k-means++ seeding

- Have seen that the k-means algorithm can output arbitrarily poor solutions, if started with a bad set of initial centroids
- k-means++ is a simple, probabilistic algorithm to compute initial centroids
- These centroids are already a reasonably good solution for the k-problem (provably)
- In practice, combining k-means++ seeding wit a few rounds of the k-means algorithm usually leads to very good solutions to the k-means problem.

Notation

- D denotes the squared Euclidean distance, $P \subset \mathbb{R}^d, |P| < \infty$
- $x \in \mathbb{R}^d, C \subset \mathbb{R}^d, |C| < \infty, D(x, C) := \min_{c \in C} D(x, c)$
- $A \subseteq P : D(A, C) := \sum_{a \in A} D(a, C)$
- $C, |C| = k$, set of centroids with corresponding set of clusters $\mathcal{C} = \{C_1, \ldots, C_k\}$, both simply called clustering.
- For $A \subseteq P$ denote by $D_{\text{opt}}(A) := D(A, C_{\text{opt}}), C_{\text{opt}} :=$ optimal k-clustering, the contribution of A to the cost of an optimal clustering.
- Write $\text{cost}_k(P)$ instead of $\text{cost}_k^D(P)$.
- If $A \in C_{\text{opt}}$, then $D_{\text{opt}}(A) = cost_1(A)$.

k-means++ seeding - distribution

k-means++ distribution

For any set $C \subset \mathbb{R}^d, |C| < \infty$, denote by $p_C(\cdot)$ the distribution on P defined by

$$\forall p \in P : p_C(p) := \frac{D(p, C)}{D(P, C)}$$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

k-means++ seeding - algorithm

K-MEANS++(P, k)

choose $c \in P$ uniformly at random, $C := \{c\}$;

repeat

chosse $c \in P$ according to distribution $p_c(\cdot)$;

$C := C \cup \{c\}$;

until $|C| = k$;

run K-MEANS on P with initial centers C;

return C;

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Theorem 4.1

For any finite set of points $P \subset \mathbb{R}^d$ and any $k \in \mathbb{N}$, algorithm k-Means++ *computes a k-clustering C of P such that*

$$E[D(P, C)] \leq 8 \cdot (2 + \ln k) \cdot opt_k(P).$$

k-means++ seeding - main lemmas

Lemma 4.2

Let $A \subseteq P$ be a cluster of C_{opt}. If $a \in A$ is chosen uniformly at random from P, then

$$E[D(A, \{a\}) | a \in A] = 2 \cdot D_{opt}(A).$$

k-means++ seeding - main lemmas

Lemma 4.2

Let $A \subseteq P$ be a cluster of C_{opt}. If $a \in A$ is chosen uniformly at random from P, then

$$E[D(A, \{a\})|a \in A] = 2 \cdot D_{opt}(A).$$

Lemma 4.3

Let $A \subseteq P$ be a cluster of C_{opt} and let $C, |C| < k$, be arbitrary. If a is chosen according to $p_C(\cdot)$, then

$$E\left[D(A, C \cup \{a\})|a \in A\right] \leq 8 \cdot D_{opt}(A).$$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Lemma 4.4

Let $0 < u < k, 0 \leq t \leq u$. Let P^u be the union of u different clusters of C_{opt} and set $P^c := P \setminus P^u$. Finally, let $B \subseteq P^c$ and set $C_0 := B$ and $C_j := C_{j-1} \cup \{a_j\}, j = 1, \ldots, t$, where a_j is chosen according to $p_{C_{j-1}}$. Then

$$E\left[D(P, C_t)\right] \leq (1 + H_t)\left(D(P^c, B) + 8 \cdot D_{opt}(P^u)\right) + \frac{u-t}{u} \cdot D(P^u, B),$$

where $H_t = \sum_{i=1}^{t} \frac{1}{i}$.