K-means FPT algorithm

The k-means problem: \[ \min \sum_{i=1}^{n} \| p_i - (\text{closest point to } p_i \text{ in } \mathcal{C}) \|^2 \]

Equivalently: find clusters \( C_1, C_2, \ldots, C_k \), and let \( C_i \) be the center/mean of \( C_i \) that partitions \( P \)

\[ \text{minimize} \quad K \sum_{i=1}^{k} \sum_{p \in C_i} \| p - C_i \|^2 \]

Note: \( C_i \) addresses \( \min \sum_{p \in C_i} \| p - C_i \|^2 \)

- The k-means (or Lloyd's algorithm) is an alternating minimization algorithm that is used widely.
- But only finds local minima
- Also may run very slowly in practice (though widely used in practice).
- Exactly: can solve in time \( K^n \) (of course).
- Also use some weighted Voronoi diagrams, can solve in \( O(n K^d) \) time.

Today: get a \( (1+\epsilon) \)-approx in time \( O(\text{nd}^2 \log n) \).

This gets \( f(K, \epsilon) \cdot \text{poly}(\text{nd}) \) time.

Another lecture: we will get runtimes that are \( f(d, \epsilon) \cdot \text{poly}(n) \).

Btw: we can assume that \( d \leq n-1 \) since we can solve SVD and bring points down to \((n-1)\)-dimensional space.

Note: NP hard to solve exactly even in \( \mathbb{R}^2 \)!
OK: Step in this: Input \( P = \{p_1, p_2, \ldots, p_n\} \subseteq \mathbb{R}^d \). For a subset \( S \subseteq \mathbb{R}^d \), define \( g(S) = \frac{1}{|S|} \sum_{x \in S} x \) = center of gravity.

Define \( \Delta(S, c) = \sum_{x \in S} \|x - c\|^2 \)

de the \( k \) means objective.

The \( k \) means problem: find partition \( C_1, C_2, \ldots, C_k \) \( (\text{OPT}) \)

\[ \text{min} \sum_{i=1}^{k} \Delta(C_i, g(C_i)) \]

or equivalently: find \( c_1, c_2, \ldots, c_k \in \mathbb{R}^d \)

\[ \text{st. min} \sum_{i=1}^{k} \min_{c \in C_i} \|p_i - c\|^2 \]

Consider \( 1 \) means problem (easy).

**Lemma (Indri et al.):**

Let \( S \) be a sample of \( \mathbb{R}^d \) \( P \). (uniformly, of size \( \Omega(n) \)) then \( \Delta(P, g(S)) \leq \left(1 + \frac{n}{\min} \right) \Delta(P, g(P)) \)

with prob \( \delta \).

**Proof:** First use an exercise:

\[ \Delta(P, x) = \Delta(P, g(P)) + \|p_i - g(P)\|^2 \]

So suffices to show that

\[ \Pr \left[ \|g(S) - g(P)\|^2 \leq \frac{\delta}{\min} \Delta(P, g(P)) \right] \geq 1 - \delta \]

Define \( g(P) = c^* = \frac{1}{|P|} \sum_{x \in P} x \)

Note: \( \mathbb{E} g(S) = c^* \)
Lemma: Let \( P \) be a set of \( n \) points. For any \( X \subseteq P \), let \( g(X) \) be the set of points in \( P \) that are closer to \( X \) than to any other point in \( P \). Then for any \( \epsilon > 0 \), there is an \( M \) such that

\[
|\Delta(P, g(X))| \leq \left(1 + \frac{1}{\delta M}\right) \Delta(P, g(P))
\]

whenever \( |X| \geq M \).

**Proof:**

Since \( \Delta(P, g(X)) = \Delta(P, g(P)) + \| g(X) - g(P) \|^2 \), it suffices to show that

\[
\| g(X) - g(P) \|^2 \leq \frac{1}{|P|} \cdot \frac{1}{\delta M} \cdot \Delta(P, g(P)).
\]

Set \( X_j \) as a random point from \( P \), \( j = 1, 2, \ldots, M \). Then

\[
E[X_j] = \frac{1}{|P|} \cdot \frac{1}{\delta M} \cdot \Delta(P, g(P)).
\]

Let \( cg(P) = \frac{1}{M} \sum_{j=1}^{M} X_j \) be the average of these \( M \) samples. Then

\[
E[|g(P) - g(X)|] = E[|X_j - g(P)|] \leq \frac{1}{M} \cdot \Delta(P, g(P)).
\]

\[
\Rightarrow P_X[\| g(X) - g(P) \|^2 \geq \delta \cdot \text{expectation}] \leq \delta.
\]

Note: If you pick points from some distribution that picks at least \( M \) points, you are still fine.
Algorithm for Balanced Case (Simplified)

$V_i = 1..K$

Pick $S_i =$ random sample (uniform) of $\frac{10k}{e}$ points from $F_i$. 

Now for each choice of $T_i, T_k$, where $T_i \leq S_i \leq T_k$.

Try the solution $a_i = c_i(T_i)$

Output the best.

Runtime = simple time $O(nk) + \left( \frac{10k^2}{e} \right)^k O(nd)$ <- for all the enumeration.

$\Pr(\text{Success}) \geq \frac{1}{4k}$.

1. For some $i$, then claim: $\Pr \left[ |S_i \cap O_i| \geq \frac{2k}{e} \right] \geq \frac{1}{2}$.

Proof: $E[|S_i \cap O_i|] = \frac{10k}{e}k = \frac{10k}{e}$.

Now: the $S_i \cap O_i$ is like a set of $\frac{10k}{e}$ copies with success probability $\frac{10k}{n}$.

So $\text{Var}(|S_i \cap O_i|) \leq E[|S_i \cap O_i|]$

$\Rightarrow$ Chebyshev says $\Pr \left[ |S_i \cap O_i| \leq \frac{1}{3} \text{ mean} \right] =$

$\leq \Pr \left[ \left| S_i \cap O_i - \text{mean} \right| > \frac{4}{3} \text{ mean} \right]$

$\leq \frac{\text{Var}(S_i \cap O_i)} \left( \frac{4}{3} \text{ mean} \right)^2 \leq \frac{1}{\left( \frac{4}{3} \text{ mean} \right)^2} \leq O(e)$

2. Conditioned on this good event, any $T \subseteq (S_i \cap O_i)$ is a random sample of $O_i$, and hence Indaba et al. says $c_i(T)$ is a good soln. + $M \geq \frac{2k}{e}$.
\[ \Rightarrow \text{w.p. } \frac{1}{4^n}, \text{ get a soln whose cost } \leq (1+\epsilon) \text{OPT} \]

\[ \Rightarrow \text{w.p. } \frac{1}{2^n}, \text{ all } |S_i \cap N_i| \text{ are large and hence } |T_i| \leq |S_i \cap N_i| \text{ of size } \frac{n}{2} \]

and then w.p. \( \frac{1}{2^n} \), all the \( cy(T_i) \) are good solns (since \( d = \frac{1}{2} \)).
What about unbalanced case?

Different clusters may have different sizes
(but the same cost)

And very far away from each other. So if you miss one cluster, poor apa!!

Solution: "D² sampling" + Iteratively find clusters.

High level algo:
1. Call JKS(C(0))

JKS(C(0), C(set of clusters))

2. If n ≤ k sample set S(0) from P, where p ∈ P chosen w.p. 0.1 ||p - C(0)||²

For all possible sets T ≤ S(0) of size O(kε)

JKS(C(0), C(0) ⊆ C(T) U ε C(T))

else if i = k,

store the solution if better than previous ones.

Runtime: \[O(k^2 \cdot \frac{k}{2})\] sampling cycles. \[\approx O(k^2 \cdot O(4^k))\]

Will show: prob of success \(\geq 2^{O(k\epsilon)}\). So can repeat this \(2^{O(k\epsilon)}\) times.

Issues: Sample S(0) is not uniform from the clusters.

Want to say that S(0) contains a subset that is (almost uniform

from among the remaining clusters, of size \(\approx C(\epsilon)\).

So will be able to get another cluster in C(0).

Fix: \(G_1, G_2, \ldots, G_k\),

with centers \(c_1, c_2, \ldots, c_k\),

\[G_1 \cup G_2 \cup \cdots \cup G_k\]

don't need this.
Simplifying assumption: (Well separatedness or stability)

Assume that \( \text{OPT}_k \leq (1 + \frac{\epsilon_1}{k^2}) \text{OPT}_{k-1} \). (\( \leq \) stable)

Suppose not; then look for solution with \( k-1 \), it is not much costlier. But this process can go on only for \( k \) steps. Eventually find \( (1+\epsilon) \) approx to problem with \( i \leq k \) centers, but that \( \text{OPT}_i \leq \prod_{j=i+1}^{k} (1 + \frac{\epsilon_i}{j^2}) \text{OPT}_k \)

\( \leq (1 + O(\epsilon)) \text{OPT}_k \).

So we still get \( (1+O(\epsilon)) \) approx!

In fact that the optimal centers are well separated.

\[ c_i^* = s_g[c_i^*] \]

**Lemma 2:** \( \|c_i^* - c_j^*\|^2 \geq \delta (r_i + r_j) \)

**Pf:** Suppose not; then say \( |c_i^*| > |c_j^*| \).

\[ \Delta(c_i^* \cup c_j^*, c_k^*) = |c_i^*|.r_i + |c_j^*|.r_j \]

\[ + |c_j^*|. \|c_i^* - c_j^*\|^2 \]

\[ \leq (1 + \delta) \left[ \Delta(c_i^*, c_j^*) + \Delta(c_j^*, c_k^*) \right] \]

a contradiction to the \( \delta \)-stability.
Now suppose: \( \exists C_1^*, C_2^*, \ldots C_k^* \) s.t. our centers \( \bar{x}_1, \bar{x}_2, \ldots \bar{x}_k \) satisfy
\[
\Delta(C_j^*, C_j) \leq (1 + \frac{3}{20}) \Delta(C_j^*, C_i^*).
\]
then we show that w.p. \( \geq \frac{1}{2} \) (some constant), set new centers \( C_i^* \) to "hit" some clusters \( C_i^* \)
and maintain invariant (*).

Define \( d_p^i = \min \| p - \text{all} \| \) be the distance of \( p \) from previous clusters.

Define \( d_p^i \) to be the distance of \( p \) from previous clusters.

Then we pick \( p \) w.p.
\[
\frac{d_p^i}{\sum d_p^i} \text{ in round } i.
\]

Claim 3:
\[
\sum_{p \in \text{opt clusters}} d_p^2 \geq \frac{1}{2} \sum_{p \in \text{opt clusters}} d_p^2.
\]

Proof: Suppose not.
\[
\sum_{p \in \text{opt clusters}} d_p^2 = \sum_{p \in \text{1st ith-1 opt clusters}} d_p^2 + \sum_{p \in \text{i-1 opt clusters}} d_p^2 < \frac{1}{2} \sum_{p \in \text{opt clusters}} d_p^2,
\]

\[
\leq (1 + \frac{3}{20}) \text{ OPT cost of } \text{opt clusters}
\]

\[
\leq (1 + \frac{3}{20}) \text{ OPT cost in total}.
\]

Now: Claim 4: each point in \( C_i^* \) is sampled at least \( \frac{\sqrt{\log n}}{\log k} \) times.

\[
\sum_{p \in \text{opt clusters}} d_p^2 \geq \frac{\sqrt{\log n}}{\log k} \sum_{p \in \text{opt clusters}} d_p^2.
\]

\[
\Rightarrow \text{ sampled overall w.p. } \frac{d_p^i}{\sum_{p \in \text{opt clusters}} d_p^i} \geq \frac{\sqrt{\log n}}{2^{100}} \cdot \frac{1}{128k \cdot 16^k}.
\]
Pf of clm (4): let p be closest to $c_0 \in C^{(a)}$. i.e. $d_p = \|p - c_0\|^2$.

Then $\sum_{p \in C^i} d_p = \Delta(C^i, C^j) + \|d^i - c_0\|^2$ by relaxed triangle inequality.

Recall: j is me of the $i$th clusters, not 1..i-1.

And a is in $i_1..i-3$.

Also: $\|p - c_0\|^2 = d_p^2 = \frac{\|p - c_0\|^2}{2} - \|c_0 - c_{j}^i\|^2$

but $p$ is closest to $c_{j}^i$ than to $c_{a}^i$ since it's in $G^i$

so $\|c_0 - c_{j}^i\|^2$ by stability

$\Rightarrow \frac{d_p^2}{\sum_{p \in C^i} d_p^2} \geq \frac{\|c_0 - c_{j}^i\|^2}{\|c_0 - c_{a}^i\|^2}$

$\Rightarrow \frac{\|c_0 - c_{j}^i\|^2}{\|c_0 - c_{a}^i\|^2} \leq \frac{\|c_0 - c_{j}^i\|^2}{\|c_0 - c_{a}^i\|^2}$

\[\Rightarrow \frac{\|c_0 - c_{j}^i\|^2}{\|c_0 - c_{a}^i\|^2} \leq \frac{1}{16}. \theta(\bar{G})\]

Finally: Need to handle all elements of $G^i$ are not sampled up. $|G^i|$, but with

$1/2$ that lies between $\frac{1}{|G^i|} \cdot \theta(\bar{G})$ but with

But by inverting that we duplicate element and sample, we conside the same

We need to sample $1/2$ elements instead of $1/3$. 

\[\theta(\bar{G})\]

or, not sample reject
Other facts: Arthur and Vassilvitskii found Ostrovski Rakin and Schulman Sweeney

Show that kmeans++

which just does $D^2$-sampling:

1. Pick a random $v$ as $C^0$.
2. Pick a random $v$ as $C^i$ by sampling up to $x^i$ (distanced-squared) from $C^i$.

This is a valid solution for k-means that is $O(kdK)$ approx.

They show that this seeding already is good, and Lloyds also can only improve.

Other results:

[Arthur Manthey Roeglin] k-means has polynomial smoothed complexity,

but [Arthur Vassilvitskii, Vattani] 3 instances where k-means converges very slowly,

(even to local optimum) ≠ global optimum.

Can extend the k-means result to other metrics which have

- approx 1-norm and approx symmetry
- centroid property (that $\sum P_{i} \| p_{i} - q \|^2 = \sum P_{i} \| p_{i} - q \|^2 + \| p_{i} - l \|^2 - \| q - l \|^2$)
- sampling property.

See the Tulsar et al paper for details.