The k-means problem and algorithm

- The k-means algorithm, also known as Lloyd's algorithm, together with its variants is probably the most popular clustering algorithm

The k-means problem and algorithm

- The k-means algorithm, also known as Lloyd's algorithm, together with its variants is probably the most popular clustering algorithm
- k-means tries to find good solutions to k-median problems by a simple two step approach

The k-means problem and algorithm

- The k-means algorithm, also known as Lloyd's algorithm, together with its variants is probably the most popular clustering algorithm
- k-means tries to find good solutions to k-median problems by a simple two step approach
- it has various shortcomings that do not seem to affect its popularity

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

The k-means problem and algorithm

- The k-means algorithm, also known as Lloyd's algorithm, together with its variants is probably the most popular clustering algorithm
- k-means tries to find good solutions to k-median problems by a simple two step approach
- it has various shortcomings that do not seem to affect its popularity
- it can be very inefficient and find poor solutions

The k-means problem and algorithm

- The k-means algorithm, also known as Lloyd's algorithm, together with its variants is probably the most popular clustering algorithm
- k-means tries to find good solutions to k-median problems by a simple two step approach
- it has various shortcomings that do not seem to affect its popularity
- it can be very inefficient and find poor solutions
- will also see a local improvement algorithms with provable approximation guarantees

$D : M \times M \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ dissimilarity measure

k-median problem

Given $P \subset M, k \in \mathbb{N}$, find $C = \{c_1, \ldots, c_k\} \subset M$ that minimizes

$$\sum_{p \in P} \min_{1 \leq i \leq k} D(p, c_i).$$

$c_i's$ called centroids. For $D = D_{\ell_2^2}$ the k-median problem is called the *k-means problem*.

The k-median problem

$D : M \times M \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ dissimilarity measure

k-median problem

Given $P \subset M, k \in \mathbb{N}$, find $C = \{c_1, \ldots, c_k\} \subset M$ that minimizes

$$\sum_{p \in P} \min_{1 \leq i \leq k} D(p, c_i).$$

$c_i's$ called centroids. For $D = D_{\ell_2^2}$ the k-median problem is called the *k-means problem*.

- Given $C = \{c_1, \ldots, c_k\}$, define

$$C_i := \{p \in P | \forall j \; D(p, c_i) \leq D(p, c_j)\}.$$

- If ties are broken, $\mathcal{C} := \{C_1, \ldots, C_k\}$ is a partition of P.

The k-median problem - notation

- $C \subset M, |C| < \infty$,

$$D(x, C) := \min_{c \in C} D(x, c).$$

- $P, C \subset M, |P|, |C| < \infty$,

$$D(P, C) := \sum_{p \in P} D(p, C)$$

(D-) cost of P with respect to C

- $k \in \mathbb{N}$,

$$\mathrm{cost}_k^D(P) := \min_{C \subset M, |C| = k} D(P, C),$$

called *k-median cost of P*.

The k-median problem - again

k-median problem

Given $P \subset M, k \in \mathbb{N}$, find $C = \{c_1, \ldots, c_k\} \subset M$ such that

$$D(P, C) = \mathrm{cost}_k^D(P).$$

k-median problem - alternative view

- Given a subset $Q \subset M$

$$c^D(Q) := \text{argmin}_{x \in M} \sum_{p \in Q} D(p, x).$$

  is called the *centroid* of set Q (with respect to D)
- For a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of set P, the cost of the partition $\mathcal{C}$ is defined as the cost of the set of centroids $C = \{c^D(C_1), \ldots, c^D(C_k)\}$.

k-median problem - alternative view

- Given a subset $Q \subset M$

$$c^D(Q) := \text{argmin}_{x \in M} \sum_{p \in Q} D(p, x).$$

  is called the *centroid* of set Q (with respect to D)
- For a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of set P, the cost of the partition $\mathcal{C}$ is defined as the cost of the set of centroids $C = \{c^D(C_1), \ldots, c^D(C_k)\}$.

k-median problem - alternative definition

Given a set of points $P \subset M$ and $k \in \mathbb{N}$, find a partition of P into k subsets or clusters $C_1, \ldots, C_k$ with corresponding set of centroids $C = \{c^D(C_1), \ldots, c^D(C_k)\}$ such that $D(P, C) = \text{cost}_k^D(P)$.

The k-means algorithm - idea

Idea of k-means

1. choose k initial centers
2. repeat the following steps until there is no improvement in cost function
   a) $C_i :=$ set of points closest to c_i
   b) $c_i :=$ centroid of C_i

The k-means algorithm - idea

Idea of k-means

1. choose k initial centers
2. repeat the following steps until there is no improvement in cost function
   a) $C_i :=$ set of points closest to c_i
   b) $c_i :=$ centroid of C_i

Questions

1. What are the centroids (with respect to D)?
2. Does k-means converge? If so, how fast?
3. How good are the solutions found by k-means?
4. For which dissimilarity measures can it be applied?

Centroids for D_{l_2}

- called Weber points
- can not be represented exactly using simple functions $(+, \times, \sqrt[d]{\cdot}, d \in \mathbb{N})$ in original points

Centroids for squared Euclidean distance

Lemma 2.1

For any finite set $X \subset \mathbb{R}^d$ the centroid of X with respect to the squared euclidean distance $D_{l_2^2}$ is given by the center of gravity of the points in X, i.e.

$$c(X) = \frac{1}{|X|} \sum_{x \in X} x.$$

More precisely, for any $y \in \mathbb{R}^d$:

$$D_{l_2^2}(X, y) = D_{l_2^2}(X, c(X)) + |X| \cdot D_{l_2^2}(c(X), y).$$

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

The k-means algorithm

K-MEANS(P)

---

choose k initial centroids $c_1, \ldots, c_k$;

repeat

  for $i = 1, \ldots, k$ **do**

    $C_i :=$ set of points in P closest to c_i;

  for $i = 1, \ldots, k$ **do**

    $c_i := c(C_i) = \frac{1}{|C_i|} \sum_{p \in C_i} p$;

until *convergence*;

return $c_1, \ldots, c_k$ and $C_1, \ldots, C_k$

The k-means algorithm

K-MEANS(P)

---

choose k initial centroids $c_1, \dots, c_k$;

repeat

  for $i = 1, \dots, k$ **do**

    $C_i :=$ set of points in P closest to c_i;

  for $i = 1, \dots, k$ **do**

    $c_i := c(C_i) = \frac{1}{|C_i|} \sum_{p \in C_i} p$;

until *convergence*;

return $c_1, \dots, c_k$ and $C_1, \dots, C_k$

---

convergence: quality of solution no longer improves k-MEANS(P)

choose k initial centroids $c_1, \dots, c_k$;

repeat

```
/* assignment step */
```

for $i = 1, \dots, k$ **do**

$C_i :=$ set of points in P closest to c_i;

for $i = 1, \dots, k$ **do**

$c_i := c(C_i) = \frac{1}{|C_i|} \sum_{p \in C_i} p$;

until *convergence*;

return $c_1, \dots, c_k$ and $C_1, \dots, C_k$

convergence: quality of solution no longer improves

The k-means algorithm

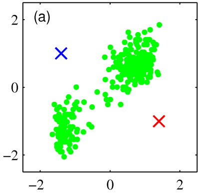
k-MEANS(P)

```
choose k initial centroids c_1, ..., c_k;
repeat
    /* assignment step                                    */
    for i = 1, ..., k do
        C_i := set of points in P closest to c_i;
    /* update step                                        */
    for i = 1, ..., k do
        c_i := c(C_i) = 1/|C_i| Σ_{p∈C_i} p;
until convergence;
return c_1, ..., c_k and C_1, ..., C_k
```

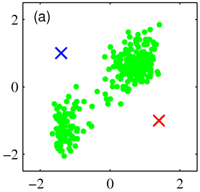
convergence: quality of solution no longer improves

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

The k-means algorithm - an example and useful lemma



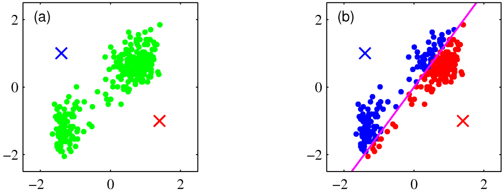
The k-means algorithm - an example and useful lemma



Lemma 2.2

Let $p, q \in \mathbb{R}^d$. The set of points x satisfying $D_{\ell_2^2}(p, x) = D_{\ell_2^2}(q, x)$ is given by the hyperplane orthogonal to $q - p$ and containing the midpoint $(p + q)/2$ of the line segment between p and q.

The k-means algorithm - an example and useful lemma

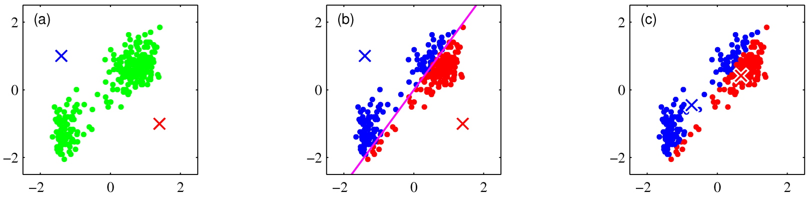


Lemma 2.2

Let $p, q \in \mathbb{R}^d$. The set of points x satisfying $D_{\ell_2^2}(p, x) = D_{\ell_2^2}(q, x)$ is given by the hyperplane orthogonal to $q - p$ and containing the midpoint $(p + q)/2$ of the line segment between p and q.

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

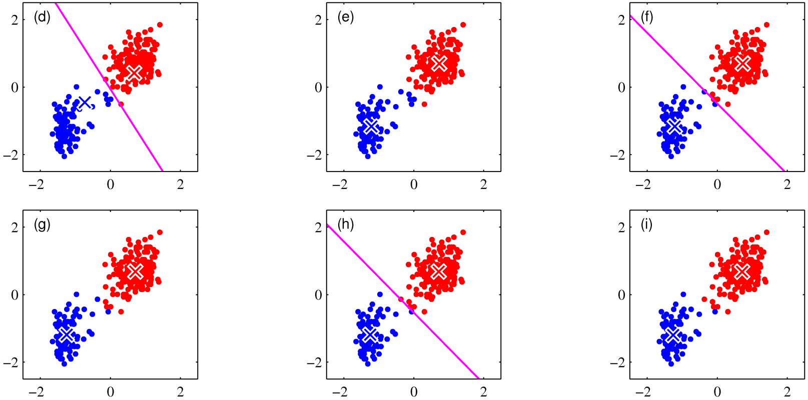
The k-means algorithm - an example and useful lemma



Lemma 2.2

Let $p, q \in \mathbb{R}^d$. The set of points x satisfying $D_{\ell_2^2}(p,x) = D_{\ell_2^2}(q,x)$ is given by the hyperplane orthogonal to $q - p$ and containing the midpoint $(p+q)/2$ of the line segment between p and q.

The k-means algorithm - an example and usefu lemma



Lemma 2.3

Algorithm k-MEANS *always halts after a finite number of steps. The number of assignment and update steps can be bounded by* $n^{O(k^2 \cdot d)}$.

Simple properties

Lemma 2.3

Algorithm K-MEANS *always halts after a finite number of steps. The number of assignment and update steps can be bounded by* $n^{O(k^2 \cdot d)}$.

Lemma 2.4

For every n *there exists a set* $P \subset \mathbb{R}^2$ *with* $n = |P|$, *a number* $k = \Theta(n)$ *and initial centroids such that on input* P K-MEANS *uses* $2^{\Omega(n)}$ *assignment and update steps. If* $k = o(n)$, *then the lower bound on the number of assignment and update steps becomes* $2^{\Omega(k)}$

Quality of solutions

Algorithm k-Means can get stuck in arbitrarily poor local minima.

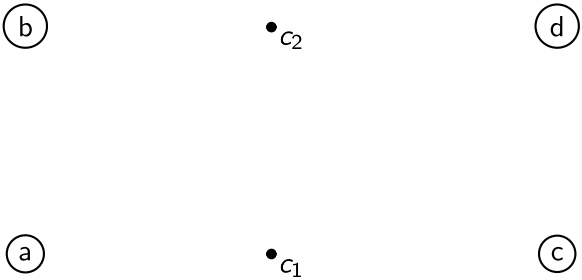
Simple properties

Quality of solutions

Algorithm k-MEANS can get stuck in arbitrarily poor local minima.

Quality of solutions

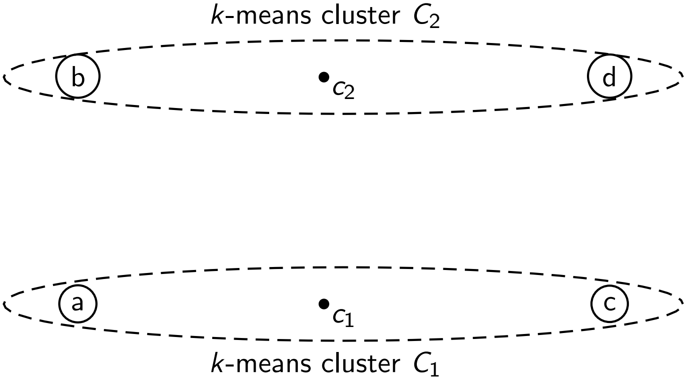
Algorithm k-Means can get stuck in arbitrarily poor local minima.



Simple properties

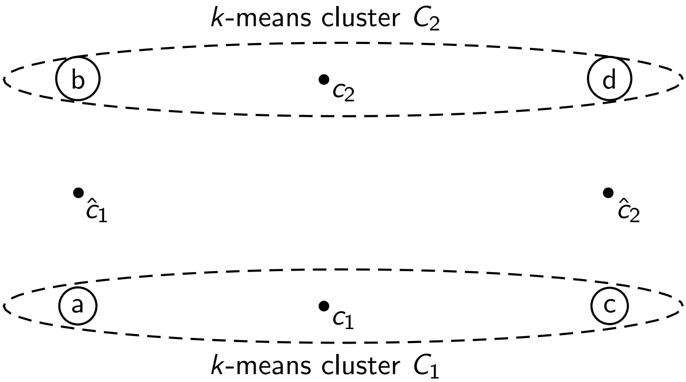
Quality of solutions

Algorithm k-Means can get stuck in arbitrarily poor local minima.



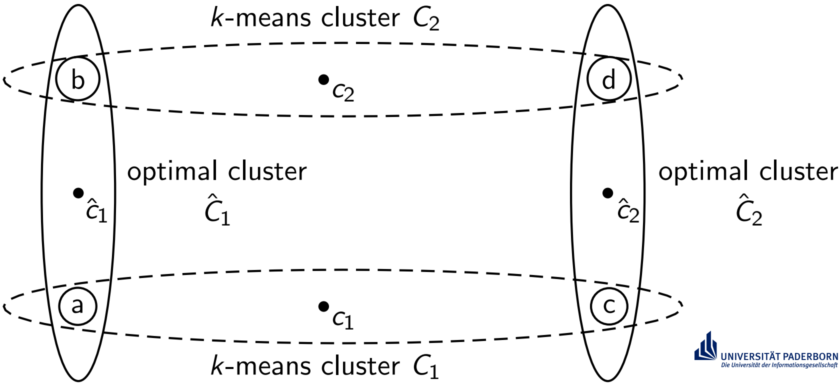
Quality of solutions

Algorithm k-Means can get stuck in arbitrarily poor local minima.



Quality of solutions

Algorithm k-MEANS can get stuck in arbitrarily poor local minima.



Complexity of k-means

Theorem 2.5

The k-means problem is **NP***-complete. This remains true if*

1. $d = 2$ *and* k *is arbitrary,*
2. $k = 2$ *and* d *is arbitrary.*

Complexity of k-means

Theorem 2.5

*The k-means problem is **NP**-complete. This remains true if*

1. $d = 2$ *and* k *is arbitrary,*
2. $k = 2$ *and* d *is arbitrary.*

Theorem 2.6

If $\mathbf{P} \neq \mathbf{NP}$*, then there is a constant* $\epsilon > 0$ *such that there is no polynomial time algorithms that for any finite point set* $P \subset \mathbb{R}^d$ *and any* $k \in \mathbb{N}$ *computes a set* $C, |C| = k$*, satisfying* $D(P, C) \leq (1 + \epsilon) \cdot cost_k^D(P)$*. Here* $D = D_{\ell_2^2}$ *is the squared euclidean distance.*

Bregman divergences

Definition 2.7

$S \subseteq \mathbb{R}^d, S \neq \emptyset$, convex, $\phi : S \to \mathbb{R}$ differentiable, strictly convex function. The Bregman divergence d_ϕ associated to ϕ is defined by

$$\begin{array}{rcl} S \times S & \to & \mathbb{R}_{\geq 0} \\ (x, y) & \mapsto & \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle. \end{array}$$

Definition 2.7

$S \subseteq \mathbb{R}^d, S \neq \emptyset$, convex, $\phi : S \to \mathbb{R}$ differentiable, strictly convex function. The Bregman divergence d_ϕ associated to ϕ is defined by

$$\begin{array}{rcl} S \times S & \to & \mathbb{R}_{\geq 0} \\ (x, y) & \mapsto & \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle. \end{array}$$

Remarks

- S convex: $\forall x, y \in S, \lambda \in [0, 1] : \lambda \cdot x + (1 - \lambda) \cdot y \in S$

Definition 2.7

$S \subseteq \mathbb{R}^d, S \neq \emptyset$, convex, $\phi : S \to \mathbb{R}$ differentiable, strictly convex function. The Bregman divergence d_ϕ associated to ϕ is defined by

$$\begin{array}{rcl} S \times S & \to & \mathbb{R}_{\geq 0} \\ (x, y) & \mapsto & \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle. \end{array}$$

Remarks

- S convex: $\forall x, y \in S, \lambda \in [0, 1] : \lambda \cdot x + (1 - \lambda) \cdot y \in S$
- ϕ strictly convex: $\forall x, y \in S, \lambda \in (0, 1)$:
  $\lambda \cdot \phi(x) + (1 - \lambda) \cdot \phi(y) > \phi(\lambda \cdot x + (1 - \lambda) \cdot y)$

Bregman divergences

Definition 2.7

$S \subseteq \mathbb{R}^d, S \neq \emptyset$, convex, $\phi : S \to \mathbb{R}$ differentiable, strictly convex function. The Bregman divergence d_ϕ associated to ϕ is defined by

$$\begin{array}{rcl} S \times S & \to & \mathbb{R}_{\geq 0} \\ (x, y) & \mapsto & \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle. \end{array}$$

Remarks

- S convex: $\forall x, y \in S, \lambda \in [0, 1] : \lambda \cdot x + (1 - \lambda) \cdot y \in S$
- ϕ strictly convex: $\forall x, y \in S, \lambda \in (0, 1)$:
  $\lambda \cdot \phi(x) + (1 - \lambda) \cdot \phi(y) > \phi(\lambda \cdot x + (1 - \lambda) \cdot y)$
- $\langle \cdot, \cdot \rangle$ denotes inner product

Bregman divergences

Definition 2.7

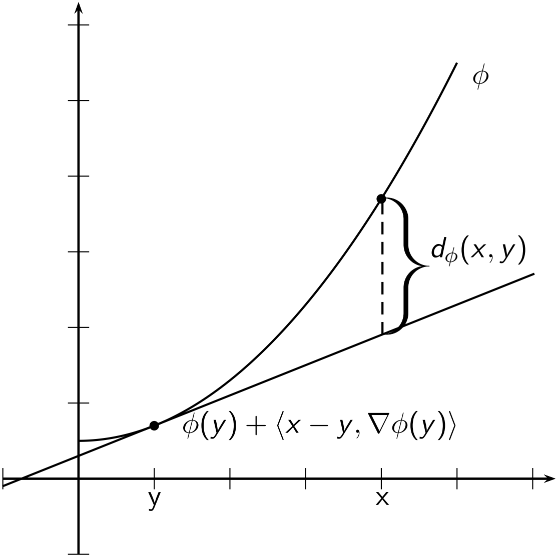
$S \subseteq \mathbb{R}^d, S \neq \emptyset$, convex, $\phi : S \to \mathbb{R}$ differentiable, strictly convex function. The Bregman divergence d_ϕ associated to ϕ is defined by

$$\begin{array}{ccc} S \times S & \to & \mathbb{R}_{\geq 0} \\ (x, y) & \mapsto & \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle. \end{array}$$

Remarks

- S convex: $\forall x, y \in S, \lambda \in [0, 1] : \lambda \cdot x + (1 - \lambda) \cdot y \in S$
- ϕ strictly convex: $\forall x, y \in S, \lambda \in (0, 1)$:
  $\lambda \cdot \phi(x) + (1 - \lambda) \cdot \phi(y) > \phi(\lambda \cdot x + (1 - \lambda) \cdot y)$
- $\langle \cdot, \cdot \rangle$ denotes inner product
- $g : \mathbb{R}^d \to \mathbb{R}, (x_1, \dots, x_d) \mapsto g(x_1, \dots, x_d)$, then
  $\nabla g(y) := \left(\frac{\partial g}{x_1}(y), \dots, \frac{\partial g}{x_d}(y)\right)$

Geometric interpretation



Observation

Suppose $\phi : (a, b) \to \mathbb{R}$ has a continuous second derivative $\phi''(\cdot)$ such that $\phi''(x) > 0$ for all $x \in (a, b)$, then ϕ is strictly convex.

Observation

Suppose $\phi : (a, b) \to \mathbb{R}$ has a continuous second derivative $\phi''(\cdot)$ such that $\phi''(x) > 0$ for all $x \in (a, b)$, then ϕ is strictly convex.

Examples

- $D_{l_2^2}$ is the Bregman divergence associated to

$$\phi(x) = \langle x, x \rangle = \sum_{i=1}^{d} x_i^2 .$$

Observation

Suppose $\phi : (a, b) \to \mathbb{R}$ has a continuous second derivative $\phi''(\cdot)$ such that $\phi''(x) > 0$ for all $x \in (a, b)$, then ϕ is strictly convex.

Examples

- $D_{l_2^2}$ is the Bregman divergence associated to

$$\phi(x) = \langle x, x \rangle = \sum_{i=1}^{d} x_i^2.$$

- $D_A, A \in \mathbb{R}^{d \times d}$ positive definite, is the Bregman divergence associated to $\phi(x) = x^T \cdot A \cdot x$.

Observation

Suppose $\phi : (a, b) \to \mathbb{R}$ *has a continuous second derivative* $\phi''(\cdot)$ *such that* $\phi''(x) > 0$ *for all* $x \in (a, b)$, *then* ϕ *is strictly convex.*

Examples

- D_{KL} is the Bregman divergence associated to

$$\phi(x) = \sum_{i=1}^{d} x_i \log x_i .$$

Properties

Lemma 2.8

Bregman divergences are positive and reflexive.

Properties

Lemma 2.8

Bregman divergences are positive and reflexive.

Lemma 2.9

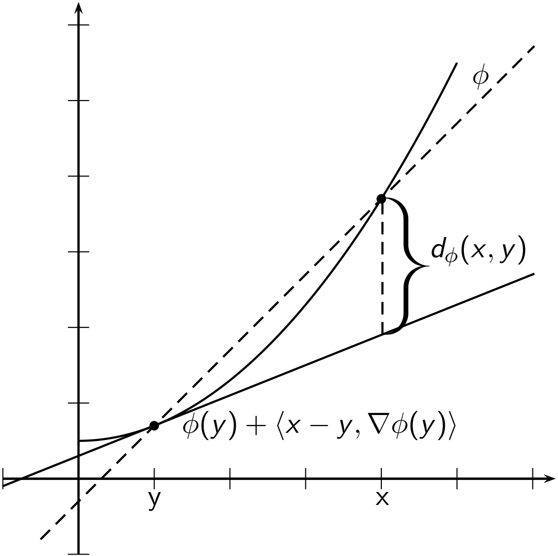
Let $d_\phi : S \times S \to \mathbb{R}_{\geq 0}$ *be a Bregman divergence and* $X \subset S, |X| < \infty$. *Then*

$$c(X) := \frac{1}{|X|} \sum_{x \in X} x = argmin_{y \in S} d_\phi(X, y).$$

More precisely, for any $y \in S$*:*

$$d_\phi(X, y) = d_\phi(X, c(X)) + |X| \cdot d_\phi(c(X), y).$$

Proof sketch for Lemma 2.8



The k-means algorithm for Bregman divergences

Bregman k-Means(P)

```
choose k initial centroids c_1, ..., c_k;
repeat
    /* assignment step */
    for i = 1, ..., k do
        C_i := set of points in P closest to c_i with respect to d_φ;
    /* update step */
    for i = 1, ..., k do
        c_i := c(C_i) = 1/|C_i| Σ_{p∈C_i} p;
until convergence;
return c_1, ..., c_k and C_1, ..., C_k
```

d_ϕ a Bregman divergence.