CSE 291: Geometric algorithms **Spring 2013**

Lecture 3 — Algorithms for k-means clustering

3.1 The k-means cost function

Although we have so far considered clustering in general metric spaces, the most common setting by far is when the data lie in an Euclidean space $\mathbb{R}^d$ and the cost function is k-means.

> k-MEANS CLUSTERING
>
> *Input:* Finite set $S \subset \mathbb{R}^d$; integer k.
>
> *Output:* $T \subset \mathbb{R}^d$ with $|T| = k$.
>
> *Goal:* Minimize $\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2$.

It is interesting that the cost function uses the square of the L_2 norm rather than L_2 norm. This is a fortuitous choice that turns out to simplify the math in many ways.

Finding the optimal k-means clustering is NP-hard even if $k = 2$ (Dasgupta, 2008) or if $d = 2$ (Vattani, 2009; Mahajan et al., 2012).

3.1.1 Voronoi regions

The representatives T induce a *Voronoi partition* of $\mathbb{R}^d$: a decomposition of $\mathbb{R}^d$ into k convex *cells*, each corresponding to some $z \in T$ and containing the region of space whose nearest representative is z.

This partition induces an optimal clustering of the data set, $S = \cup_{z \in T} C_z$, where

$$C_z = \{x \in S : \text{the closest representative of } x \text{ is } z\}.$$

Thus the k-means cost function can equally be written

$$\text{cost}(T) = \sum_{z \in T} \sum_{x \in C_z} \|x - z\|^2.$$

In analyzing algorithms, we'll sometimes need to consider suboptimal partitions $C_1, \ldots, C_k$ of S. To this end, define

$$\text{cost}(C_{1:k}, z_{1:k}) = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2.$$

3.1.2 A single cluster

Suppose a single cluster C is assigned representative z. The cost is then

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

This is minimized when $z = \text{mean}(C)$. Moreover, the additional cost incurred by picking $z \neq \text{mean}(C)$ can be characterized very simply:

Lemma 1. *For any set $C \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$,*

$$\mathrm{cost}(C, z) = \mathrm{cost}(C, \mathrm{mean}(C)) + |C| \cdot \|z - \mathrm{mean}(C)\|^2.$$

Contrast this with the case of k-center, where there is no closed-form expression for the optimal center of a cluster. This is a perfect example of the benefit of using average distortion and the squared L_2 norm.

Lemma 1 follows from a generic bias-variance decomposition of random vectors.

Lemma 2. *Let $X \in R^d$ be any random variable. For any $z \in \mathbb{R}^d$,*

$$\mathbb{E}\|X - z\|^2 = \mathbb{E}\|X - \mathbb{E}X\|^2 + \|z - \mathbb{E}X\|^2.$$

Proof. Expanding the right-hand side,

$$\begin{aligned} \mathbb{E}\|X - \mathbb{E}X\|^2 + \|z - \mathbb{E}X\|^2 &= \mathbb{E}\left[\|X\|^2 + \|\mathbb{E}X\|^2 - 2X \cdot \mathbb{E}X\right] + \left[\|z\|^2 + \|\mathbb{E}X\|^2 - 2z \cdot \mathbb{E}X\right] \\ &= \mathbb{E}\|X\|^2 + \|\mathbb{E}X\|^2 - 2\mathbb{E}X \cdot \mathbb{E}X + \|z\|^2 + \|\mathbb{E}X\|^2 - 2z \cdot \mathbb{E}X \\ &= \mathbb{E}\|X\|^2 + \|z\|^2 - 2z \cdot \mathbb{E}X \ = \ \mathbb{E}\|X - z\|^2. \end{aligned}$$

□

To see how this implies Lemma 1, let X denote a uniform random draw from cluster C. Then

$$\mathbb{E}\|X - z\|^2 = \sum_{x \in C} \frac{1}{|C|} \|x - z\|^2 = \frac{1}{|C|} \mathrm{cost}(C, z)$$

and

$$\mathbb{E}\|X - \mathbb{E}X\|^2 = \frac{1}{|C|} \mathrm{cost}(C, \mathrm{mean}(C)).$$

3.2 The k-means algorithm

The name "k-means" is applied both to the clustering task defined above and to a specific algorithm that attempts (with mixed success) to solve it. Here's how the algorithm works, given a data set $S \subset \mathbb{R}^d$ and an integer k:

```
initialize centers $z_1, \ldots, z_k \in \mathbb{R}^d$ and clusters $C_1, \ldots, C_k$ in any way
repeat until there is no further change in cost:
   for each $j$:  $C_j \leftarrow \{x \in S$ whose closest center is $z_j\}$
   for each $j$:  $z_j \leftarrow \mathrm{mean}(C_j)$
```

This is simple enough, and takes $O(k|S|)$ time per iteration. What can one possibly prove about it?

3.2.1 Convergence

Lemma 3. *During the course of the k-means algorithm, the cost monotonically decreases.*

Proof. Let $z_1^{(t)}, \ldots, z_k^{(t)}, C_1^{(t)}, \ldots, C_k^{(t)}$ denote the centers and clusters at the start of the t^{th} iteration of k-means. The first step of the iteration assigns each data point to its closest center; therefore

$$\mathrm{cost}(C_{1:k}^{(t+1)}, z_{1:k}^{(t)}) \leq \mathrm{cost}(C_{1:k}^{(t)}, z_{1:k}^{(t)}).$$

On the second step, each cluster is re-centered at its mean; by Lemma 1,

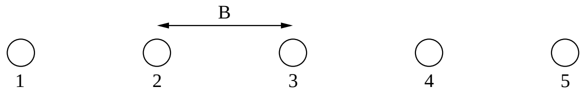
$$\mathrm{cost}(C_{1:k}^{(t+1)}, z_{1:k}^{(t+1)}) \leq \mathrm{cost}(C_{1:k}^{(t+1)}, z_{1:k}^{(t)}).$$

□

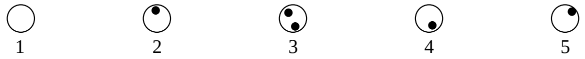
3.2.2 Initialization

We've seen that the k-means algorithm converges to a local optimum of its cost function. The quality of its final clustering depends heavily on the manner of initialization. If this isn't done right, things could go horribly wrong.

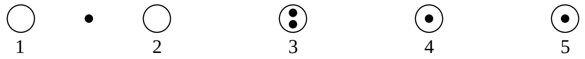
Here's an example. Suppose the data set consists of n points in five tight clusters (of some tiny radius δ) arranged in a line, with some large distance B between them:



The optimal 5-clustering has cost roughly $\delta^2 n$. If we initialize k-means by choosing five centers at random from the data, there is some chance that we'd end up with no centers from cluster 1, two centers from cluster 3, and one center each from clusters 2, 4, and 5:



In the first round of k-means, all points in clusters 1 and 2 will be assigned to the leftmost center. The two centers in cluster 3 will end up sharing that cluster. And the centers in clusters 4 and 5 will move roughly to the centers of those clusters.



Thereafter, no further changes will occur. This local optimum has cost $\Omega(B^2 n)$. We can make this arbitrarily far away from the optimum cost by setting B large enough. Thus, good initialization is crucial.

An interesting problem is to characterize the various kinds of local optima into which the k-means algorithm can fall. A possible starting point is to consider the case of clusters that are well separated from each other.

3.3 The k-means++ initializer

One idea for initializing k-means is to use a farthest-first traversal on the data set, to pick k points that are far away from each other. However, this is too sensitive to outliers. Instead, Arthur and Vassilvitskii (2007) suggest the following procedure, called k-means++: pick the k centers one at a time, but instead of always choosing the point farthest from those picked so far, choose each point at random, with probability proportional to its squared distance from the centers chosen already. Here's the algorithm (given S and k).

```
pick $x \in S$ uniformly at random and set $T \leftarrow \{x\}$
while $|T| < k$:
   pick $x \in S$ at random, with probability proportional to $\mathrm{cost}(x, T) = \min_{z \in T} \|x - z\|^2$
   $T \leftarrow T \cup \{x\}$
```

This initialization takes time $O(k|S|)$, about the same as a single iteration of k-means. Arthur and Vassilvitskii (2007) show that this initialization is itself a pretty good clustering. And subsequent iterations of k-means can only improve things.

Theorem 4. *Let T be the initial centers chosen by k-means++. Let T^* be the optimal centers. Then $\mathbb{E}[\mathrm{cost}(T)] \leq \mathrm{cost}(T^*) \cdot O(\log k)$, where the expectation is over the randomness in the initialization procedure.*

3.3.1 Analysis: preliminaries

Let $T^* = \{z_1, \ldots, z_k\}$ denote the optimal k-means solution, with corresponding clusters $C_1, \ldots, C_k$ (in particular, this implies $z_i = \text{mean}(C_i)$). We will see that T isn't too much worse than this.

One limitation of the centers T is that they are not arbitrary points in $\mathbb{R}^d$ but are constrained to be points from the data set itself. It turns out that in general, one loses at most a factor of two on account of this. The following lemma can be applied to an individual cluster within the optimal k-means clustering.

Lemma 5. *For any $C \subset \mathbb{R}^d$, let $\mu = mean(C)$. If Z is chosen uniformly at random from C, then*

$$\mathbb{E}[\text{cost}(C, Z)] = 2\,\text{cost}(C, \mu).$$

Proof. By Lemma 1,

$$\begin{aligned}\mathbb{E}[\text{cost}(C, Z)] \;&=\; \sum_{z \in C} \frac{1}{|C|} \text{cost}(C, z) = \frac{1}{|C|} \sum_{z \in C} \left(\text{cost}(C, \mu) + |C| \cdot \|z - \mu\|^2\right) \\ &= \text{cost}(C, \mu) + \sum_{z \in C} \|z - \mu\|^2 \;=\; 2\text{cost}(C, \mu).\end{aligned}$$

□

In particular, suppose the *first* center chosen by k-means++ lies in cluster C_i. It is a uniform-random point in C_i, and thus, in expectation, incurs a cost twice that of the optimal center z_i.

Lemma 5 does not apply to subsequent centers chosen by k-means++, since they are not uniform-random draws from their respective clusters. To see this, suppose a few centers T have already been chosen, and the next center z lies in cluster C_i. The sampling distribution biases z towards the region of C_i that is least well-explained by the existing T. Luckily this non-uniformity cannot hurt too much: in expectation, the cost for C_i will be at most 8 times optimal.

Lemma 6. *If some centers T have already been chosen by k-means++ and $Z \in C_i$ is added next, then*

$$\mathbb{E}[\text{cost}(C_i, T \cup \{Z\}) \mid T, \{Z \in C_i\}] \leq 8\,\text{cost}(C_i, z_i).$$

Proof. Let's write out the left-hand expectation in full.

$$\begin{aligned}\mathbb{E}[\text{cost}(C_i, T \cup \{Z\}) \mid T, \{Z \in C_i\}] &= \sum_{z \in C_i} \Pr(\text{Choose } z|T)\,\text{cost}(C_i, T \cup \{z\}) \\ &= \sum_{z \in C_i} \frac{\text{cost}(z, T)}{\text{cost}(C_i, T)} \sum_{x \in C_i} \min(\text{cost}(x, T), \|x - z\|^2)\end{aligned}$$

First let's get an upper bound on $\text{cost}(z, T)$. Pick any $x \in C_i$, and let t be its closest center in T. We have by the triangle inequality that

$$\begin{aligned}\text{cost}(z, T) \leq \|z - t\|^2 &\leq (\|z - x\| + \|x - t\|)^2 \\ &\leq 2\|z - x\|^2 + 2\|x - t\|^2 = 2\|x - z\|^2 + 2\text{cost}(x, T)\end{aligned}$$

where we have used $(A+B)^2 \leq 2A^2 + 2B^2$. Averaging over all $x \in C_i$ yields

$$\mathrm{cost}(z,T) \leq \frac{2}{|C_i|}\sum_{x\in C_i}\|x-z\|^2 + \frac{2}{|C_i|}\sum_{x \in C_i}\mathrm{cost}(x,T) = \frac{2}{|C_i|}\left(\mathrm{cost}(C_i,z)+\mathrm{cost}(C_i,T)\right).$$

Now let's substitute this bound into our earlier expression for the expected value.

$$\begin{aligned}\mathbb{E}[\mathrm{cost}(C_i, T\cup\{Z\}) \mid T, \{Z \in C_i\}] &\leq \sum_{z\in C_i}\frac{\frac{2}{|C_i|}\left(\mathrm{cost}(C_i,T)+\mathrm{cost}(C_i,z)\right)}{\mathrm{cost}(C_i,T)}\sum_{x\in C_i}\min(\mathrm{cost}(x,T), \|x-z\|^2) \\ &\leq \left(\sum_{z\in C_i}\frac{2}{|C_i|}\sum_{x\in C_i}\|x-z\|^2\right) + \left(\sum_{z\in C_i}\frac{2}{|C_i|}\frac{\mathrm{cost}(C_i,z)}{\mathrm{cost}(C_i,T)}\sum_{x\in C_i}\mathrm{cost}(x,T)\right) \\ &= \frac{2}{|C_i|}\sum_{z\in C_i}\mathrm{cost}(C_i,z) + \frac{2}{|C_i|}\sum_{z\in C_i}\mathrm{cost}(C_i,z) \\ &= \frac{4}{|C_i|}\sum_{z\in C_i}\left(\mathrm{cost}(C_i,z_i)+|C_i|\|z-z_i\|^2\right) \; = \; 8\,\mathrm{cost}(C_i,z_i)\end{aligned}$$

where the last line invokes Lemma 1. □

So: the first center we pick blows up the cost of that cluster by a factor of 2, while subsequent centers blow up the costs of the corresponding clusters by a factor of 8. Why is the overall approximation factor $O(\log k)$ instead of 8? Because we may fail to hit all the clusters.

3.3.2 Main analysis

The k-means++ algorithm picks some centers T and incurs $\mathrm{cost}(T)$. Instead of paying this full amount at the end, let's pay it in installments, a little bit on each of the k iterations.

First let's establish some notation. At the end of the tth iteration, t centers have been chosen: call these T_t. Define H_t to be the clusters (of the optimal solution) that have already been "hit," that is, $H_t = \{1 \leq i \leq k : C_i \cap T_t \neq \emptyset\}$. Let $U_t = [k] \setminus H_t$ be the remaining "uncovered" clusters. Also, let W_t denote the number of "wasted" iterations so far, iterations on which a new cluster was not hit; so $W_t = t - |H_t|$. Finally, let $\mathrm{cost}_t(A)$ be a shorthand for $\mathrm{cost}(A, T_t)$, the cost function using the centers at time t.

Here's the payment plan. We'll make sure that by the end of the tth iteration, we have paid a total amount of

$$\mathrm{cost}_t(H_t) + \frac{W_t \mathrm{cost}_t(U_t)}{|U_t|}.$$

Let's look at this briefly. The first term is simply the cost of the clusters we have already hit. As we've seen, in expectation this will be at most 8 times the optimal cost for these clusters. The second term is nonzero when $W_t > 0$. If we have wasted W_t iterations, it means that at the very end, we will have at least W_t uncovered clusters. The average cost of each such cluster can be upper-bounded by roughly $\mathrm{cost}_t(U_t)/|U_t|$.

At the very beginning, when $t = 0$, we have $H_t = \emptyset$ and $W_t = 0$, so no payment is due. At the very end, when $t = k$, we have $W_k = |U_k|$ and thus the final total is $\mathrm{cost}_k(H_k) + \mathrm{cost}_k(U_k) = \mathrm{cost}(T)$, as expected.

As we mentioned, the term $\mathrm{cost}_t(H_t)$ is easy to bound. The following is an immediate consequence of Lemma 6.

Lemma 7. *For any $t \leq k$, we have $\mathbb{E}[\mathrm{cost}_t(H_t)] \leq 8\,\mathrm{cost}(T^*)$.*

So let's focus upon the second term in the payment,

$$\Psi_t = \frac{W_t \text{cost}_t(U_t)}{|U_t|}.$$

We'll bound the expected increase in this amount from iteration t to $t+1$, that is, $\mathbb{E}[\Psi_{t+1} - \Psi_t]$.

Suppose that the $(t+1)$st center lies in cluster C_I. We'll consider two cases: when this is a new (uncovered) cluster, and when it is a cluster that has previously been hit. The first case is the desirable situation, and produces no additional charge (in expectation). In what follows, let $\mathcal{F}_t$ denote the history of all random events upto and including iteration t.

Lemma 8. *Suppose that in iteration $t+1$, center $Z \in C_I$ is chosen.*

$$\mathbb{E}[\Psi_{t+1} - \Psi_t \mid \mathcal{F}_t, \{I \in U_t\}] \leq 0.$$

Proof. When $I \in U_t$, we have $H_{t+1} = H_t \cup \{I\}$, $W_{t+1} = W_t$, and $U_{t+1} = U_t \setminus \{I\}$. Hence

$$\Psi_{t+1} = \frac{W_{t+1}\text{cost}_{t+1}(U_{t+1})}{|U_{t+1}|} \leq \frac{W_t(\text{cost}_t(U_t) - \text{cost}_t(C_I))}{|U_t| - 1}.$$

Let's bound $\text{cost}_t(C_I)$, for I randomly chosen from U_t.

$$\mathbb{E}[\text{cost}_t(C_I)|\mathcal{F}_t, \{I \in U_t\}] \;=\; \sum_{i \in U_t} \frac{\text{cost}_t(C_i)}{\text{cost}_t(U_t)}\text{cost}_t(C_i) \;\geq\; \frac{\text{cost}_t(U_t)}{|U_t|}$$

using the Cauchy-Schwarz inequality. Thus

$$\begin{aligned}\mathbb{E}[\Psi_{t+1}|\mathcal{F}_t, \{I \in U_t\}] &\leq \frac{W_t}{|U_t| - 1}\left(\text{cost}_t(U_t) - \mathbb{E}[\text{cost}_t(C_I)|\mathcal{F}_t, \{I \in U_t\}]\right) \\ &\leq \frac{W_t}{|U_t| - 1}\left(\text{cost}_t(U_t) - \frac{\text{cost}_t(U_t)}{|U_t|}\right) \;=\; \Psi_t.\end{aligned}$$

□

Now, let's move to the bad case, when the $(t+1)$st center lies in a cluster C_I that has already been hit: that is, $I \in H_t$.

Lemma 9. *If $I \in H_t$, then $\Psi_{t+1} - \Psi_t \leq \text{cost}_t(U_t)/|U_t|$.*

Proof. When $I \in H_t$, we have $H_{t+1} = H_t$, $W_{t+1} = W_t + 1$, and $U_{t+1} = U_t$. Thus

$$\Psi_{t+1} - \Psi_t \;=\; \frac{W_{t+1}\text{cost}_{t+1}(U_{t+1})}{|U_{t+1}|} - \frac{W_t\text{cost}_t(U_t)}{|U_t|} \;\leq\; \frac{(W_t+1)\text{cost}_t(U_t)}{|U_t|} - \frac{W_t\text{cost}_t(U_t)}{|U_t|} \;=\; \frac{\text{cost}_t(U_t)}{|U_t|}.$$

□

Putting these two cases together gives the expected additional payment on round $t+1$.

Lemma 10. *For any $t \geq 0$, we have $\mathbb{E}[\Psi_{t+1} - \Psi_t|\mathcal{F}_t] \leq \text{cost}_t(H_t)/(k-t)$.*

Proof. Using Lemmas 8 and 9, we have

$$\begin{aligned}\mathbb{E}[\Psi_{t+1} - \Psi_t | \mathcal{F}_t] &= \mathbb{E}[\Psi_{t+1} - \Psi_t | \mathcal{F}_t, \{I \in U_t\}]\Pr(I \in U_t|\mathcal{F}_t) + \mathbb{E}[\Psi_{t+1} - \Psi_t | \mathcal{F}_t, \{I \in H_t\}]\Pr(I \in H_t|\mathcal{F}_t) \\ &\leq 0 + \frac{\text{cost}_t(U_t)}{|U_t|} \cdot \frac{\text{cost}_t(H_t)}{\text{cost}(T_t)} \;\leq\; \frac{\text{cost}_t(H_t)}{|U_t|} \;\leq\; \frac{\text{cost}_t(H_t)}{k-t}.\end{aligned}$$

□

Adding up these expected installments gives the overall payment.

Theorem 11. *If T are the centers returned by the k-means++ algorithm, then*

$$\mathbb{E}[\text{cost}(T)] \leq 8(2 + \ln k)\text{cost}(T^*).$$

Proof. Using $\text{cost}(T) = \text{cost}_k(H_k) + \text{cost}_k(U_k) = \text{cost}_k(H_k) + \Psi_k$ and Lemmas 7 and 10, we get

$$\begin{aligned}\mathbb{E}[\text{cost}(T)] &= \mathbb{E}[\text{cost}_k(H_k)] + \sum_{t=0}^{k-1} \mathbb{E}[\Psi_{t+1} - \Psi_t] \\ &\leq \mathbb{E}[\text{cost}_k(H_k)] + \sum_{t=0}^{k-1} \frac{\mathbb{E}[\text{cost}_t(H_t)]}{k-t} \leq 8\text{cost}(T^*)\left(1 + 1 + \frac{1}{2} + \cdots + \frac{1}{k}\right).\end{aligned}$$

The harmonic sum can upper-bounded by $1 + \frac{1}{2} + \cdots + \frac{1}{k} \leq 1 + \int_1^k \frac{1}{x} dx = 1 + \ln k$. □

3.4 Discussion

There are some constant-factor approximation algorithms—for instance, the local search method of Kanungo et al. (2004)—but these haven't been as popular in practice as using k-means++ as an initializer for the regular k-means algorithm. It is also unclear what factors are achievable, since there don't seem to be any hardness of approximation results.

Bibliography

Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*.

Dasgupta, S. (2008). The hardness of k-means clustering. *UCSD CSE Technical Report 2008-0916*.

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., and Wu, A. (2004). A local search approximation algorithm for k-means clustering. *Computational Geometry: Theory and Applications*, 28:89–112.

Mahajan, M., Nimbhorkar, P., and Varadarajan, K. (2012). The planar k-means problem is NP-hard. *Theoretical Computer Science*, 442:13–21.

Vattani, A. (2009). The hardness of k-means clustering in the plane.