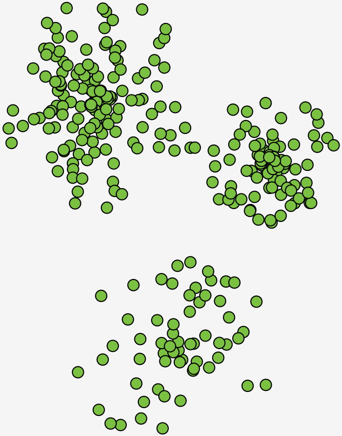
Dimensionality reductions for k-means

Melanie Schmidt

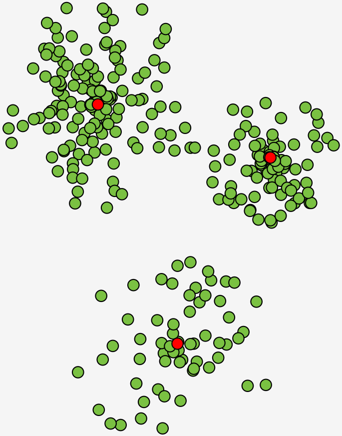
19.05.2015

The *k*-means problem



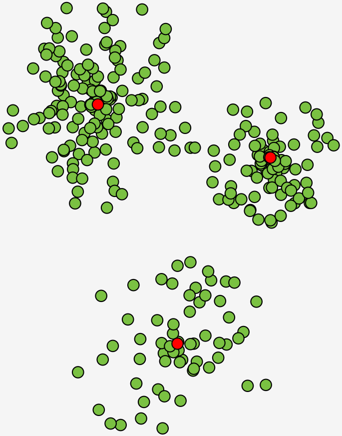
The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,



The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,
- compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers

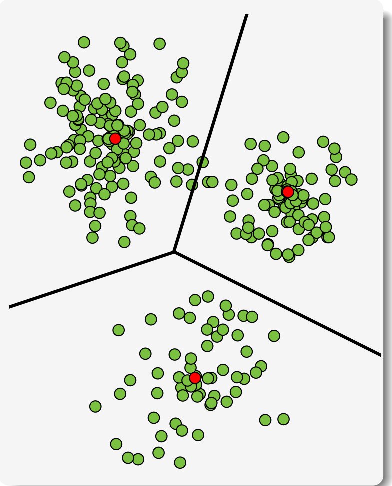


The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,
- compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers
- which minimizes cost(P, C)

$$= \sum_{p \in P} \min_{c \in C} ||p - c||^2,$$

the sum of the squared distances.



The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,
- compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers
- which minimizes cost(P, C)

$$= \sum_{p \in P} \min_{c \in C} ||p - c||^2,$$

the sum of the squared distances.

- induces a partitioning of the input point set

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957
- various algorithms for the *k*-means problem

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the *k*-means problem

k-means is still widely studied, new theoretical insights

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03], now $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03], now $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]
- k-means++ [AV07]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03], now $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]
- k-means++ [AV07]

Hardness of k-means

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03], now $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]
- k-means++ [AV07]

Hardness of k-means

- k-means is NP-hard for $k = 2$ [ADHP09]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03], now $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]
- k-means++ [AV07]

Hardness of k-means

- k-means is NP-hard for $k = 2$ [ADHP09], also for $d = 2$ [MNV09]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03], now $\mathcal{O}(nd + 2^{\mathrm{poly}(1/\varepsilon,k)})$ [FL11]
- k-means++ [AV07]

Hardness of k-means

- k-means is NP-hard for $k = 2$ [ADHP09], also for $d = 2$ [MNV09]
- it is NP-hard to c-approximate for a small $c > 1$ [ACKS15]

k-means is studied since the 50s

- defined around 1950
- Lloyd's algorithm in 1957, 2500 new citations since 2011
- various algorithms for the k-means problem

k-means is still widely studied, new theoretical insights

- Analysis of Lloyd's algorithm (running time) in [AV06,V11]
- constant approximation algorithms [JV01],[KM+04]
- PTAS for constant k [M00,dlV+03], now $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon, k)})$ [FL11]
- k-means++ [AV07]

Hardness of k-means

- k-means is NP-hard for $k = 2$ [ADHP09], also for $d = 2$ [MNV09]
- it is NP-hard to c-approximate for a small $c > 1$ [ACKS15]
- ($c \geq 1.001418$)

High dimensional data

- Assume that d is much larger than k
- Do we need to solve a d-dimensional problem?

High dimensional data

- Assume that d is much larger than k
- Do we need to solve a d-dimensional problem?

Some answers

- [JL84] $\mathcal{O}(\varepsilon^{-2} \log n)$ dimensions suffice for a $(1+\varepsilon)$-approximation
- [DF+99] k dimensions suffice for a 2-approximation

High dimensional data

- Assume that d is much larger than k
- Do we need to solve a d-dimensional problem?

Some answers

- [JL84] $\mathcal{O}(\varepsilon^{-2} \log n)$ dimensions suffice for a $(1+\varepsilon)$-approximation
- [DF+99] k dimensions suffice for a 2-approximation

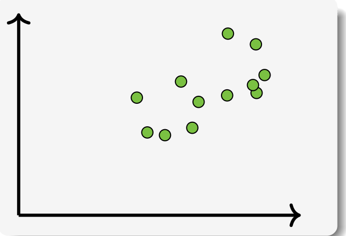
How many dimensions do we need to approximately solve k-means?

Dimensionality reduction

Replace P by a point set Q of smaller intrinsic dimension

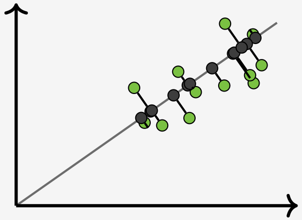
Dimensionality reduction

Replace P by a point set Q of smaller intrinsic dimension



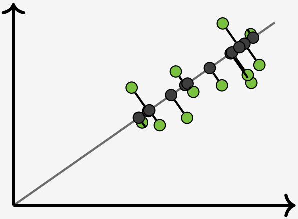
Dimensionality reduction

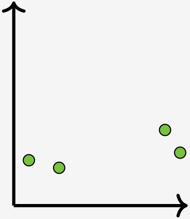
Replace P by a point set Q of smaller intrinsic dimension



Dimensionality reduction

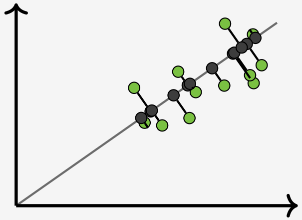
Replace P by a point set Q of smaller intrinsic dimension

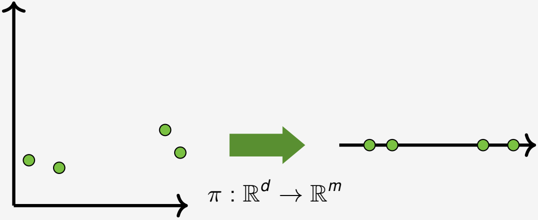




Dimensionality reduction

Replace P by a point set Q of smaller intrinsic dimension

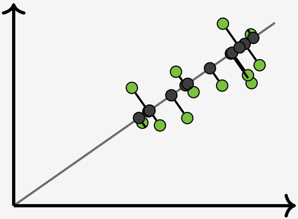


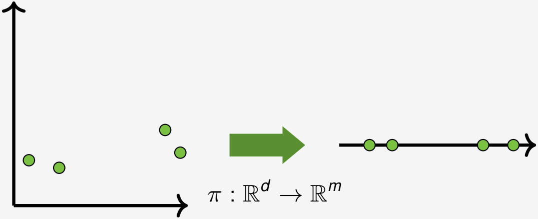


$\pi : \mathbb{R}^d \rightarrow \mathbb{R}^m$

Dimensionality reduction

Replace P by a point set Q of smaller intrinsic dimension





Dimensionality reduction

$P \subset \mathbb{R}^d$ is replaced by $Q \subset \mathbb{R}^d$ of smaller intrinsic dimension such that

$$|\mathrm{cost}(Q, C) - \mathrm{cost}(P, C)| \leq \varepsilon \cdot \mathrm{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

In the following

In the following

- Facts on k-means and JL result

In the following

- Facts on k-means and JL result
- Joint work with Dan Feldman and Christian Sohler

In the following

- Facts on k-means and JL result
- Joint work with Dan Feldman and Christian Sohler
- STOC '15 paper due to Cohen et. al.

Fact 1 [Foklore?]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

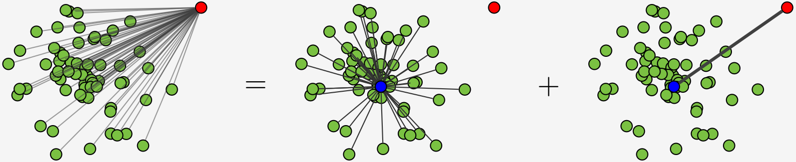
where $\mu(P) = \sum_{x \in P} x / |P|$ is the centroid of P.

Fact 1 [Foklore?]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.

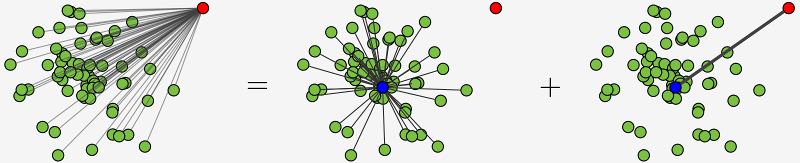


Fact 1 [Foklore?]

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.



Implications

- centroid is always the optimal 1-means solution
- optimal solution consists of centroids of subsets

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.

Corollary (Fact 2)

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x - y||^2.$$

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.

Corollary (Fact 2)

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x - y||^2.$$

$$\sum_{z \in P} \sum_{x \in P} ||x - z||^2$$

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x / |P|$ is the centroid of P.

Corollary (Fact 2)

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x - y||^2.$$

$$\sum_{z \in P} \sum_{x \in P} ||x - z||^2 = \sum_{z \in P}$$

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.

Corollary (Fact 2)

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x - y||^2.$$

$$\sum_{z \in P} \sum_{x \in P} ||x - z||^2 = \sum_{z \in P} \left(\sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2 \right)$$

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x / |P|$ is the centroid of P.

Corollary (Fact 2)

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x - y||^2.$$

$$\sum_{z \in P} \sum_{x \in P} ||x - z||^2 = \sum_{z \in P} \Big(\sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2 \Big)$$

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2,$$

where $\mu(P) = \sum_{x \in P} x/|P|$ is the centroid of P.

Corollary (Fact 2)

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x - y||^2.$$

$$\sum_{z \in P} \sum_{x \in P} ||x - z||^2 = \sum_{z \in P} \Big(\sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2 \Big)$$
$$= |P| \sum_{x \in P} ||x - \mu(P)||^2 + |P| \sum_{z \in P} \cdot ||\mu(P) - z||^2$$

Magic formula

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2$$

and $\mu(P) = \sum_{x \in P} x/|P|$ is the optimal 1-means solution.

Magic formula

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2$$

and $\mu(P) = \sum_{x \in P} x/|P|$ is the optimal 1-means solution.

Corollary

The optimal k-means solution consists of centroids of subsets of P.

Magic formula

It holds for any $P \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$ that

$$\sum_{x \in P} ||x - z||^2 = \sum_{x \in P} ||x - \mu(P)||^2 + |P| \cdot ||\mu(P) - z||^2$$

and $\mu(P) = \sum_{x \in P} x/|P|$ is the optimal 1-means solution.

Corollary

The optimal k-means solution consists of centroids of subsets of P.

Corollary

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x - y||^2.$$

Johnson, Lindenstrauss, 1984

Given $\varepsilon \in (0, 1)$, there is an $r \in \mathcal{O}(\varepsilon^{-2} \log n)$ and a linear map $\pi : \mathbb{R}^d \to \mathbb{R}^r$ such that for all $x, y \in P$:

$$(1 - \varepsilon)||x - y||^2 < ||\pi(x) - \pi(y)||^2 < (1 + \varepsilon)||x - y||^2.$$

Such a map can be found in randomized polynomial time.

Johnson, Lindenstrauss, 1984

Given $\varepsilon \in (0,1)$, there is an $r \in \mathcal{O}(\varepsilon^{-2} \log n)$ and a linear map $\pi : \mathbb{R}^d \to \mathbb{R}^r$ such that for all $x, y \in P$:

$$(1-\varepsilon)||x-y||^2 < ||\pi(x) - \pi(y)||^2 < (1+\varepsilon)||x-y||^2.$$

Such a map can be found in randomized polynomial time.

Corollary

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x - \mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x-y||^2.$$

Johnson, Lindenstrauss, 1984

Given $\varepsilon \in (0,1)$, there is an $r \in \mathcal{O}(\varepsilon^{-2} \log n)$ and a linear map $\pi : \mathbb{R}^d \to \mathbb{R}^r$ such that for all $x, y \in P$:

$$(1-\varepsilon)||x-y||^2 < ||\pi(x)-\pi(y)||^2 < (1+\varepsilon)||x-y||^2.$$

Such a map can be found in randomized polynomial time.

Corollary

The optimal 1-means cost of any $P \subset \mathbb{R}^d$ is given by

$$\sum_{x \in P} ||x-\mu(P)||^2 = \frac{1}{2|P|} \sum_{x \in P} \sum_{y \in P} ||x-y||^2.$$

Implies dimensionality reduction for k-means with $r \in \mathcal{O}(\varepsilon^{-2} \log n)$.

Johnson, Lindenstrauss, 1984

Given $\varepsilon \in (0,1)$, there is an $r \in \mathcal{O}(\varepsilon^{-2} \log n)$ and a linear map $\pi : \mathbb{R}^d \to \mathbb{R}^r$ such that for all $x, y \in P$:

$$(1-\varepsilon)||x-y||^2 < ||\pi(x)-\pi(y)||^2 < (1+\varepsilon)||x-y||^2.$$

Such a map can be found in randomized polynomial time.

Lower Bound for JL-type results: Larsen, Nelson, 2014

For any $d > 1$ and $\varepsilon \in (0, 1/2)$, there is a point set $X \subset \mathbb{R}^d$ such that

- $|X| = d^{O(1)}$
- if a linear $\pi : \mathbb{R}^d \to \mathbb{R}^r$ provides the JL guarantee for X, then $r \in \Omega(\min\{d, \varepsilon^{-2} \log n\})$

Implies dimensionality reduction for k-means with $r \in \mathcal{O}(\varepsilon^{-2} \log n)$.

Lower Bound of $\Omega(\varepsilon^{-2} \log n)$ for JL-type results

$\rightsquigarrow$ Is this a lower bound for the k-means problem, too?

Lower Bound of $\Omega(\varepsilon^{-2} \log n)$ for JL-type results

$\rightsquigarrow$ Is this a lower bound for the k-means problem, too?

No!

But dimensionality reduction must not preserve pairwise distances!

Lower Bound of $\Omega(\varepsilon^{-2} \log n)$ for JL-type results

$\rightsquigarrow$ Is this a lower bound for the k-means problem, too?

No!

But dimensionality reduction must not preserve pairwise distances!

Recall: k-means cost function

$$\text{cost}(P, C) = \sum_{p \in P} \min_{c \in C} ||p - c||^2$$

Dimensionality reduction

$P \subset \mathbb{R}^d$ is replaced by $Q \subset \mathbb{R}^d$ of smaller intrinsic dimension such that

$$|\text{cost}(Q, C) - \text{cost}(P, C)| \leq \varepsilon \cdot \text{cost}(P, C)$$

holds for all sets $C \subset \mathbb{R}^d$ of k centers.

Idea

Use the Singular Value Decomposition!

Idea

Use the Singular Value Decomposition!

SVD-based results for k-means

- [Drineas, Frieze, Kannan, Vempala, Vinay, 1999]
  2-approximation algorithm that projects to k dimensions by SVD
- [McSherry, 2001], [Awashti, Sheffet, 2014]
  4-guarantee with k dimensions based on SVD

Idea

Use the Singular Value Decomposition!

SVD-based results for k-means

- [Drineas, Frieze, Kannan, Vempala, Vinay, 1999]
  2-approximation algorithm that projects to k dimensions by SVD
- [McSherry, 2001], [Awashti, Sheffet, 2014]
  4-guarantee with k dimensions based on SVD

More precise idea

Project to more than k dimensions based on SVD!

Idea

Use the Singular Value Decomposition!

SVD-based results for k-means

- [Drineas, Frieze, Kannan, Vempala, Vinay, 1999]
  2-approximation algorithm that projects to k dimensions by SVD
- [McSherry, 2001], [Awashti, Sheffet, 2014]
  4-guarantee with k dimensions based on SVD
- [Boutsidis, Mahoney, Drineas, 2009]
  $(2+\varepsilon)$-guarantee with $\tilde{\Theta}(k/\varepsilon^2)$ dimensions (SVD+sampling)

More precise idea

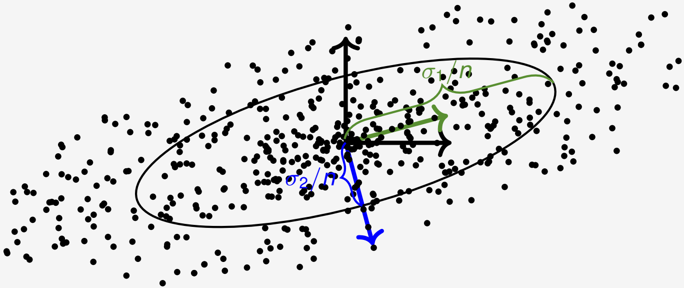
Project to more than k dimensions based on SVD!

Utilizing the Singular Value Decomposition (SVD)

- singular vectors $v_1, \ldots, v_d$, form a basis
- ordered according to singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$

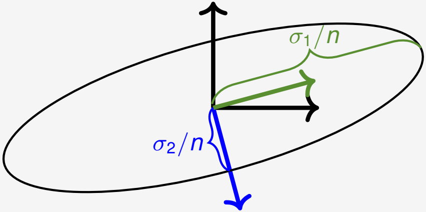
Utilizing the Singular Value Decomposition (SVD)

- singular vectors $v_1, \ldots, v_d$, form a basis
- ordered according to singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$



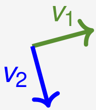
Utilizing the Singular Value Decomposition (SVD)

- singular vectors $v_1, \ldots, v_d$, form a basis
- ordered according to singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$



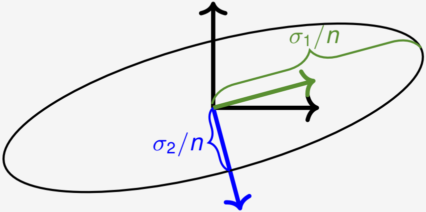
Utilizing the Singular Value Decomposition (SVD)

- singular vectors $v_1, \ldots, v_d$, form a basis
- ordered according to singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$



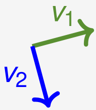
$$\sigma_i^2 = \sum_{x \in P} (x^t v_i)^2$$

$$\sum_{i=1}^{r} \sigma_i^2 = \sum_{x \in P} ||x||^2$$



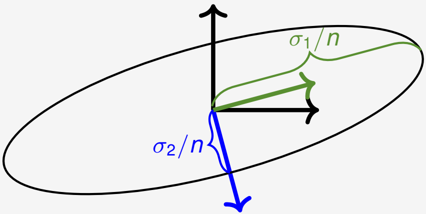
Utilizing the Singular Value Decomposition (SVD)

- singular vectors $v_1, \ldots, v_d$, form a basis
- ordered according to singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$



$$\sigma_i^2 = \sum_{x \in P} (x^t v_i)^2$$

$$\sum_{i=1}^{r} \sigma_i^2 = \sum_{x \in P} ||x||^2$$



SVD-based projections

$\rightsquigarrow$ Project to the span of the first m singular vectors, V_m.

Deal with an easier problem first

$\rightsquigarrow$ Subspace Approximation

Deal with an easier problem first

$\rightsquigarrow$ Subspace Approximation

The Subspace Approximation Problem

Given $P \subset \mathbb{R}^d$, find a k-dimensional subspace V that minimizes

$$\sum_{x \in P} ||x - \pi_V(x)||^2$$

where $\pi_V(x)$ is the perpendicular projection of x to V.

Deal with an easier problem first

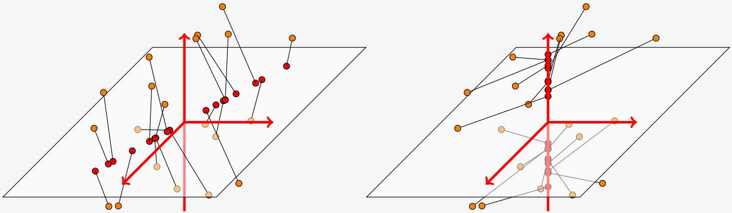
$\rightsquigarrow$ Subspace Approximation

The Subspace Approximation Problem

Given $P \subset \mathbb{R}^d$, find a k-dimensional subspace V that minimizes

$$\sum_{x \in P} ||x - \pi_V(x)||^2$$

where $\pi_V(x)$ is the perpendicular projection of x to V.



This talk is not about optimizing cost functions!

This talk is not about optimizing cost functions!

If we wanted to solve the subspace approximation problem...

The span of the first k singular vectors V_k is the optimal solution!

This talk is not about optimizing cost functions!

If we wanted to solve the subspace approximation problem...

The span of the first k singular vectors V_k is the optimal solution!

Dimensionality reduction for subspace approximation

$P \subset \mathbb{R}^d$ is replaced by $Q \subset \mathbb{R}^d$ of smaller intrinsic dimension such that

$$\left| \sum_{y \in Q} ||y - \pi_V(y)||^2 \qquad - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

holds for all k-dimensional subspaces V.

This talk is not about optimizing cost functions!

If we wanted to solve the subspace approximation problem...

The span of the first k singular vectors V_k is the optimal solution!

Dimensionality reduction for subspace approximation

$P \subset \mathbb{R}^d$ is replaced by $Q \subset \mathbb{R}^d$ of smaller intrinsic dimension such that

$$\left| \sum_{y \in Q} ||y - \pi_V(y)||^2 + \Delta - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

holds for all k-dimensional subspaces V.

This talk is not about optimizing cost functions!

If we wanted to solve the subspace approximation problem...

The span of the first k singular vectors V_k is the optimal solution!

Dimensionality reduction for subspace approximation

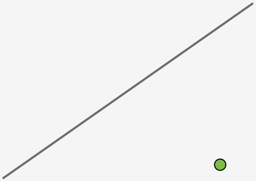
$P \subset \mathbb{R}^d$ is replaced by $Q \subset \mathbb{R}^d$ of smaller intrinsic dimension such that

$$\left| \sum_{y \in Q} ||y - \pi_V(y)||^2 + \Delta - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

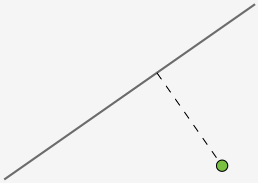
holds for all k-dimensional subspaces V.

$\rightsquigarrow$ want to provide an oracle that can answer subpace queries

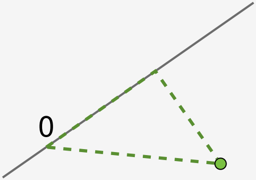
What is the squared distance between a subspace and a point?



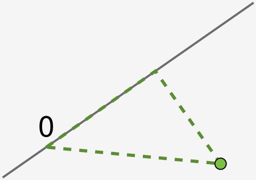
What is the squared distance between a subspace and a point?



What is the squared distance between a subspace and a point?

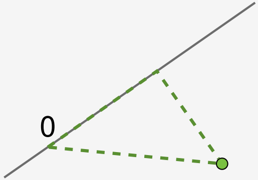


What is the squared distance between a subspace and a point?



$$||x - \pi_V(x)||^2 = ||x||^2 - ||\pi_V(x)||^2$$

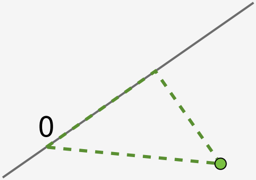
What is the squared distance between a subspace and a point?



$$||x - \pi_V(x)||^2 = ||x||^2 - ||\pi_V(x)||^2$$

- gets closer to $||x||^2$ if k is small compared to d
- subspace 'chooses' k directions where the length is disregarded

What is the squared distance between a subspace and a point?

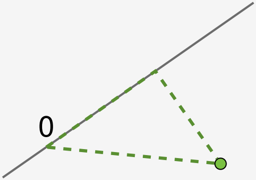


$$||x - \pi_V(x)||^2 = ||x||^2 - ||\pi_V(x)||^2$$

- gets closer to $||x||^2$ if k is small compared to d
- subspace ‘chooses’ k directions where the length is disregarded

- First idea: Just say $\sum_{x \in P} ||x||^2$!

What is the squared distance between a subspace and a point?

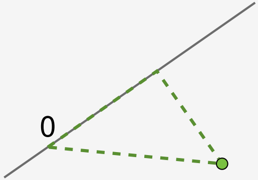


$$||x - \pi_V(x)||^2 = ||x||^2 - ||\pi_V(x)||^2$$

- gets closer to $||x||^2$ if k is small compared to d
- subspace ‘chooses’ k directions where the length is disregarded

- First idea: Just say $\sum_{x \in P} ||x||^2$!
- Problem: P lies within k dimensions $\rightarrow$ true answer can be 0

What is the squared distance between a subspace and a point?

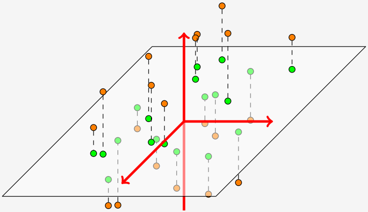


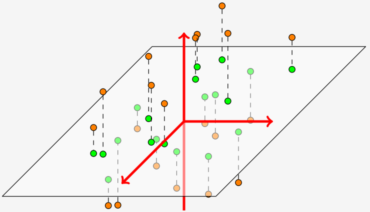
$$||x - \pi_V(x)||^2 = ||x||^2 - ||\pi_V(x)||^2$$

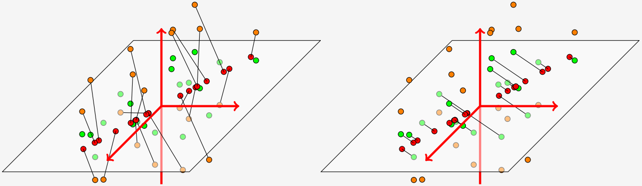
- gets closer to $||x||^2$ if k is small compared to d
- subspace ‘chooses’ k directions where the length is disregarded

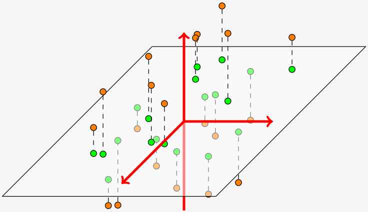
- First idea: Just say $\sum_{x \in P} ||x||^2$!
- Problem: P lies within k dimensions $\rightarrow$ true answer can be 0

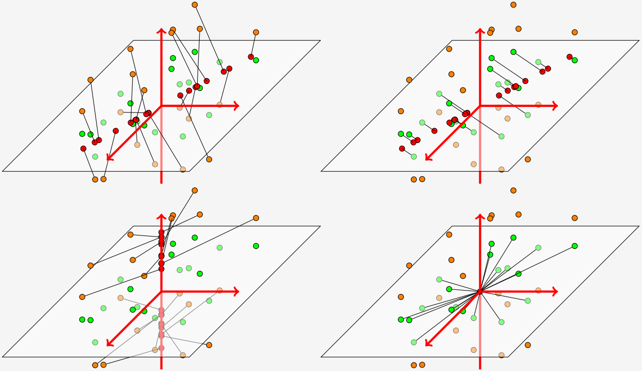
- Second idea: Store most important dimensions and lost length!
- $\rightsquigarrow$ Project points to V_m for some nice m, set $\Delta := \sum_{i=m+1}^{r} \sigma_i^2$.



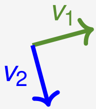






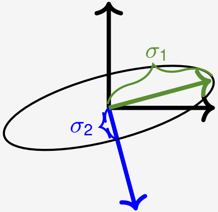


The Singular Value Decomposition (SVD)

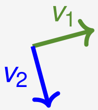


$$\sigma_i^2 = \sum_{x \in P} (x^t v_i)^2$$

$$\sum_{i=1}^{r} \sigma_i^2 = \sum_{x \in P} ||x||^2$$

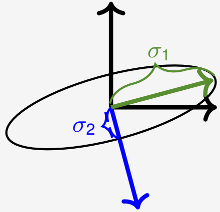


The Singular Value Decomposition (SVD)



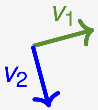
$$\sigma_i^2 = \sum_{x \in P} (x^t v_i)^2$$

$$\sum_{i=1}^{r} \sigma_i^2 = \sum_{x \in P} ||x||^2$$



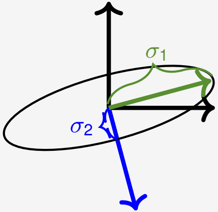
- distance to subspace gets closer to $||x||^2$ if k is small compared to d
- subspace ‘chooses’ k directions where the length is disregarded

The Singular Value Decomposition (SVD)



$$\sigma_i^2 = \sum_{x \in P} (x^t v_i)^2$$

$$\sum_{i=1}^{r} \sigma_i^2 = \sum_{x \in P} ||x||^2$$



- distance to subspace gets closer to $||x||^2$ if k is small compared to d
- subspace ‘chooses’ k directions where the length is disregarded

Assumption *for this talk*

Query subspace is spanned by singular vectors

Dimensionality reduction

Project P to V_m, store $\sum_{i=m+1}^{r} \sigma_i^2$!

Task: Report distance to a given query subspace

- Query subspace 'disregards' length in k directions
- we want to report $\sum ||x||^2 -$ *disregarded length*

Dimensionality reduction

Project P to V_m, store $\sum_{i=m+1}^{r} \sigma_i^2$!

Task: Report distance to a given query subspace

- Query subspace 'disregards' length in k directions
- we want to report $\sum ||x||^2 -$ *disregarded length*

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \dots \sigma_k^2 \quad \sigma_{k+1}^2 \dots \sigma_{2k}^2 \dots \sigma_m^2 \quad \sigma_{m+1}^2 \dots \sigma_{m+k}^2 \dots \sigma_{r-1}^2 \quad \sigma_r^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

Dimensionality reduction

Project P to V_m, store $\sum_{i=m+1}^{r} \sigma_i^2$!

Task: Report distance to a given query subspace

- Query subspace 'disregards' length in k directions
- we want to report $\sum ||x||^2 -$ *disregarded length*

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_{r-1}^2 \quad \sigma_r^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

Dimensionality reduction

Project P to V_m, store $\sum_{i=m+1}^{r} \sigma_i^2$!

Task: Report distance to a given query subspace

- Query subspace 'disregards' length in k directions
- we want to report $\sum ||x||^2 -$ *disregarded length*

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_{r-1}^2 \quad \sigma_r^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded

Dimensionality reduction

Project P to V_m, store $\sum_{i=m+1}^{r} \sigma_i^2$!

Task: Report distance to a given query subspace

- Query subspace ‘disregards’ length in k directions
- we want to report $\sum ||x||^2 -$ *disregarded length*

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_{r-1}^2 \quad \sigma_r^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded $\leq \sum_{i=m+1}^{m+k} \sigma_i^2$

Dimensionality reduction

Project P to V_m, store $\sum_{i=m+1}^{r} \sigma_i^2$!

Task: Report distance to a given query subspace

- Query subspace ‘disregards’ length in k directions
- we want to report $\sum ||x||^2 -$ *disregarded length*

$$\sigma_1^2 \quad \sigma_2^2 \quad \sigma_3^2 \ldots \sigma_k^2 \quad \sigma_{k+1}^2 \ldots \sigma_{2k}^2 \ldots \sigma_m^2 \quad \sigma_{m+1}^2 \ldots \sigma_{m+k}^2 \ldots \sigma_{r-1}^2 \quad \sigma_r^2$$

- we report $\sum_{i=m+1}^{d} \sigma_i^2$ plus correct contribution of first m
- Error: Dimensions we report but are disregarded $\leq \sum_{i=m+1}^{m+k} \sigma_i^2$

Core idea

Make m large enough such that $\sigma_{m+1}^2 + \ldots + \sigma_{m+k}^2$
is small compared to $\sigma_{k+1}^2 + \sigma_2^2 \ldots + \ldots + \sigma_r^2$! $\quad \rightarrow m \geq \lceil k/\varepsilon \rceil$

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil k/\varepsilon \rceil$, there exists a Q with intrinsic dimension $\lceil k/\varepsilon \rceil$ and a constant Δ such that

$$\left| \sum_{x \in Q} ||y - \pi_V(y)||^2 + \Delta - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

holds for all k-dimensional subspaces V.

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil k/\varepsilon \rceil$, there exists a Q with intrinsic dimension $\lceil k/\varepsilon \rceil$ and a constant Δ such that

$$\left| \sum_{x \in Q} ||y - \pi_V(y)||^2 + \Delta - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

holds for all k-dimensional subspaces V.

- Q is the projection of P to V_m with $m = \lceil k/\varepsilon \rceil$

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil k/\varepsilon \rceil$, there exists a Q with intrinsic dimension $\lceil k/\varepsilon \rceil$ and a constant Δ such that

$$\left| \sum_{x \in Q} ||y - \pi_V(y)||^2 + \Delta - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

holds for all k-dimensional subspaces V.

- Q is the projection of P to V_m with $m = \lceil k/\varepsilon \rceil$ $\quad A_m$

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil k/\varepsilon \rceil$, there exists a Q with intrinsic dimension $\lceil k/\varepsilon \rceil$ and a constant Δ such that

$$\left| \sum_{x \in Q} ||y - \pi_V(y)||^2 + \Delta - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

holds for all k-dimensional subspaces V.

- Q is the projection of P to V_m with $m = \lceil k/\varepsilon \rceil$ A_m
- Δ is the lost squared length $\sum_{i=m+1}^{r} \sigma_i^2$

Theorem

For any $P \in \mathbb{R}^d$, k, $\varepsilon \in (0,1)$, $n, d \geq k + \lceil k/\varepsilon \rceil$, there exists a Q with intrinsic dimension $\lceil k/\varepsilon \rceil$ and a constant Δ such that

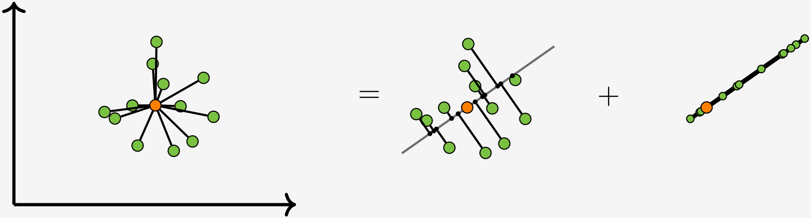
$$\left| \sum_{x \in Q} ||y - \pi_V(y)||^2 + \Delta - \sum_{x \in P} ||x - \pi_V(x)||^2 \right| \leq \varepsilon \sum_{x \in P} ||x - \pi_V(x)||^2$$

holds for all k-dimensional subspaces V.

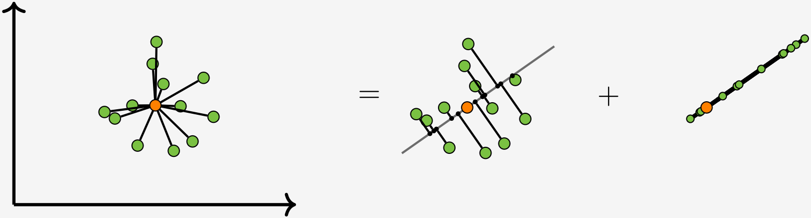
- Q is the projection of P to V_m with $m = \lceil k/\varepsilon \rceil$ $\quad A_m$
- Δ is the lost squared length $\sum_{i=m+1}^{r} \sigma_i^2$
- maximum error is $\sum_{i=m+1}^{m+k} \sigma_i^r \leq \varepsilon \sum_{i=k+1}^{r} \sigma_i^r$

How does this help for *k*-means?

Our idea: Split *k*-means cost into two terms

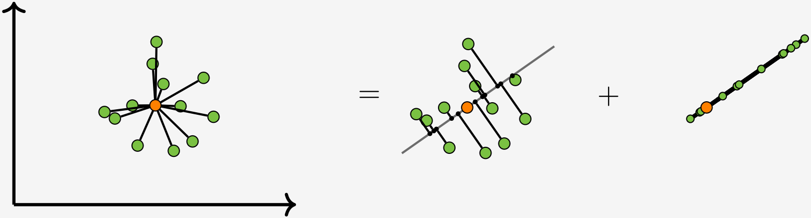


Our idea: Split k-means cost into two terms



For any k-dimensional subspace,
approximate squared distances to and within the subspace!

Our idea: Split k-means cost into two terms



For any k-dimensional subspace,
approximate squared distances to and within the subspace!

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Let $P \subseteq \mathbb{R}^d$, let C be k centers

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Let $P \subseteq \mathbb{R}^d$, let C be k centers

- Store the points as rows of a matrix $A \in \mathbb{R}^{n \times d}$

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Let $P \subseteq \mathbb{R}^d$, let C be k centers

- Store the points as rows of a matrix $A \in \mathbb{R}^{n \times d}$
- Define $(X_C)_{ij} = \begin{cases} 1/\sqrt{|C_j|} & \text{if } x_i \in C_j \\ 0 & \text{else} \end{cases} \rightsquigarrow (n \times k)$-matrix

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Let $P \subseteq \mathbb{R}^d$, let C be k centers

- Store the points as rows of a matrix $A \in \mathbb{R}^{n \times d}$
- Define $(X_C)_{ij} = \begin{cases} 1/\sqrt{|C_j|} & \text{if } x_i \in C_j \\ 0 & \text{else} \end{cases} \rightsquigarrow (n \times k)$-matrix
- $X_C X_C^T A = \begin{pmatrix} \dots \\ \mu(C(x_j)) \\ \dots \end{pmatrix}$

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Let $P \subseteq \mathbb{R}^d$, let C be k centers

- Store the points as rows of a matrix $A \in \mathbb{R}^{n \times d}$
- Define $(X_C)_{ij} = \begin{cases} 1/\sqrt{|C_j|} & \text{if } x_i \in C_j \\ 0 & \text{else} \end{cases} \rightsquigarrow (n \times k)$-matrix
- $X_C X_C^T A = \begin{pmatrix} \dots \\ \mu(C(x_j)) \\ \dots \end{pmatrix}$
- $\rightsquigarrow \sum_{j=1}^{n} ||x_j - \mu(C(x_j))||^2 = ||A - X_C X_C^T A||_F^2$

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Let $P \subseteq \mathbb{R}^d$, let C be k centers

- Store the points as rows of a matrix $A \in \mathbb{R}^{n \times d}$
- Define $(X_C)_{ij} = \begin{cases} 1/\sqrt{|C_j|} & \text{if } x_i \in C_j \\ 0 & \text{else} \end{cases} \rightsquigarrow (n \times k)$-matrix
- $X_C X_C^T A = \begin{pmatrix} \dots \\ \mu(C(x_j)) \\ \dots \end{pmatrix}$
- $\rightsquigarrow \sum_{j=1}^{n} ||x_j - \mu(C(x_j))||^2 = ||A - X_C X_C^T A||_F^2$
- $X_C X_C^T$ is a projection matrix and has rank k!

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Better Plan

Let $P \subseteq \mathbb{R}^d$, let C be k centers

- Store the points as rows of a matrix $A \in \mathbb{R}^{n \times d}$
- Define $(X_C)_{ij} = \begin{cases} 1/\sqrt{|C_j|} & \text{if } x_i \in C_j \\ 0 & \text{else} \end{cases} \rightsquigarrow (n \times k)$-matrix
- $X_C X_C^T A = \begin{pmatrix} \cdots \\ \mu(C(x_j)) \\ \cdots \end{pmatrix}$
- $\rightsquigarrow \sum_{j=1}^{n} ||x_j - \mu(C(x_j))||^2 = ||A - X_C X_C^T A||_F^2$
- $X_C X_C^T$ is a projection matrix and has rank k!
- Theorem already works for X_C, result for k-means immediate

Cohen, Elder, Musco, Musco, Persu, 2015:

This is unnecessary, we are already done!

Boutsidis, Mahoney, Drineas, 2009

The k-means problem is equivalent to a
constraint subspace approximation problem

Boutsidis, Mahoney, Drineas, 2009

The k-means problem is equivalent to a
constraint subspace approximation problem in $\mathbb{R}^n$!

Boutsidis, Mahoney, Drineas, 2009

The k-means problem is equivalent to a
constraint subspace approximation problem in $\mathbb{R}^n$!

Fits the columns of A to a k-dimensional subspace.

Boutsidis, Mahoney, Drineas, 2009

The k-means problem is equivalent to a
constraint subspace approximation problem in $\mathbb{R}^n$!

Fits the columns of A to a k-dimensional subspace.

- Apply dimensionality reduction for subspace approximation
- Result is a $(n \times d)$-matrix of rank m

Boutsidis, Mahoney, Drineas, 2009

The k-means problem is equivalent to a constraint subspace approximation problem in $\mathbb{R}^n$!

Fits the columns of A to a k-dimensional subspace.

- Apply dimensionality reduction for subspace approximation
- Result is a $(n \times d)$-matrix of rank m

Dimensionality reduction for k-means to $\lceil k/\varepsilon \rceil$ dimensions!

Lower Bound, Cohen, Elder, Musco, Musco, Persu, 2015

For any $\varepsilon > 0$ there exist n, d, k and a point set $P \subseteq \mathbb{R}^d$ such that

- projecting to V_m with $m := \lceil k/\varepsilon \rceil - 1$
- and computing optimal centers on V_m
- does not give a $(1+\varepsilon)$-approximation

Lower Bound, Cohen, Elder, Musco, Musco, Persu, 2015

For any $\varepsilon > 0$ there exist n, d, k and a point set $P \subseteq \mathbb{R}^d$ such that

- projecting to V_m with $m := \lceil k/\varepsilon \rceil - 1$
- and computing optimal centers on V_m
- does not give a $(1+\varepsilon)$-approximation

Construction

- points with $\lceil k/\varepsilon \rceil + k - 1$ dimensions
- place simplex in $k - 1$ dimensions
- place a Gaussian cloud in remaining $\lceil k/\varepsilon \rceil$ dimensions

Optimal solution: One center for Gaussian cloud, $k - 1$ for simplex

Lower Bound, Cohen, Elder, Musco, Musco, Persu, 2015

For any $\varepsilon > 0$ there exist n, d, k and a point set $P \subseteq \mathbb{R}^d$ such that

- projecting to V_m with $m := \lceil k/\varepsilon \rceil - 1$
- and computing optimal centers on V_m
- does not give a $(1 + \varepsilon)$-approximation

Construction

- points with $\lceil k/\varepsilon \rceil + k - 1$ dimensions
- place simplex in $k - 1$ dimensions
- place a Gaussian cloud in remaining $\lceil k/\varepsilon \rceil$ dimensions

Optimal solution: One center for Gaussian cloud, $k - 1$ for simplex

Parameters are adjusted such that whp

- largest $\lceil k/\varepsilon \rceil$ singular vectors lie in the cloud
- $\rightsquigarrow$ simplex collapses to origin $\rightsquigarrow$ too high clustering cost

Lower Bound, Cohen, Elder, Musco, Musco, Persu, 2015

For any $\varepsilon > 0$ there exist n, d, k and a point set $P \subseteq \mathbb{R}^d$ such that

- projecting to V_m with $m := \lceil k/\varepsilon \rceil - 1$
- and computing optimal centers on V_m
- does not give a $(1 + \varepsilon)$-approximation

Construction

- points with $\lceil k/\varepsilon \rceil + k - 1$ dimensions
- place simplex in $k - 1$ dimensions
- place a Gaussian cloud in remaining $\lceil k/\varepsilon \rceil$ dimensions

Optimal solution: One center for Gaussian cloud, $k - 1$ for simplex

Thank you for your attention!