Assignment-Based Clustering

Note Title 2/10/2016

(class starting at 3:05)

k-means

$X \subset \mathbb{R}^d$

$d(a,b) = \|a-b\|_2$

X set of data points

$d: X \times X \to \mathbb{R}_+$

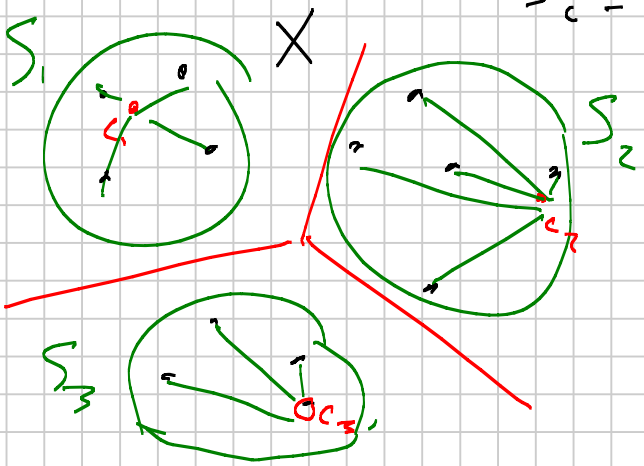
Goal: Find a set of k centers C

Minimize a cost $(C, X) = f\left(\text{cost}(c_i, S_i)\right)$

$$\phi_C(x) = \arg\min_{c_i \in C} d(x, c_i)$$

k clusters $S_1, \ldots S_k$

$$S_i = \{x \in X \mid \phi_C(x) = c_i\}$$



k-means

$$\text{cost}_2(C, X) = \sum_{x \in X} d(x, \phi_C(x))^2$$

$$= \sum_{c_i \in C} \sum_{x \in S_i} d(x, c_i)^2$$

k-center

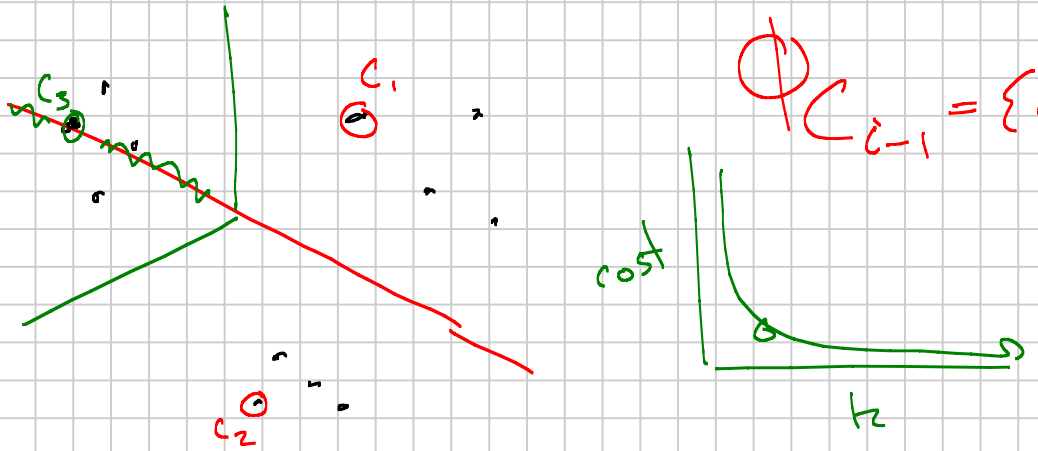
$$\text{cost}_\infty(C, X) = \max_{x \in X} d(x, \phi_C(x))$$

k-median $\text{cost}_1(C, X) = \sum_{x \in X} d(x, \phi_C(x))$

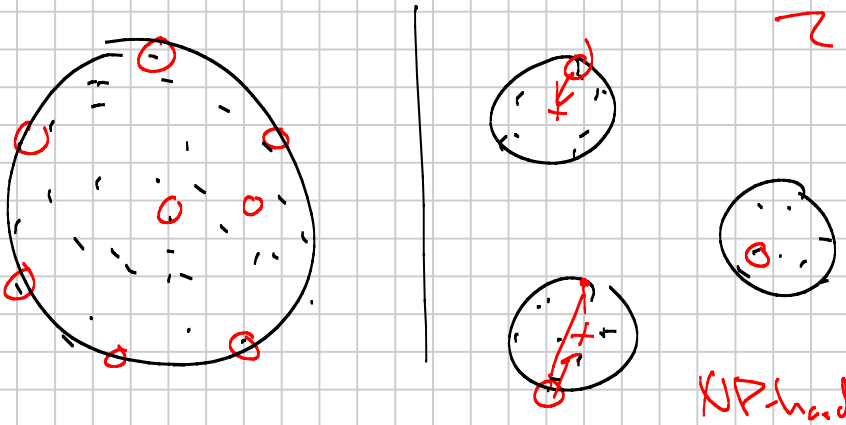
k-medoid $\text{cost}_1(C, X)$ s.t. $C \subset X$

Algorithm for k-center (Gonzalez)

1. Choose $c_1 \in X$ arbitrarily — $C_1 = \{c_1\}$
2. for $i = 2$ to k do — $C_2 = \{c_1, c_2\}$, $C_i = \{c_1, c_2, \ldots, c_i\}$
3. Set $c_i = \arg\max_{x \in X} d(x, \phi_{C_{i-1}}(x))$



$\phi_{C_{i-1}} = \{c_1, c_2, \ldots, c_{i-1}\}$



2-approximation

$cost_\infty(X, C) \leq 2\, cost_\infty(X, C^*)$

C^*: optimal center

NP-hard to $(2-\varepsilon)$-apx

Lloyd's Algorithm for k-means

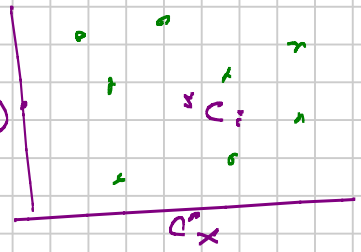
1957 → publish 1982

1. Choose k points $C \subset X$ [arbitrary?]
2. repeat
   a) for $x \in X$, find $\phi_C(x)$ — $O(kn)$ ← reassign to closest center
   b) for $i \in [k]$, $c_i = \text{average}(S_i) = \{x \in X \mid \phi_C(x) = c_i\}$ — $O(n)$
3. until (converged) ← if cost change $\leq \tau$ ← threshold

$c_i = \text{average}\{S_i\}$

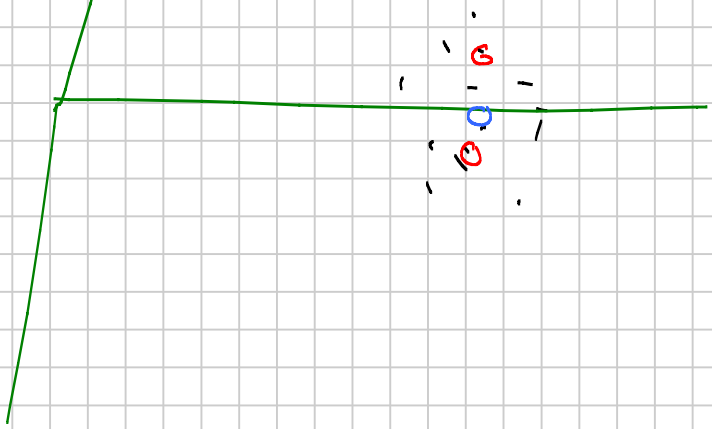
$\Leftrightarrow$

$c_i = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{s \in S_i} \|s - z\|^2$



$c_x = \frac{1}{|S_i|} \sum_{s \in S_i} s_x$

$\text{cost}(X, C) = \sum_{x \in X} \|x - \phi_C(x)\|^2$



k^n possible clusters

n data points

$\{1, 2, 3, \ldots k\}$

How to choose initial C centers

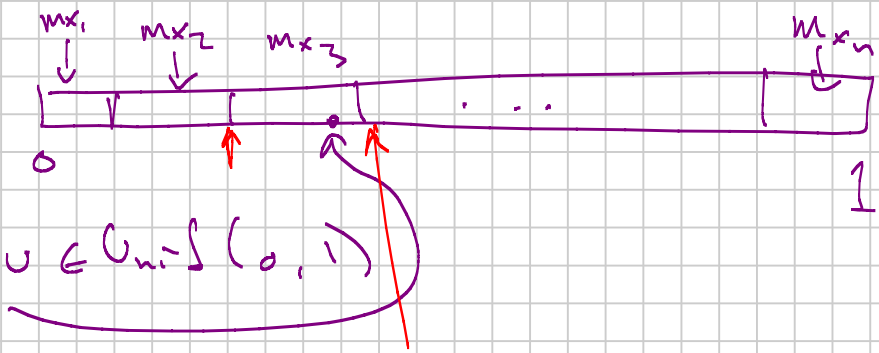
- random (keep lowest cost solution after running Lloyds)
- Coupon Collectors
  Run Lloyds w/ $k \log k$ centers
  Regroup w/ HAC
- Gonzalez k-center Algo

k-means++

1. Choose $c_1 \in X$ arbitrarily $\quad C_1 = \{c_1, c_2 \ldots\}$
2. for $i = 2$ to k
   Choose c_i from X with probability proportional to $d(x, \phi_{C_{i-1}}(x))^2$

$$M = \sum_{x \in X} \left(d(x, \phi_{C_{i-1}}(x))^2 = m_x \right)$$

$$\text{Prob}[\text{choosing } x] = \frac{m_x}{M}$$



$u \in \text{Unif}(0, 1)$

end of m_{x_3} $= \sum_{i=1}^{3} m_{x_i}$