Clustering

Anil Maheshwari

anil@scs.carleton.ca
School of Computer Science
Carleton University
Canada

Introduction

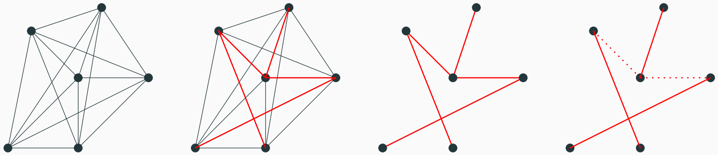
Clustering Problem

Input: A set $X = \{x_1, \ldots, x_n\}$ of n objects. For every pair $x_i, x_j \in X$, we have the distance $d(x_i, x_j) \geq 0$ such that $d(x_i, x_i) = 0$ and $d(x_i, x_j) = d(x_j, x_i)$.

Problem: Divide objects in X into k non-empty groups such that the *gap* between the groups is as large as possible. Distance between two groups is defined as the smallest distance between pair of points, where in the pair points belong to different groups.

Solution

1. Define a complete graph $G = (V = X, E)$, where each edge $e = (x_i, x_j)$ has a weight $d(x_i, x_j)$.
2. Construct a minimum spanning tree T of G
3. Delete $k - 1$ most expensive edges from T
4. Output the resulting k-connected components $C_1, \ldots, C_k$



Claim

The components $C_1, \ldots, C_k$ constitute a k-clustering of X that maximizes the gap.

K-Means Clustering Problem

Input: A set $X = \{x_1, \ldots, x_n\}$ of n-points in $\mathbb{R}^d$. An integer $0 < k \leq n$.

Problem: Partition X into k non-empty clusters $C_1, \ldots, C_k$. Points within a cluster should be *close* to each other compared to points outside the cluster.

Potential Function

Let $C_1, \dots, C_k$ be a k-clustering of X with centers $\mathcal{C} = \{c_1, \dots, c_k\}$, where $c_i \in \mathbb{R}^d$.

Define the potential function $\Phi(\mathcal{C}) = \sum_{x \in X} \min_{c \in \mathcal{C}} d(x, c)^2 = \sum_{x \in X} \min_{c \in \mathcal{C}} ||x - c||^2$

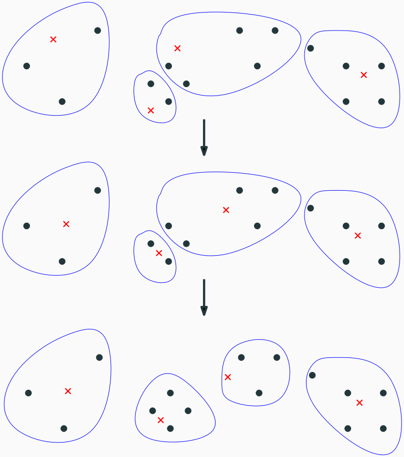
$\Phi(\mathcal{C})$= Sum of the squared distance between each point x in X to its nearest center in $\mathcal{C}$

Problem:
Given X and k, find k-centers $\mathcal{C}$ such that the corresponding clustering $C_1, \dots, C_k$ minimizes $\Phi(\mathcal{C})$

Llyod's Heuristic

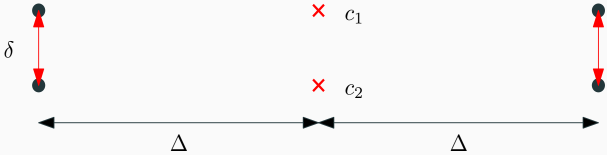
1. Select Initial *Centers*: Arbitrary choose k-centers and initialize $\mathcal{C} = \{c_1, \ldots, c_k\}$
2. Partition X: Compute sets $C_1, \ldots, C_k$ with respect to centers in $\mathcal{C}$. Point $x \in X$ is assigned to the cluster C_i if x's nearest center in $\mathcal{C}$ is c_i.
3. Recompute Centers: For each $i \in \{1, \ldots, k\}$, set c_i (the new cluster center) to be the center of the mass of points in C_i
4. Repeat Steps 2 & 3 till $\mathcal{C}$ no longer changes

An illustration of a Phase of Llyod’s Algorithm



Observations

Let $\Phi(\mathcal{C}^*)$ be the potential of an optimal clustering and let $\Phi(\mathcal{C})$ be the potential of the clustering returned by Llyod's heuristic.



Competitive Ratio

Competitive ratio of Llyod's heuristic is unbounded, i.e. $\frac{\Phi(\mathcal{C})}{\Phi(\mathcal{C}^*)} \to \infty$

Observations

Decrease in Potential

Each execution of Steps 2 & 3 decrease the value of the potential function.

Proof: We will use the following Lemma.

Lemma 1

Consider a set of points S. Let m^* denote the center of mass of S. Let z be an arbitrary point. Define $\Delta(S, z) = \sum_{x \in S} d(x, z)^2$. Then $\Delta(S, z) = \Delta(S, m^*) + |S| d(m^*, z)^2$

Corollary 1

If S is a single cluster with initial center z, then moving the cluster center to m^* reduces the potential as $\Delta(S, z) - \Delta(S, m^*) = |S| d(m^*, z)^2 \geq 0$

Proof of Lemma

Lemma 1

$\Delta(S,z) = \Delta(S,m^*) + |S| d(m^*,z)^2$, where m^* is center of mass of S and z is an arbitrary point

Proof: Assume we are in 2-dimensions. Let $z = (z_x, z_y)$ and $S = \{p_1, \ldots, p_n\}$, where $p_i = (x_i, y_i)$.

1. $m^* = \left(\frac{\sum\limits_i x_i}{n}, \frac{\sum\limits_i y_i}{n} \right)$
2. $\Delta(S,z) = \sum\limits_{p \in S} d(p,z)^2 = \sum\limits_{p=(x_i,y_i) \in S} \left((x_i - z_x)^2 + (y_i - z_y)^2 \right)$
3. $\Delta(S,m^*) = \sum\limits_{p \in S} d(p,m^*)^2 = \sum\limits_i \left(x_i - \frac{\sum x_i}{n} \right)^2 + \sum\limits_i \left(y_i - \frac{\sum y_i}{n} \right)^2$
4. $\Delta(S,z) - \Delta(S,m^*)$
   $= nz_x^2 + nz_y^2 - 2z_x \sum x_i - 2z_y \sum y_i + n\left(\frac{\sum x_i}{n}\right)^2 + n\left(\frac{\sum y_i}{n}\right)^2$
   $= n\left[z_x^2 + z_y^2 - 2z_x \frac{\sum x_i}{n} - 2z_y \frac{\sum y_i}{n} + \left(\frac{\sum x_i}{n}\right)^2 + \left(\frac{\sum y_i}{n}\right)^2 \right]$
   $= |S| d(m^*,z)^2$

□

How to choose initial centers?

Question: How to choose initial centers so that we are guaranteed to have some bounded competitive ratio with respect to optimum?

k-means++ Algorithm

Let $D(x)$= Shortest distance from x to the nearest center among the current set of centers.

$k++$ **Means Algorithm:**

Step 1: Choose an initial cluster center c_1 uniformly at random from X.

Step 2: (**Randomization Step**) Choose the next center c_i by selecting a point $x \in X$ with probability $\frac{D(x)^2}{\sum\limits_{y \in X} D(y)^2}$

Step 3: Repeat Step 2 till k centers are chosen

Step 4: Execute Llyod's Heuristic by choosing $\{c_1, \ldots, c_k\}$ as the initial centers

Observations

1. In the Randomization Step, the points of X that are farther from the currently chosen centers have a higher chance of being selected.
2. The algorithm is $8(\ln k + 2)$-competitive. Let the k-centers returned by k-means++ algorithm be $\mathcal{C}$. Then, $E[\Phi(\mathcal{C})] \leq 8(\ln k + 2)\Phi(\mathcal{C}^*)$.
3. Claim holds for the clustering obtained after Step 3. Step 4 may further improve.
4. Proof is not easy. Consider clusters of an optimal solution $\mathcal{C}^*$. The authors show
   - The algorithm is 2-competitive w.r.t. the points in the optimal cluster, say A, from where the first center c_1 is chosen by the algorithm
   - The algorithm is 8-competitive in all those clusters of optimal from which the algorithm chooses a center.
   - If $\mathcal{C}$ doesn't have centers from some of the clusters of the optimal, then the algorithm is $8(\ln k + 2)$-competitive.

Useful Notations

1. Let $\mathcal{C}$ be the clustering computed by the k-means++ algorithm
2. Let $\mathcal{C}^*$ be an optimal clustering
3. $d(x, c) = ||x - c||$ is the Euclidean distance between points x and c
4. Let $D(x) =$ Shortest distance from x to the nearest center in $\mathcal{C}$ (or $\mathcal{C}^*$).
5. $\Phi(\mathcal{C}) = \Phi_{\mathcal{C}}(X)$ refers to potential with respect to the point set X.
   Formally, $\Phi(\mathcal{C}) = \sum_{x \in X} \min_{c \in \mathcal{C}} d(x, c)^2 = \sum_{x \in X} D(x)^2$
   I.e. $\Phi(\mathcal{C})$= Sum of the squared distance between each point in X to its nearest center in $\mathcal{C}$
6. For a subset $A \subseteq X$, define $\Phi_{\mathcal{C}}(A) = \sum_{x \in A} \min_{c \in \mathcal{C}} d(x, c)^2 = \sum_{x \in A} D(x)^2$.

1st Center from an Optimal Cluster

Claim 1

Let A be an arbitrary cluster in optimal $\mathcal{C}^*$. Let $\mathcal{C}$ be the clustering with exactly one center that is chosen from A uniformly at random. Then, $E[\Phi_{\mathcal{C}}(A)] = 2\Phi_{\mathcal{C}^*}(A)$.

Proof: By definition of expected value, $E[\Phi_{\mathcal{C}}(A)] = \sum_{a_0 \in A} \frac{1}{|A|} \sum_{a \in A} ||a - a_0||^2$

From Corollary 1, in $\mathcal{C}^*$, cluster center of A will be its center of mass, say m^*.

$$\begin{aligned} E[\Phi_{\mathcal{C}}(A)] &= \frac{1}{|A|} \sum_{a_0 \in A} \left(\left(\sum_{a \in A} ||a - m^*||^2 \right) + |A| ||a_0 - m^*||^2 \right) \text{ (By Lemma 1} \\ &= 2 \sum_{a \in A} ||a - m^*||^2 \\ &= 2\Phi_{\mathcal{C}^*}(A) \end{aligned}$$

□

Other Centers from Optimal Clusters

Claim 2

Let A be an arbitrary cluster in optimal $\mathcal{C}^*$. Let $\mathcal{C}$ be an arbitrary clustering. Suppose the next center to $\mathcal{C}$ in the k-means++ algorithm is added from A, $E[\Phi_{\mathcal{C}}(A)] \leq 8\Phi_{\mathcal{C}^*}(A)$.

Proof Sketch: By triangle inequality we have for all a and a_0, $D(a_0) \leq D(a) + ||a - a_0||$.

Note that for reals x and y, $\frac{1}{2}(x+y)^2 \leq x^2 + y^2$

Thus, we have $\frac{1}{2}(D(a_0))^2 \leq \frac{1}{2}(D(a) + ||a - a_0||)^2 \leq D(a)^2 + ||a - a_0||^2$

Equivalently, $D(a_0)^2 \leq 2D(a)^2 + 2||a - a_0||^2$

Summing over all elements of A, we have

$\sum_{a \in A} D(a_0)^2 \leq \sum_{a \in A} \left(2D(a)^2 + 2||a - a_0||^2\right)$

Or, $D(a_0)^2 \leq \frac{2}{|A|} \sum_{a \in A} D(a)^2 + \frac{2}{|A|} \sum_{a \in A} (a - a_0)^2$

Other Centers from Optimal Clusters (contd.)

Probability of choosing $a_0 \in A$ as a center $= \frac{D(a_0)^2}{\sum\limits_{a \in A} D(a)^2}$

$$E[\Phi_{\mathcal{C}}(A)] = \sum_{a_0 \in A} \frac{D(a_0)^2}{\sum\limits_{a \in A} D(a)^2} \sum_{a \in A} \min(D(a), (a - a_0))^2$$

Substituting the expression for $D(a_0)^2$ in $E[\Phi_{\mathcal{C}}(A)]$ we obtain:

$$E[\Phi_{\mathcal{C}}(A)] \leq \frac{2}{|A|} \sum_{a_0 \in A} \frac{\sum\limits_{a \in A} D(a)^2}{\sum\limits_{a \in A} D(a)^2} \sum_{a \in A} \min(D(a), (a - a_0))^2 +$$

$$\frac{2}{|A|} \sum_{a_0 \in A} \frac{\sum\limits_{a \in A} (a - a_0)^2}{\sum\limits_{a \in A} D(a)^2} \sum_{a \in A} \min(D(a), (a - a_0))^2$$

Substitute for $\min(D(a), (a - a_0))^2 \leq (a - a_0)^2$ in 1st part and $\min(D(a), (a - a_0))^2 \leq D(a)^2$ in 2nd part and we obtain:

$E[\Phi_{\mathcal{C}}(A)] \leq \frac{4}{|A|} \sum\limits_{a_0 \in A} \sum\limits_{a \in A} (a - a_0)^2 = 8\Phi_{\mathcal{C}^*}(A)$ (By Claim 1).

□

Centers Not from Optimal Clusters

Claim 3

Let $\mathcal{C}$ be an arbitrary clustering. Choose $u > 0$ uncovered clusters from $\mathcal{C}^*$. Let X_u be the points in these clusters and define $X_c = X - X_u$. Assume that the algorithm adds $t \leq u$ random centers to $\mathcal{C}$. Let $\mathcal{C}'$ be the resulting clustering and Φ' be its potential. The following inequality holds $E[\Phi'] \leq (\Phi(X_c) + 8\Phi^*(X_u))\,(1 + H_t) + \frac{u-t}{u}\Phi(X_u)$.

Note that $H_t = \sum\limits_{i=1}^{t} \frac{1}{i} \approx \ln t$.

Proof Idea: Please see the paper by Arthur and Vassilvitskii, k-means++, 8th ACM-SIAM Symposium on Discrete algorithms, 2007 for details.

Proof is based on induction on values of (t, u). It is shown that if it holds for $(t-1, u)$ and $(t-1, u-1)$ then it also holds for (u, t).

Main Result

Theorem

Let the k-centers returned by k-means++ algorithm be $\mathcal{C}$. Then, $E[\Phi(\mathcal{C})] \leq 8(\ln k + 2)\Phi(\mathcal{C}^*)$, i.e. the algorithm is $8(\ln k + 2)$-competitive.

Proof: Let A be the cluster of $\mathcal{C}^*$ from where the first center was chosen by k-means++ algorithm.

Now set $t = u = k - 1$ and use Claim 3.

We have $X_c = A$, and $X_u = X - A$. We obtain

$$\begin{aligned} E[\Phi(\mathcal{C})] &\leq (\Phi(A) + 8\Phi_{\mathcal{C}^*}(X_u))(1 + H_t) + \frac{u-t}{u}\Phi(X_u) \\ &= (\Phi(A) + 8\Phi_{\mathcal{C}^*}(X) - 8\Phi_{\mathcal{C}^*}(A))(1 + H_{k-1}) \text{ (As } X = X_c \cup X_u\text{)} \\ &\leq 8(1 + H_{k-1})\Phi_{\mathcal{C}^*}(X) \text{ (By Claim 1)} \\ &\leq 8(\ln k + 2)\Phi(\mathcal{C}^*) \text{ (as } H_{k-1} \leq 1 + \ln k\text{)} \end{aligned}$$

□

Summary

Implications:

If the centers in k-means++ algorithm are chosen from each cluster of $\mathcal{C}^*$,
$\implies$ Algorithm is 8-competitive.

What if the algorithm doesn't choose centers from some of clusters of $\mathcal{C}^*$?

- This part introduces $8(\ln k + 2)$-factor in the analysis

Theorem (Arthur and Vassilvitskii 2007)

The k-means++ algorithm is $8(\ln k + 2)$-competitive