Clustering with the k-means algorithm I

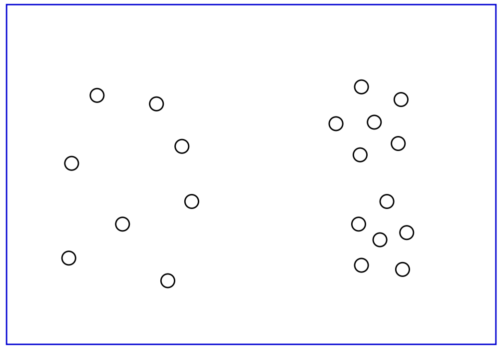
Sanjoy Dasgupta

University of California, San Diego

Topics we'll cover

1. The clustering problem
2. Two uses of clustering
3. The k-means cost function and algorithm
4. Initializing Lloyd's algorithm

Clustering in $\mathbb{R}^d$



Two common uses of clustering:

- Vector quantization
  Find a finite set of representatives that provides good coverage of a complex, possibly infinite, high-dimensional space.
- Finding meaningful structure in data
  Finding salient grouping in data.

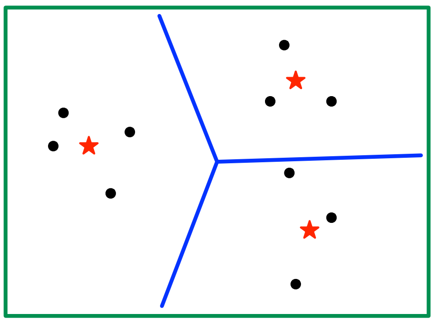
Widely-used clustering methods

1. K-means and its many variants
2. EM for mixtures of Gaussians
3. Agglomerative hierarchical clustering

The k-means optimization problem

- Input: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; integer k
- Output: "Centers", or representatives, $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$
- Goal: Minimize average squared distance between points and their nearest representatives:

$$\text{cost}(\mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \min_j \|x_i - \mu_j\|^2$$

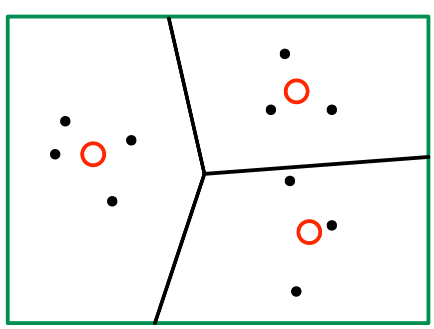


The centers partition $\mathbb{R}^d$ into k convex regions: μ_j's region consists of points for which it is the closest center.

Lloyd's k-means algorithm

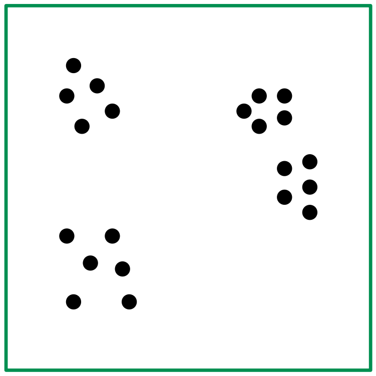
The k-means problem is NP-hard. Most popular heuristic: "k-means algorithm".

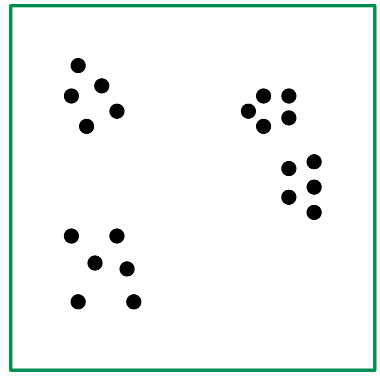
- Initialize centers $\mu_1, \ldots, \mu_k$ in some manner.
- Repeat until convergence:
  - Assign each point to its closest center.
  - Update each μ_j to the mean of the points assigned to it.



Each iteration reduces the cost $\Rightarrow$ convergence to a local optimum.

Initialization matters





Initializing the k-means algorithm

Typical practice: choose k data points at random as the initial centers.

Another common trick: start with extra centers, then prune later.

A particularly good initializer: ***k*-means++**

- Pick a data point x at random as the first center
- Let $C = \{x\}$ (centers chosen so far)
- Repeat until desired number of centers is attained:
  - Pick a data point x at random from the following distribution:

$$\Pr(x) \propto \text{dist}(x, C)^2,$$

  where $\text{dist}(x, C) = \min_{z \in C} \|x - z\|$
  - Add x to C