar
- Often used in data clustering/statistics
- Resources:
  - MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations";
  - http://www.autonlab.org/tutorials/kmeans.html

Standard k-Center



- Complete graph, edge costs satisfy tri. ineq.
- Minimize worst case distance of vertex to centroid
- Centroid in data set
  - Resources: Vazirani (2003), *Approximation Alogrithms*

2-Approximation Algorithm

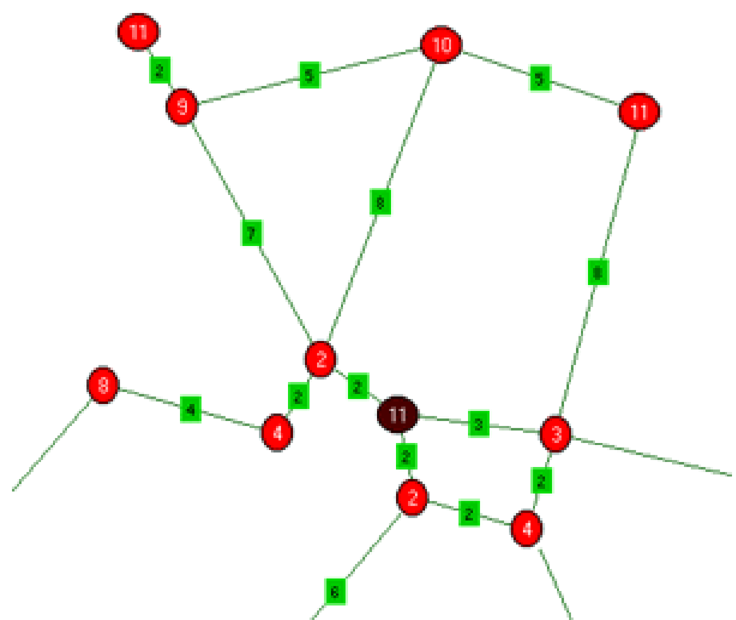
1) Order all edges e_i by cost
2) Construct graphs G_i containing all edges up to e_i
3) Construct square graphs G_i^2
4) Compute maximal independent set M_i of G_i^2
5) Find smallest i s.t. $|M_i| <= k$, say j
6) Return M_j

- Best possible polynomial time approximation: 2
- At least $O(n^3)$
  - Resources: Vazirani (2003), *Approximation Alogrithms*

2-Approximation Algorithm

1) Order all edges e_i by cost
2) Construct graphs G_i containing all edges up to e_i
3) **Construct square graphs G_i^2**
4) Compute maximal independent set M_i of G_i^2
5) Find smallest i s.t. $|M_i| <= k$, say j
6) Return M_j

- Square graph contains a one-hop connection wherever base graph had a one- or two-hop connection



2-Approximation Algorithm

1) Order all edges e_i by cost
2) Construct graphs G_i containing all edges up to e_i
3) Construct square graphs G_i^2
4) **Compute maximal independent set M_i of G_i^2**
5) Find smallest i s.t. $|M_i| <= k$, say j
6) Return M_j

- Maximal independent set
  - A set S such that every edge of the graph has at least one endpoint not in S and every vertex not in S has at least one neighbor in S
  - aka independent dominating set

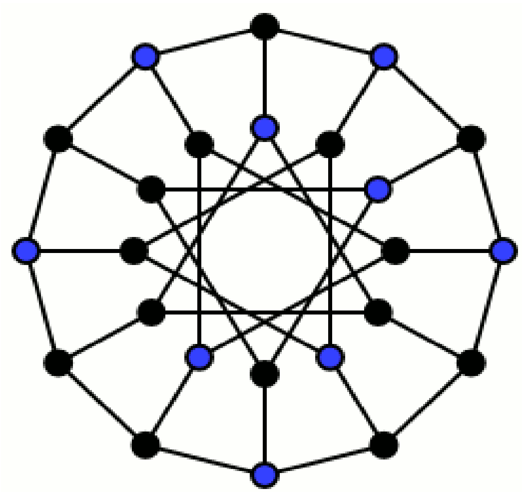


Image from en.wikipedia.org

Lloyd algorithm

```
1) Pick initial
   centroids
2) Given centroids,
   compute clusters
3) Given clusters,
   compute new
   centroids
4) Repeat 2 & 3 until
   “convergence”
   (centroids don't
   move very much)
```

- Most commonly used heuristic solver
  - Nearly synonymous with k-means
  - aka Voronoi iteration
  - Over 2,500 hits on G scholar
- Converges quickly to a good approximation in practice
  - Num iterations often << *n*
- Many applications
- Poor theoretical bounds

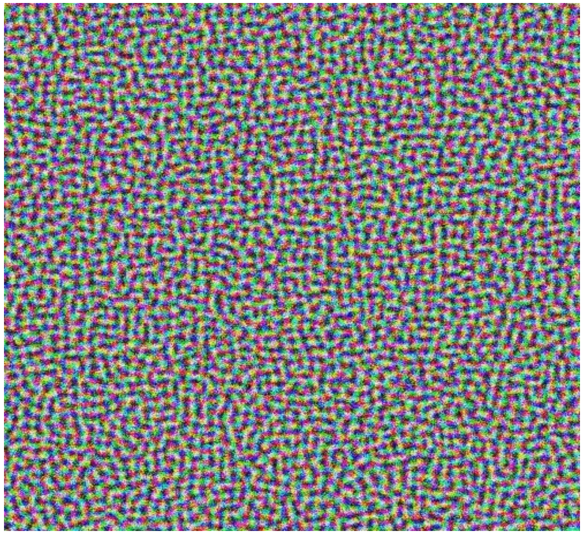
Lloyd algorithm

```
1) Pick initial
   centroids
2) Given centroids,
   compute clusters
3) Given clusters,
   compute new
   centroids
4) Repeat 2 & 3 until
   "convergence"
   (centroids don't
   move very much)
```

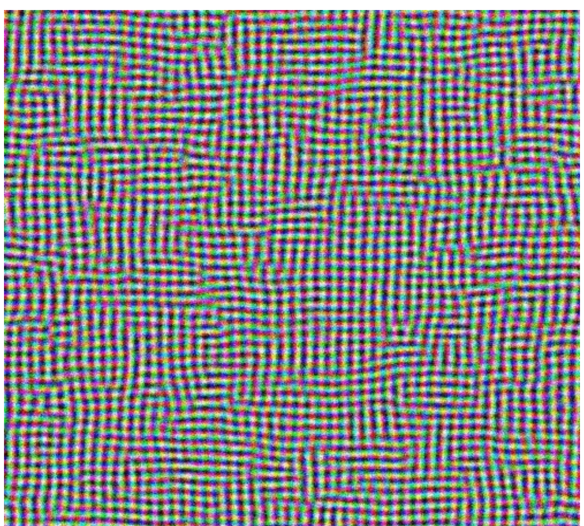
- Bad bounds
  - Time: super-polynomial in *n*
  - Approximation: can get stuck in local minimum
- "Seeding" initial centroids very important
  - Many complex methods for picking initial centroids
- Resources:
  - Lloyd (1957), "Least squares quantization in PCM"
  - Arthur & Vassilvitskii (2006), "How Slow is the k-means Method?"
  - Arthur & Vassilvitskii (2007), "k-means++ The Advantages of Careful Seeding"

Simulated Annealing

Fast cooling



Slow cooling

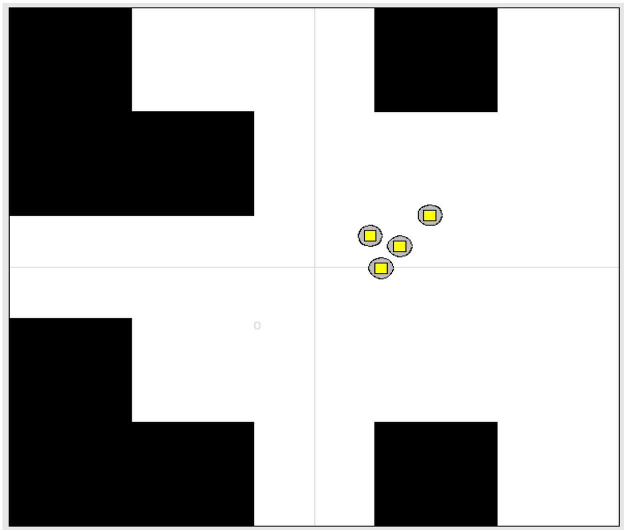


Images from en.wikipedia.org

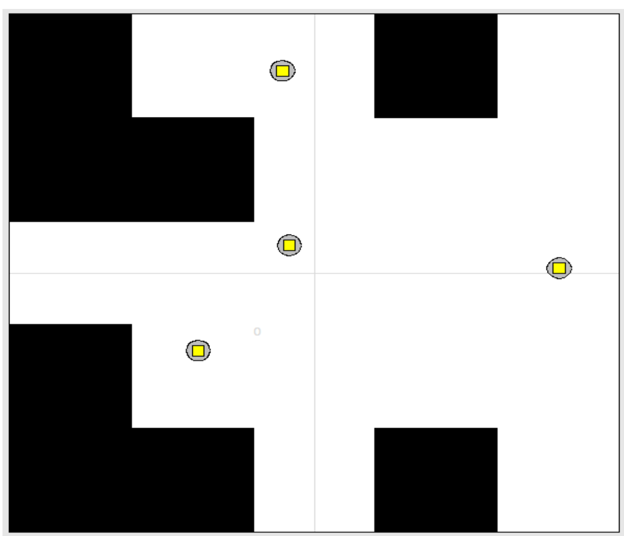
- Lloyd algorithm with added randomness
  - "Temperature" T controls level of randomness
  - At high temperature, bypasses local minima
- T is decreased on a schedule
  - Schedule affects result
  - Ideal cooling rate cannot be pre-computed
- Resources:
  - Kirkpatrick, Gelatt and Vecchi (1983), "Optimization by Simulated Annealing"

Deterministic Annealing

High T solution



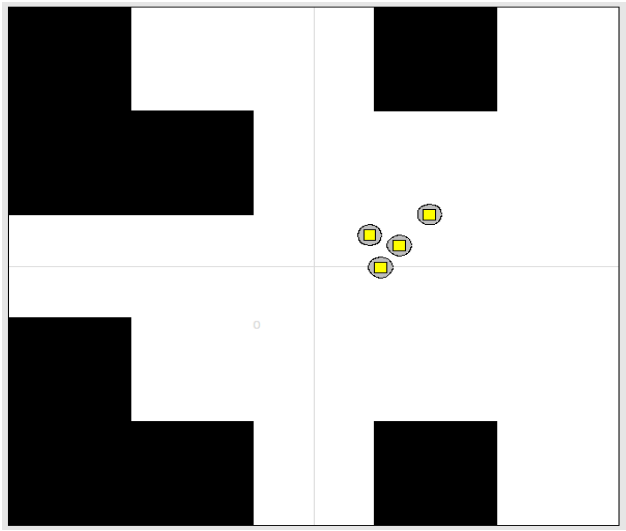
Low T solution



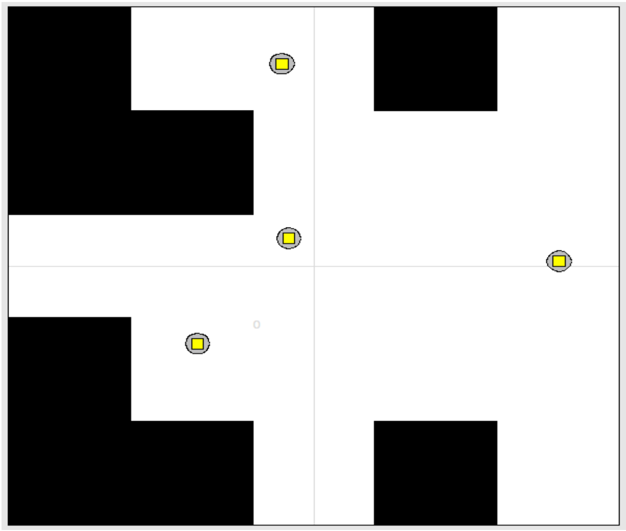
- Not stochastic!
  - Fractional ownership of vertices based on "temperature" *T*
- *T* controls centroid greed
  - At T = inf, every centroid claims every vertex equally
  - At T = 0, like Lloyd
- Resources:
  - Rose (1998), "Deterministic annealing for clustering, ..."

Deterministic Annealing

High T solution



Low T solution



- Like S.A., at high *T* D.A. bypasses local minima
  - Without randomness
- Still requires a temperature schedule
  - Again, determining an ideal schedule is complex
  - Depends on topography

Summary: k-Center Variations

	k-center	k-means	k-medoids
Datapoints in:	Graph	Cont. space	Cont. space
Centroids	In set	Not in set	In set
Distance norms	Max or 1	2	2

Summary: Solvers

	Approx. alg.	Lloyd alg.	Simulated Annealing	Deterministic Annealing
Approx. factor	2	?	?	?
Running time	Long	Short to very long	(# iter)*(lloyd)	(# iter)*(lloyd)
Stuck in local min	NA	Yes	No with good T schedule	No with good T schedule
Seeding importance	NA	High	Low	Low