離散最適化基礎論　第 4 回
クラスタリング (1)：k-センター

岡本 吉央
okamotoy@uec.ac.jp

電気通信大学

2017 年 11 月 10 日

最終更新：2017 年 11 月 10 日　11:38

主題

離散最適化のトピックの１つとして幾何的被覆問題を取り上げ,
その数理的側面と計算的側面の双方を意識して講義する

なぜ講義で取り扱う？

- 「離散最適化」と「計算幾何学」の接点として重要な役割を果たしているから
- 様々なアルゴリズム設計技法・解析技法を紹介できるから
- 応用が多いから

スケジュール 前半 (予定)

1. 幾何的被覆問題とは？ (10/6)
- ⋆ 国内出張のため休み (10/13)
2. 最小包囲円問題 (1)：基本的な性質 (10/20)
3. 最小包囲円問題 (2)：乱択アルゴリズム (10/27)
- ⋆ 文化の日のため休み (11/3)
4. クラスタリング (1)：k-センター (11/10)
5. 幾何ハイパーグラフ (1)：VC 次元 (11/17)
- ⋆ 調布祭 のため 休み (11/24)
6. 幾何ハイパーグラフ (2)：ε ネット (12/1)

注意：予定の変更もありうる

スケジュール 後半 (予定)

7 幾何的被覆問題 (1)：線形計画法の利用 (12/8)
8 幾何的被覆問題 (2)：シフト法 (12/15)
9 幾何的被覆問題 (3)：局所探索法 (12/22)
10 幾何的被覆問題 (4)：局所探索法の解析 (1/5)
⋆ センター試験準備 のため 休み (1/12)
11 幾何ハイパーグラフ (3)：ε ネット定理の証明 (1/19)
12 幾何アレンジメント (1)：合併複雑度と ε ネット (1/26)
13 幾何アレンジメント (2)：合併複雑度の例 (2/2)
14 最近のトピック (2/9)
15 期末試験 (2/16？)

注意：予定の変更もありうる

今日の内容

クラスタリング

- クラスタリング：様々な最適化モデル
  - k-センター，k-メディアン，k-ミーンズ
- k-センター：近似アルゴリズム
- k-センター：近似アルゴリズムの限界

連続型単位円被覆問題 (continuous unit disk cover problem)

入力

- 平面上の点集合 $P = \{p_1, p_2, \ldots, p_n\}$

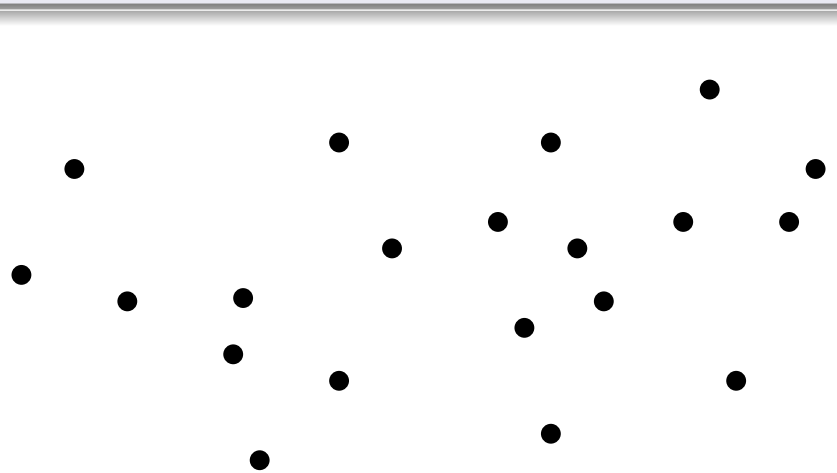
出力

- 単位円の集合 $\mathcal{D}'$ で次を満たすもの ($\mathcal{D}'$ が P を被覆する)

  任意の $p \in P$ に対して，ある $D \in \mathcal{D}'$ が存在して，$p \in D$

目的

- $|\mathcal{D}'|$ の最小化



連続型単位円被覆問題 (continuous unit disk cover problem)

入力

- 平面上の点集合 $P = \{p_1, p_2, \ldots, p_n\}$

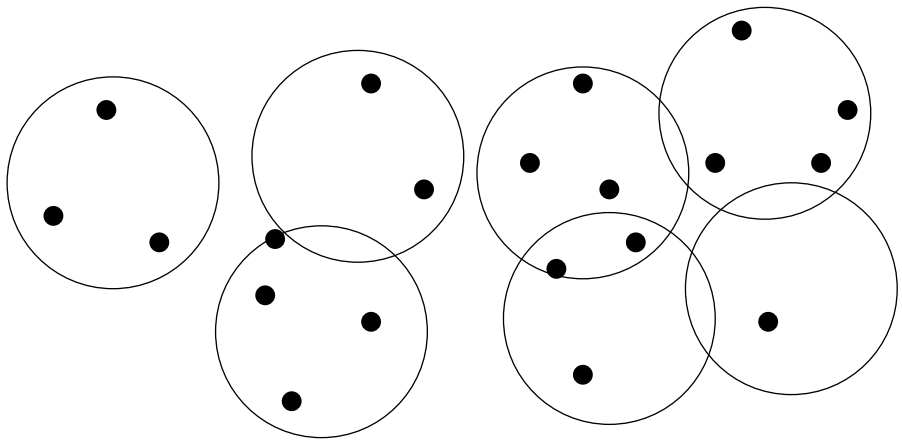
出力

- 単位円の集合 $\mathcal{D}'$ で次を満たすもの ($\mathcal{D}'$ が P を被覆する)

  任意の $p \in P$ に対して，ある $D \in \mathcal{D}'$ が存在して，$p \in D$

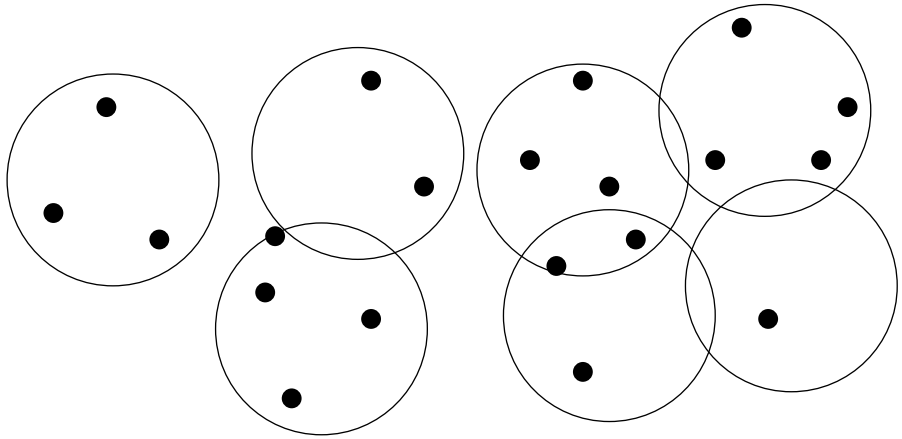
目的

- $|\mathcal{D}'|$ の最小化



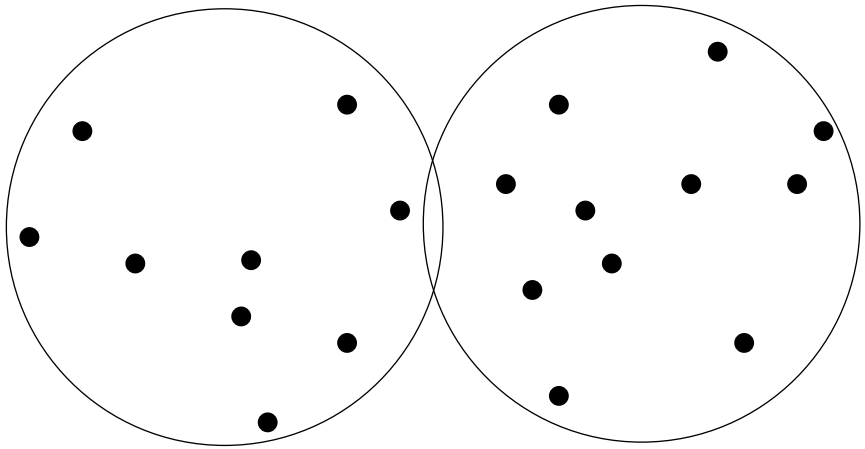
単位円ではなく，異なる半径の円を用いると？

用いる円の半径を大きくすると，
より少ない円で十分かもしれない (多くなることはない)



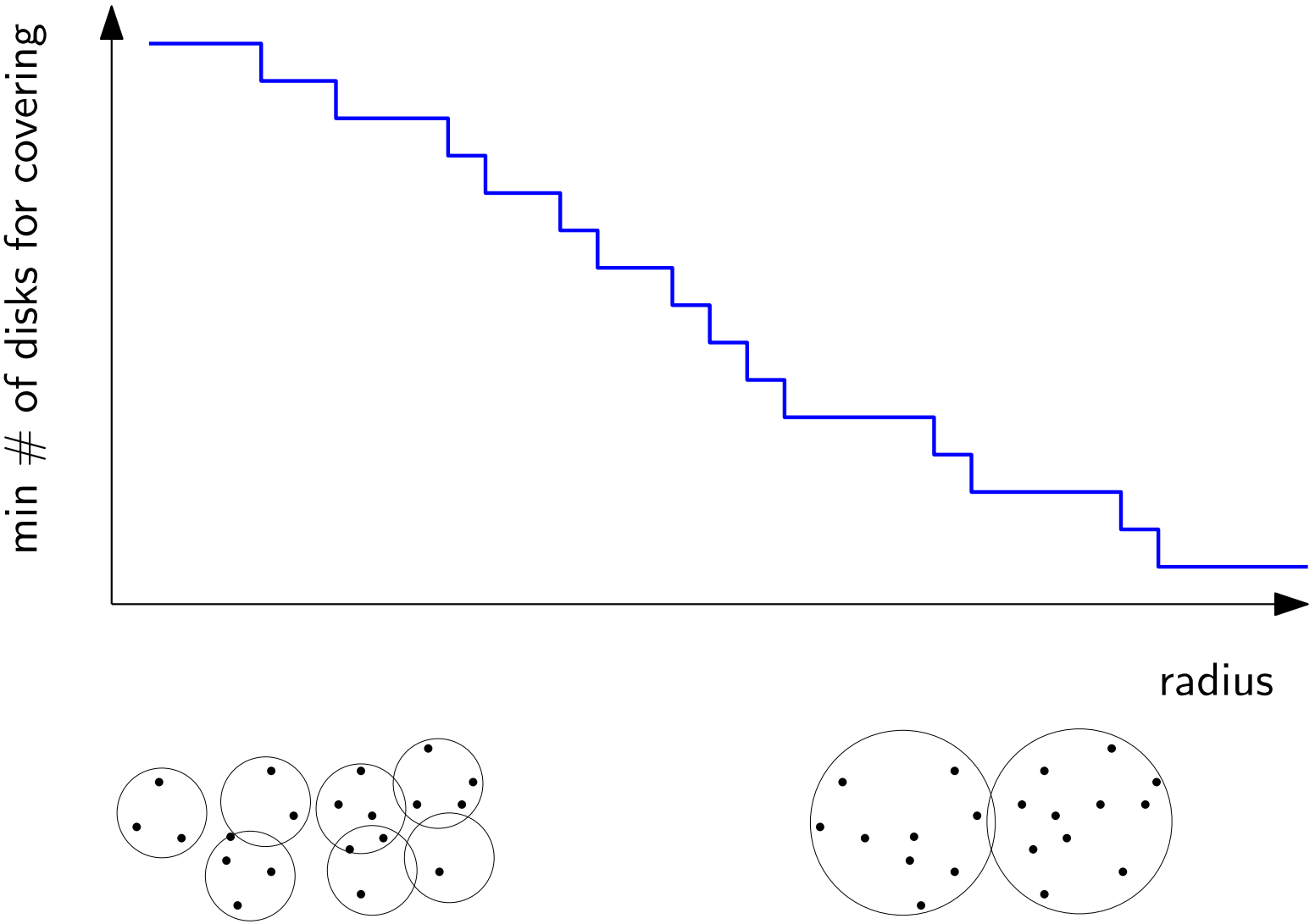
単位円ではなく，異なる半径の円を用いると？

用いる円の半径を大きくすると，
より少ない円で十分かもしれない (多くなることはない)



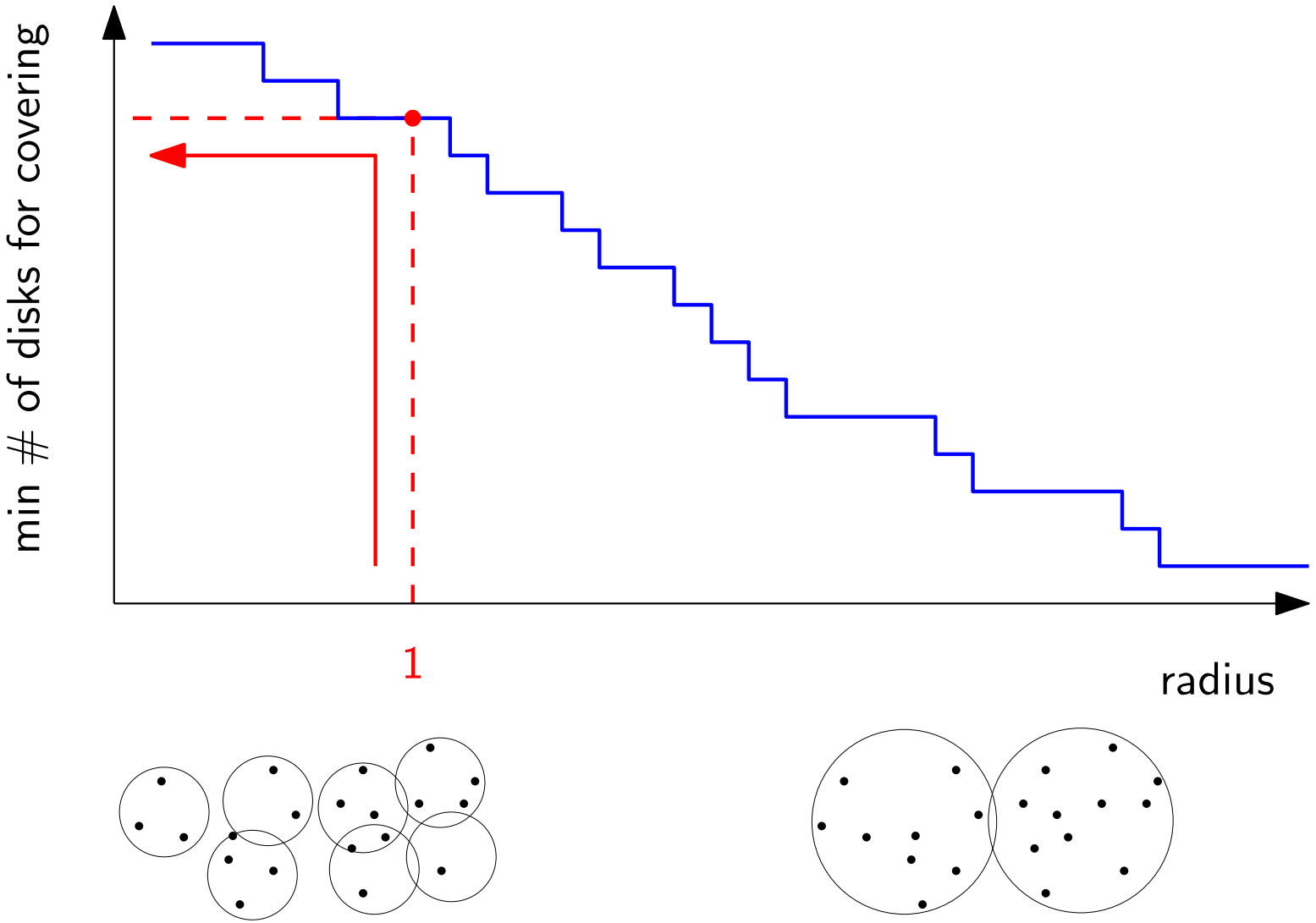
半径と最適な円の数の関係 (1)

半径を大きくすると，被覆に用いる円の最小数は単調非増加



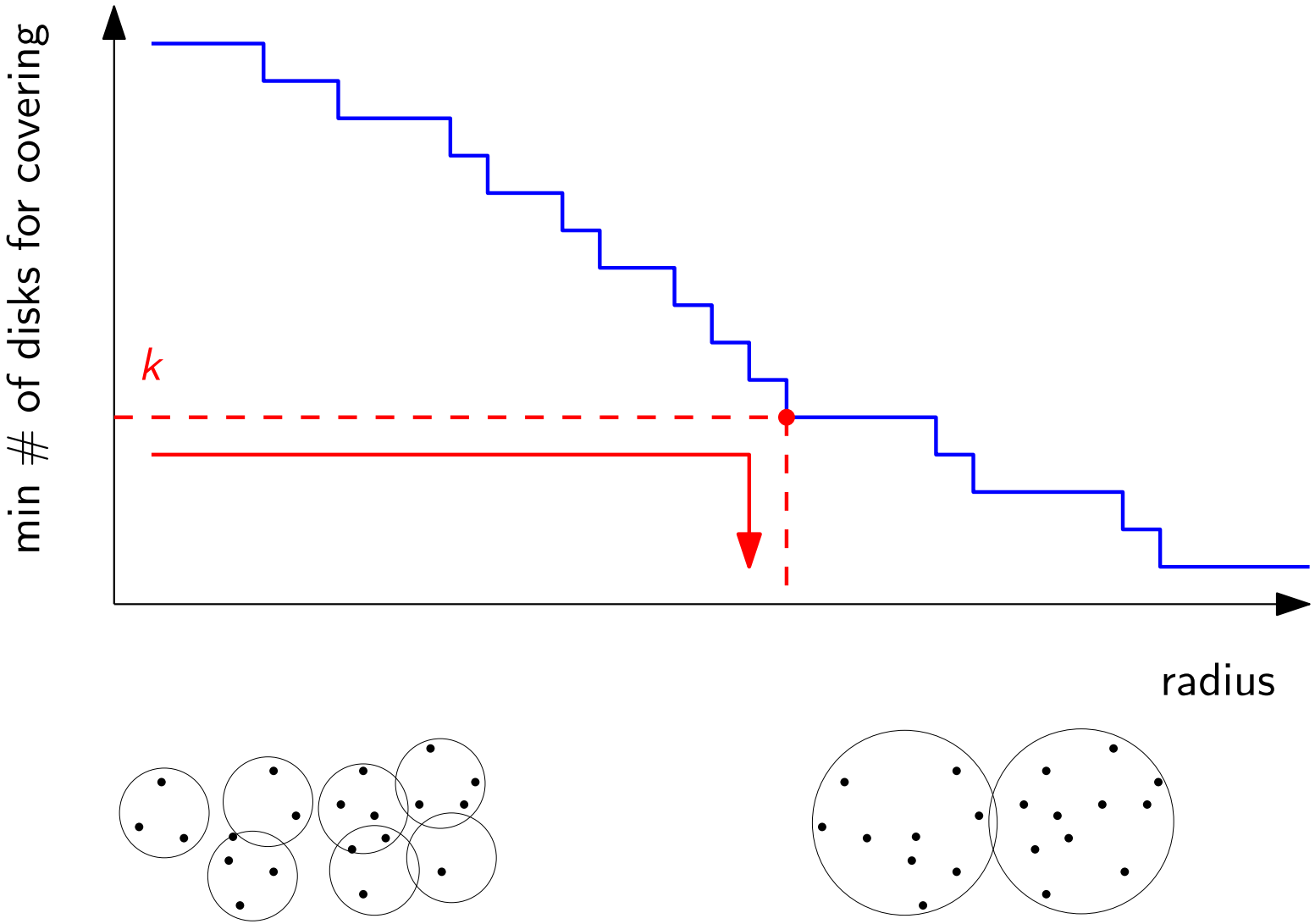
半径と最適な円の数の関係 (2)

単位円被覆問題：半径を 1 としたとき，円の最小数を問う



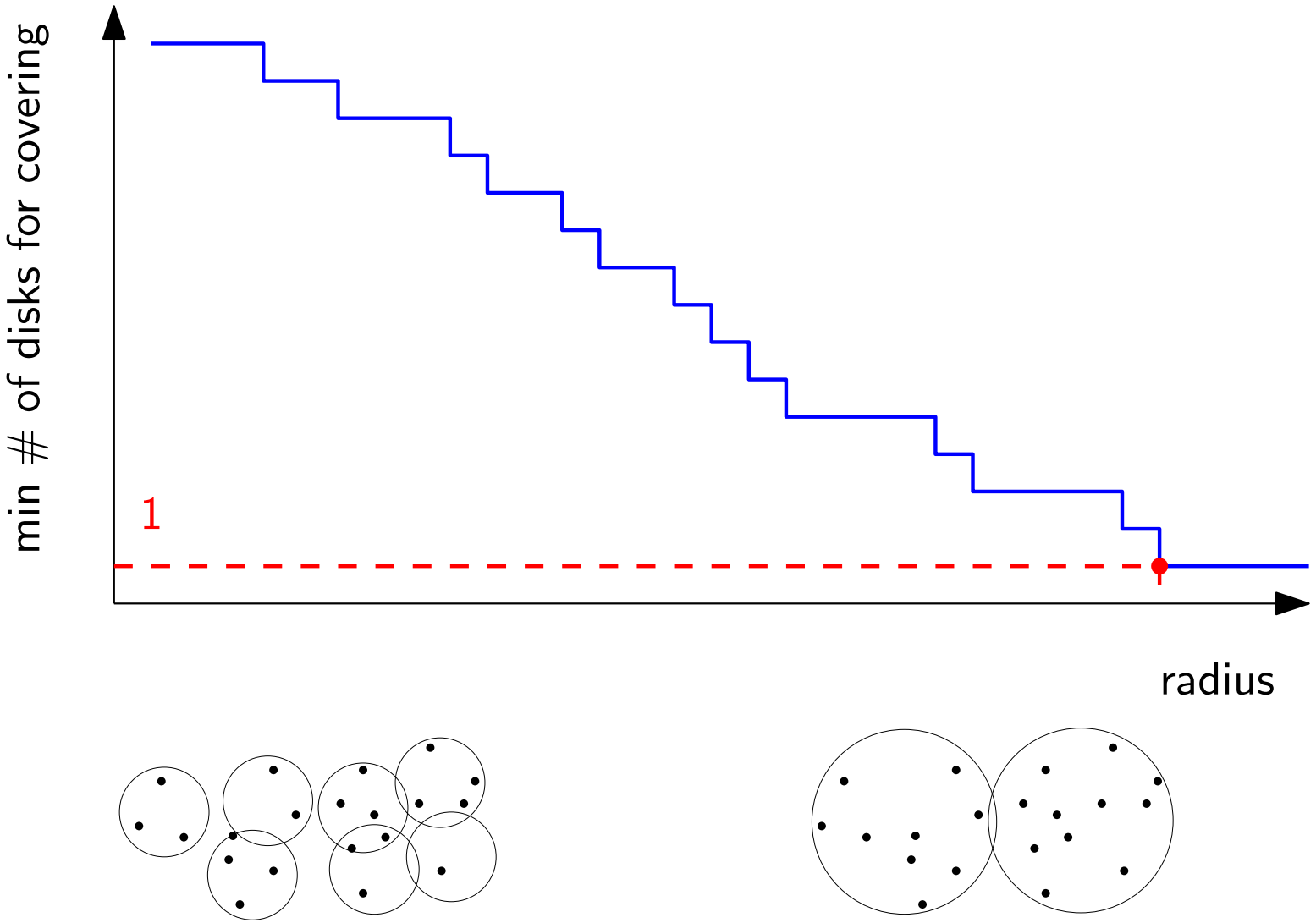
半径と最適な円の数の関係 (3)

k-センター問題：円の数を k としたとき，最小半径を問う



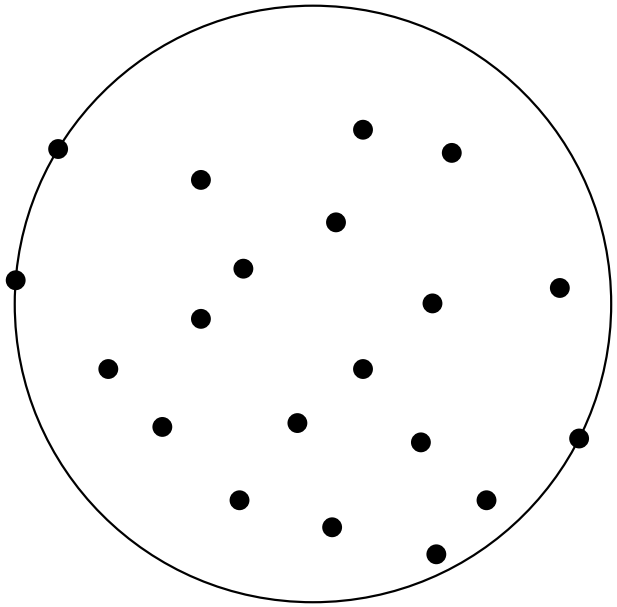
半径と最適な円の数の関係 (3)

1-センター問題 = 最小包囲円問題



最小包囲円問題

平面上にいくつか点が与えられたとき
それらをすべて含む円の中で面積が最小のものを求めよ



注意：円に対しては，面積の最小化 $\Leftrightarrow$ 半径の最小化

目次

クラスタリングの設定

- X：データがとられる集合 (母集団)
- $S \subseteq X$：とられたデータ (標本) の集合
- 各要素 x, y の非類似度 $d(x, y)$ が定められている

クラスタリングの目標

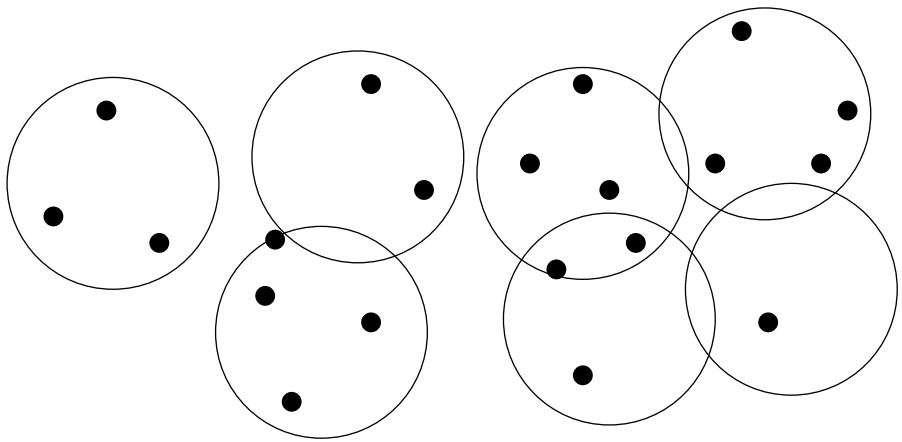
- ある規準に基づいて，点をいくつか選ぶ

$\rightsquigarrow$ 「規準」によって，異なる名称が用いられる

クラスタリング：先ほどの設定

- $X = \mathbb{R}^2$, $S \subseteq X$
- 非類似度 d：ユークリッド距離 (直線距離)
- 規準：k 個の点 $c_1, \ldots, c_k \in X$ を選んで，次の量の最小化

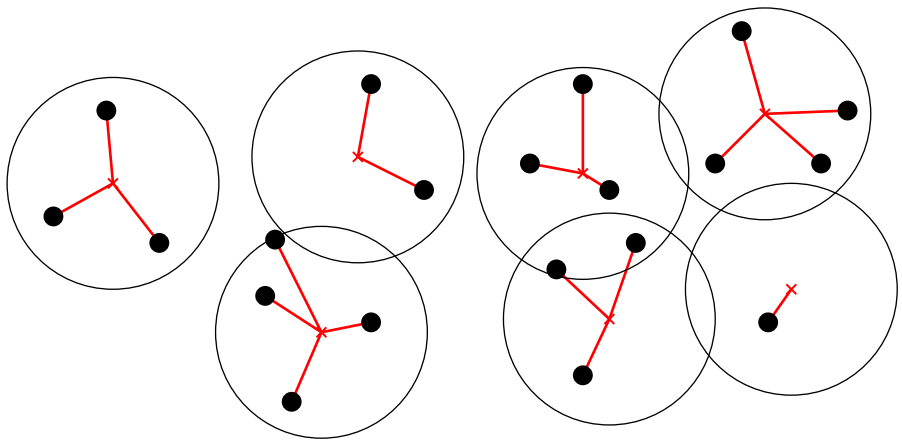
$$\max_{x \in S} \min_{i=1,\ldots,k} d(c_i, x)$$



クラスタリング：先ほどの設定

- $X = \mathbb{R}^2$, $S \subseteq X$
- 非類似度 d：ユークリッド距離 (直線距離)
- 規準：k 個の点 $c_1, \ldots, c_k \in X$ を選んで，次の量の最小化

$$\max_{x \in S} \min_{i=1,\ldots,k} d(c_i, x)$$



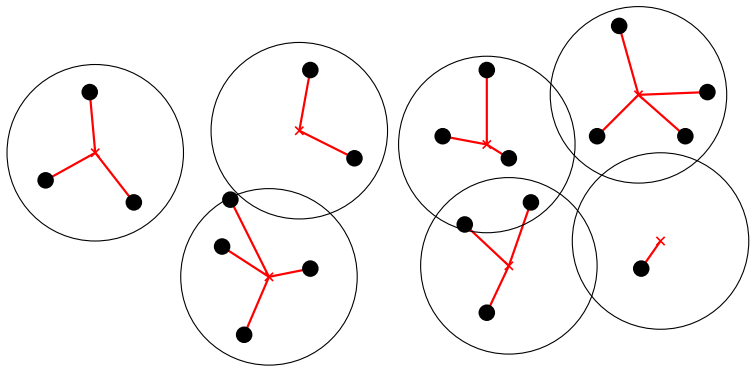
規準によって，様々な最適化モデルが得られる

規準：k 個の要素 $c_1, \ldots, c_k \in X$ を選んで，次の量の最小化

$$\max_{x \in S} \min_{i=1,\ldots,k} d(c_i, x) \qquad \text{連続型 } k\text{-センター問題}$$

$$\sum_{x \in S} \min_{i=1,\ldots,k} d(c_i, x) \qquad \text{連続型 } k\text{-メディアン問題}$$

$$\sum_{x \in S} \min_{i=1,\ldots,k} d(c_i, x)^2 \qquad \text{連続型 } k\text{-ミーンズ問題}$$



「k」ではなく「p」を使うことも多い

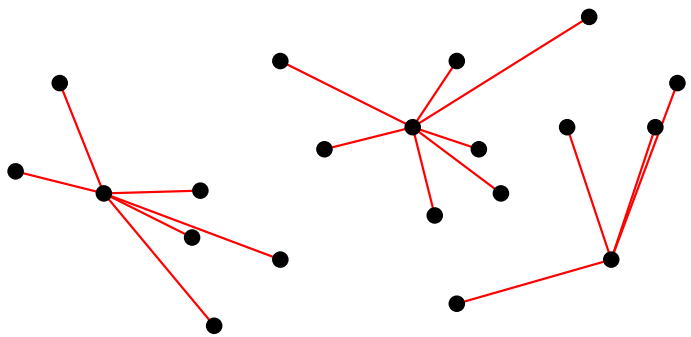
規準によって，様々な最適化モデルが得られる

規準：k 個の要素 $c_1, \ldots, c_k \in S$ を選んで，次の量の最小化

$$\max_{x \in S} \min_{i=1,\ldots,k} d(c_i, x) \qquad \text{離散型 } k\text{-センター問題}$$

$$\sum_{x \in S} \min_{i=1,\ldots,k} d(c_i, x) \qquad \text{離散型 } k\text{-メディアン問題}$$

$$\sum_{x \in S} \min_{i=1,\ldots,k} d(c_i, x)^2 \qquad \text{離散型 } k\text{-ミーンズ問題}$$



「k」ではなく「p」を使うことも多い

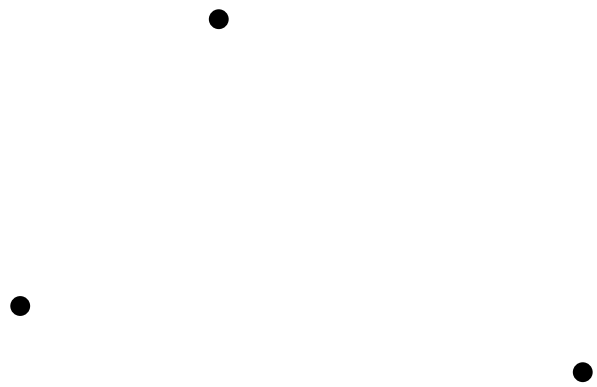
非類似度 $d\colon X^2 \to \mathbb{R}$ は次の性質を満たすものとする

1. 任意の $x, y \in X$ に対して，$d(x, y) \geq 0$
2. 任意の $x, y \in X$ に対して，$d(x, y) = 0 \Leftrightarrow x = y$
3. 任意の $x, y \in X$ に対して，$d(x, y) = d(y, x)$ (対称性)
4. 任意の $x, y, z \in X$ に対して，$d(x, y) \leq d(x, z) + d(z, y)$ (三角不等式)

これら 4 つの性質を持つ関数は距離 (metric) と呼ばれる

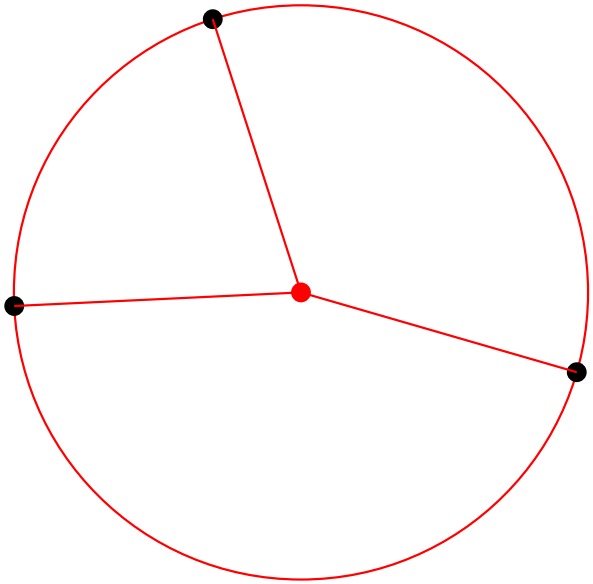
ユークリッド平面上の 3 点が鋭角三角形を成す場合

連続型 1-センター問題の最適解 $\rightsquigarrow$ 外接円 (最小包囲円)

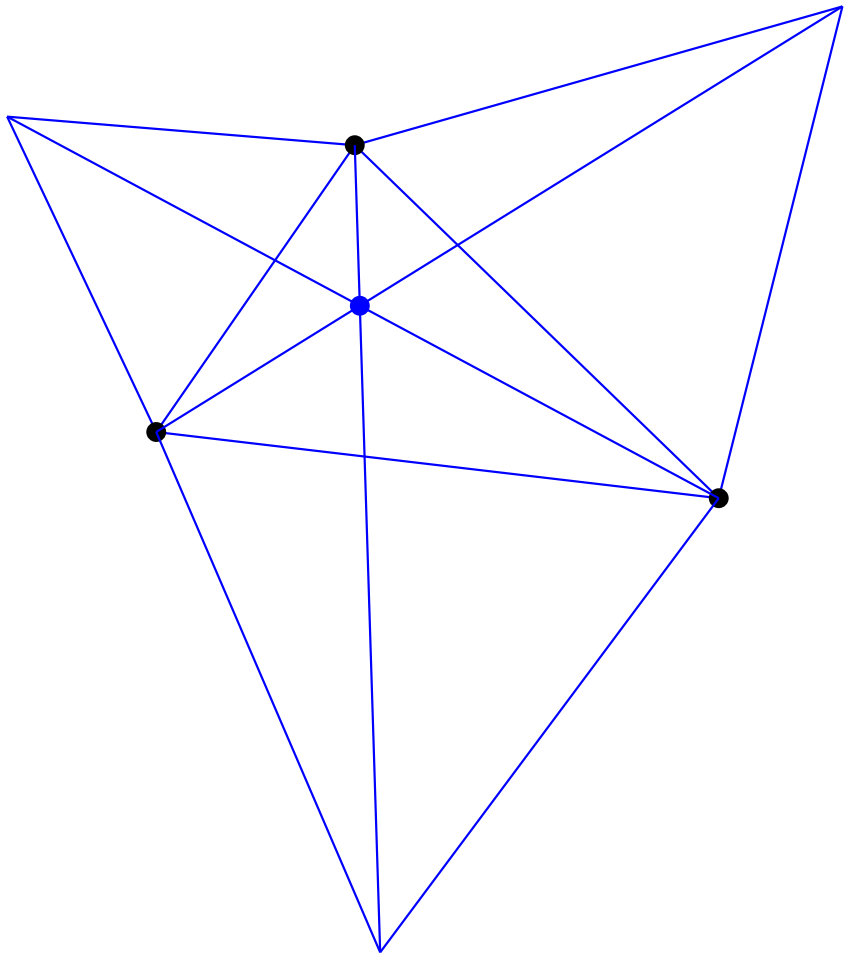


ユークリッド平面上の 3 点が鋭角三角形を成す場合

連続型 1-センター問題の最適解 $\rightsquigarrow$ 外接円 (最小包囲円)

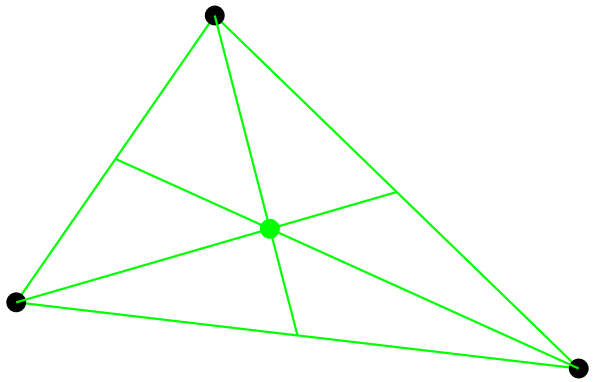


連続型 1-メディアン問題の最適解 $\rightsquigarrow$ フェルマー点 (トリチェリ点)

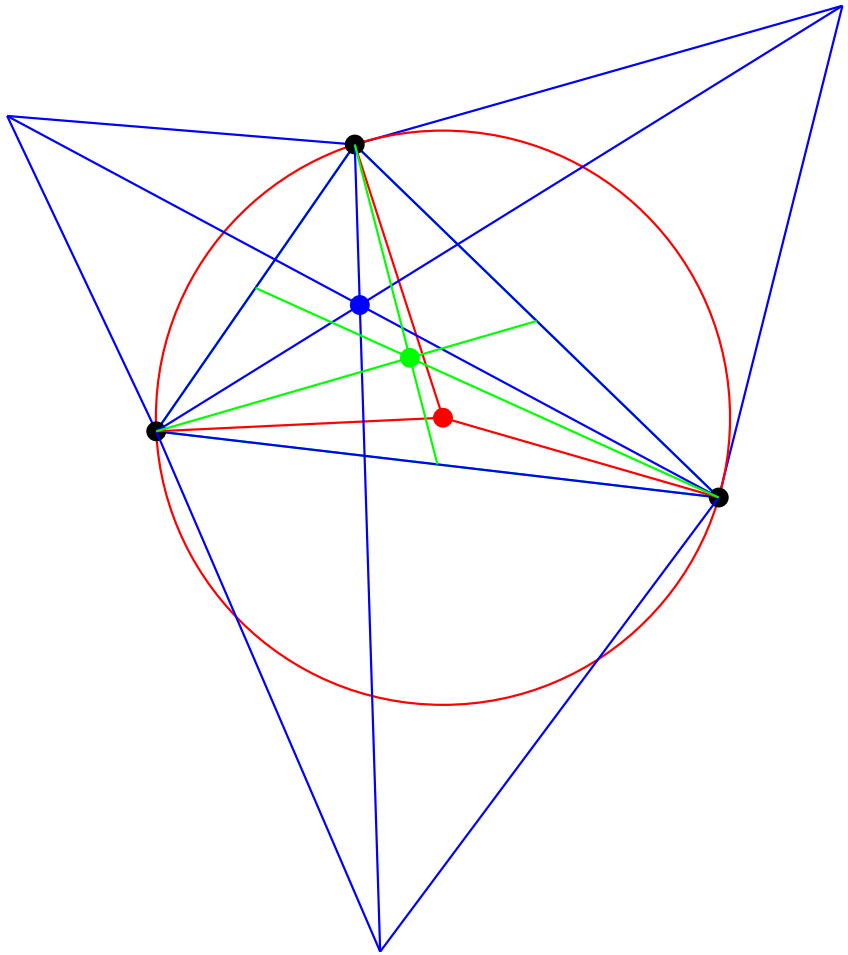


ユークリッド平面上の 3 点が鋭角三角形を成す場合

連続型 1-ミーンズ問題の最適解 $\rightsquigarrow$ 重心



比較



目次

k-センター問題に対する近似アルゴリズム

ここからの目標

離散型 k-センター問題に対する近似アルゴリズムを設計すること

ここで与えるアルゴリズムは
連続型 k-センター問題に対しても近似アルゴリズムとなる

知られていること

離散型 k-センター問題は NP 困難 (ユークリッド平面上の問題でも)

$\alpha \geq 1$ とする

定義：α 近似解

最小化問題に対する α 近似解とは，その問題に対する解 X で

$$\text{最適値} \quad \leq \quad X \text{ に対する目的関数値} \quad \leq \quad \alpha \cdot \text{最適値}$$

を満たすもののこと (この α のことを近似比と呼ぶことがある)

定義：α 近似アルゴリズム

最小化問題に対する α 近似アルゴリズムとは，
必ず α 近似解を出力するアルゴリズムのこと

アイディア

$$\alpha \text{ 近似解がよい近似} \quad \Longleftrightarrow \quad \alpha \text{ が小さい}$$

つまり，α が小さい近似アルゴリズムを設計することが目的

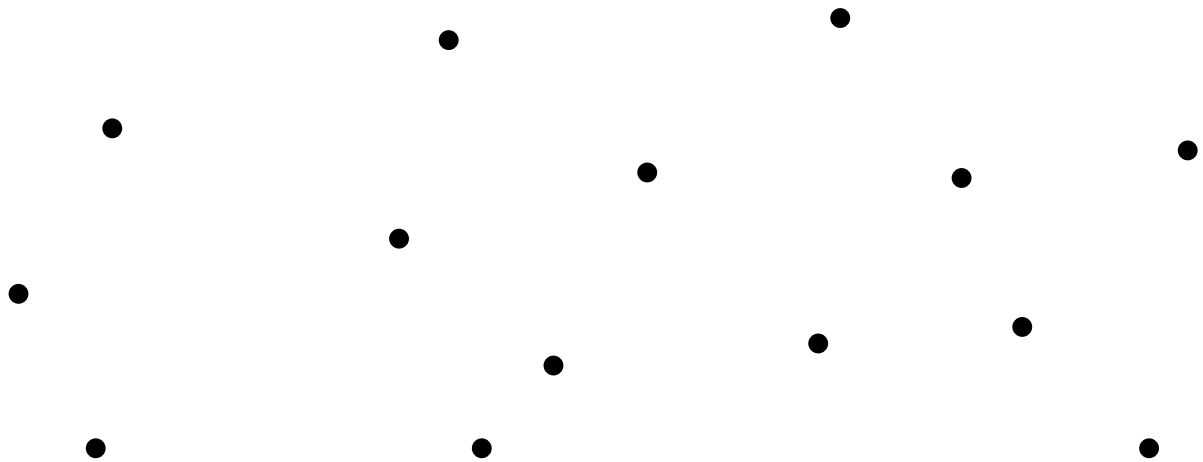
アルゴリズム (1)

離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- はじめは，任意の要素を選択する

$k = 4$ のときの例



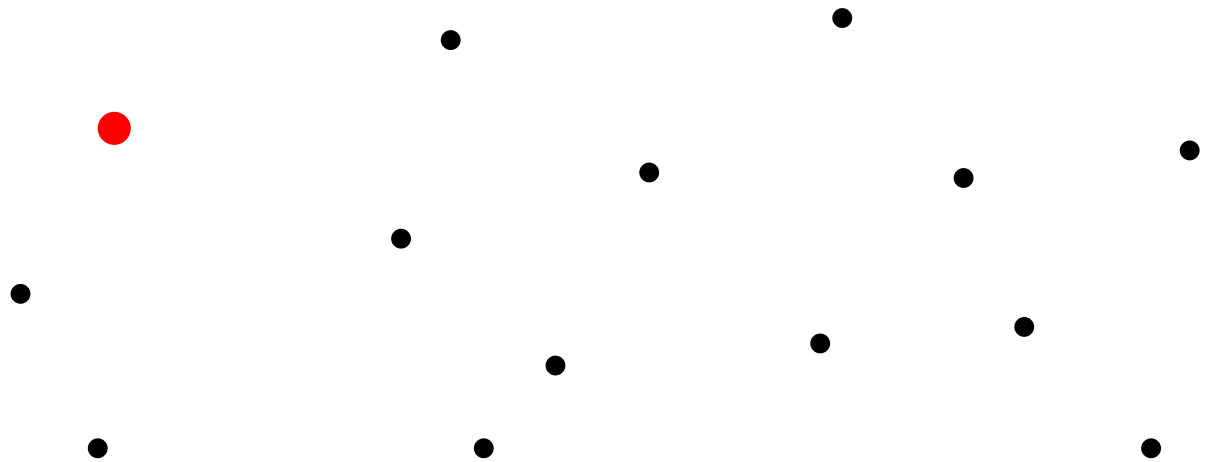
アルゴリズム (1)

離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- はじめは，任意の要素を選択する

$k = 4$ のときの例



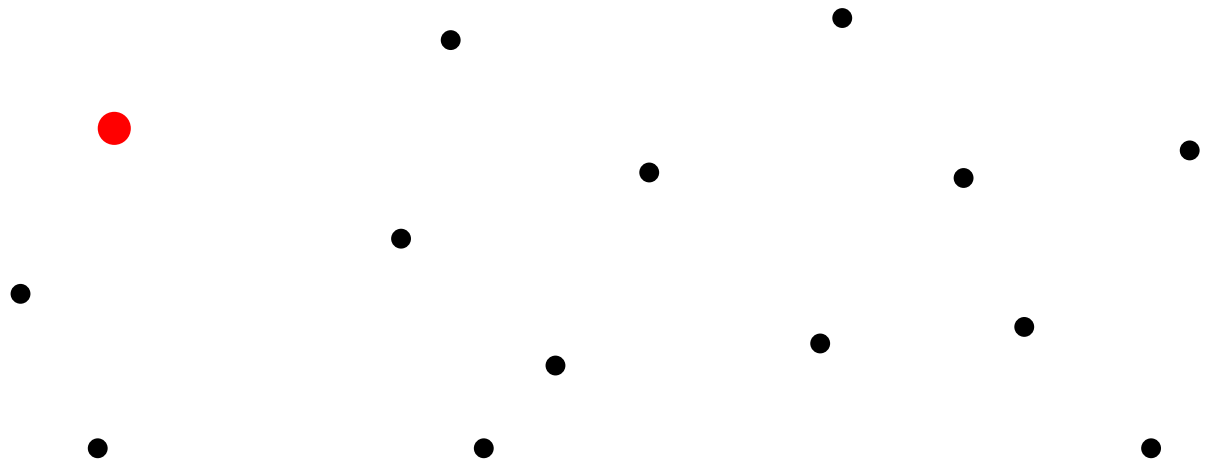
アルゴリズム (2)

離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- 2 個目からは，今まで選択した要素から最も遠い S の要素を選択する

$k = 4$ のときの例



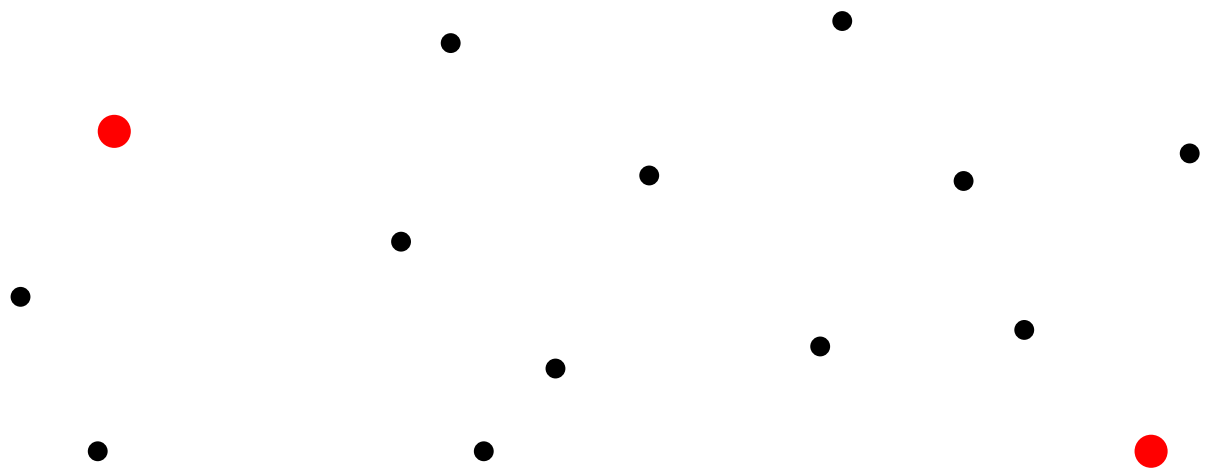
アルゴリズム (2)

離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- 2 個目からは，今まで選択した要素から最も遠い S の要素を選択する

$k = 4$ のときの例



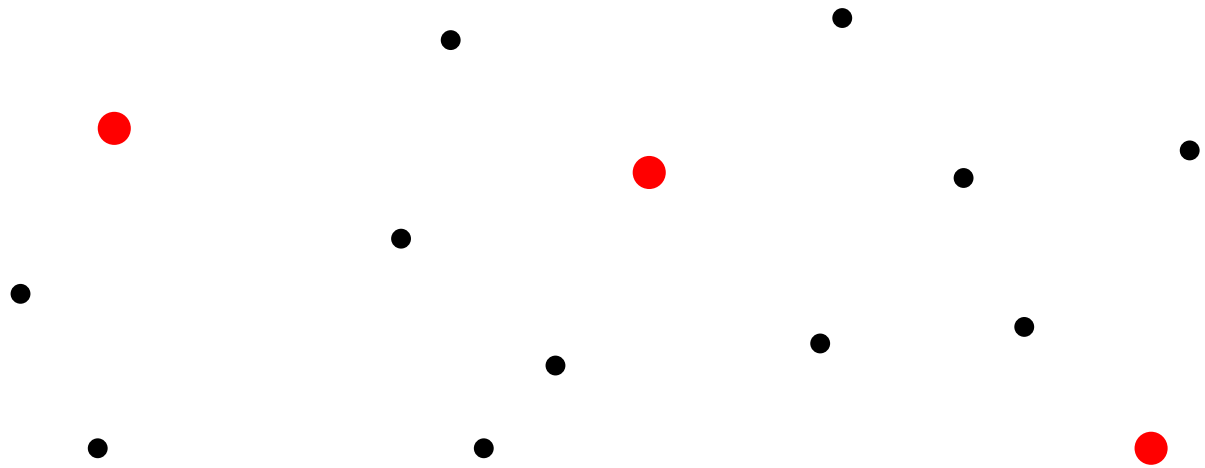
アルゴリズム (2)

離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- 2 個目からは，今まで選択した要素から最も遠い S の要素を選択する

$k = 4$ のときの例



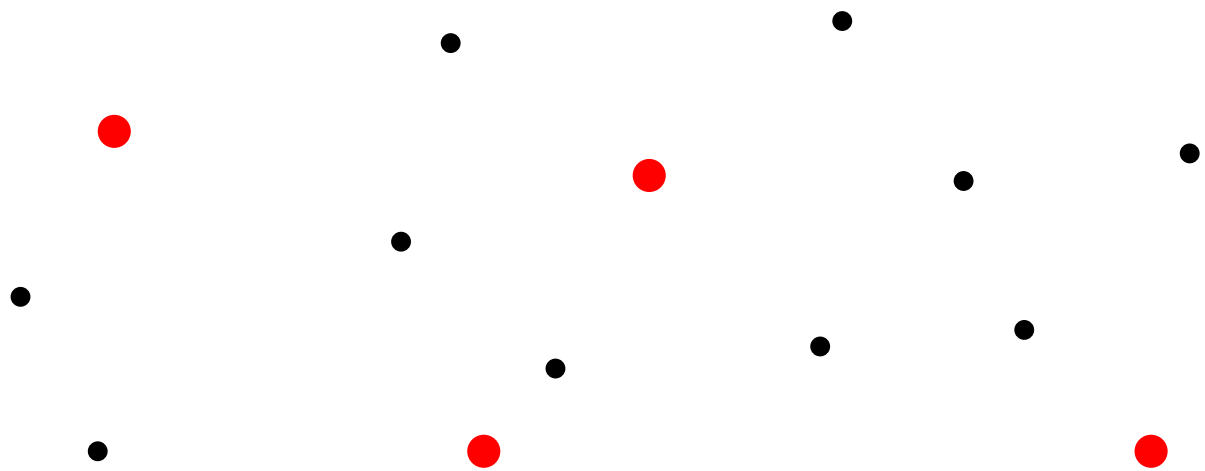
アルゴリズム (2)

離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- 2 個目からは，今まで選択した要素から最も遠い S の要素を選択する

$k = 4$ のときの例



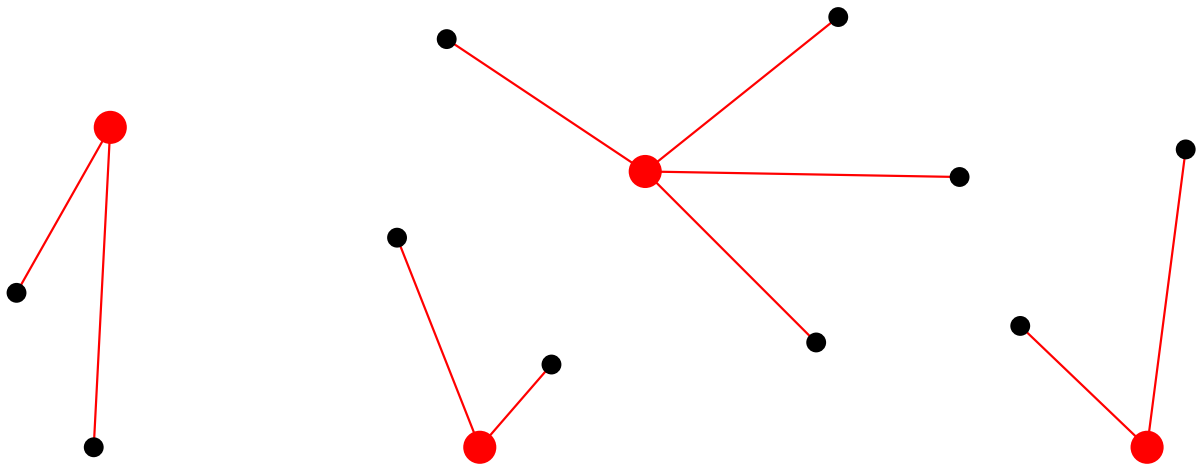
アルゴリズム (2)

離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- 2 個目からは，今まで選択した要素から最も遠い S の要素を選択する

$k = 4$ のときの例

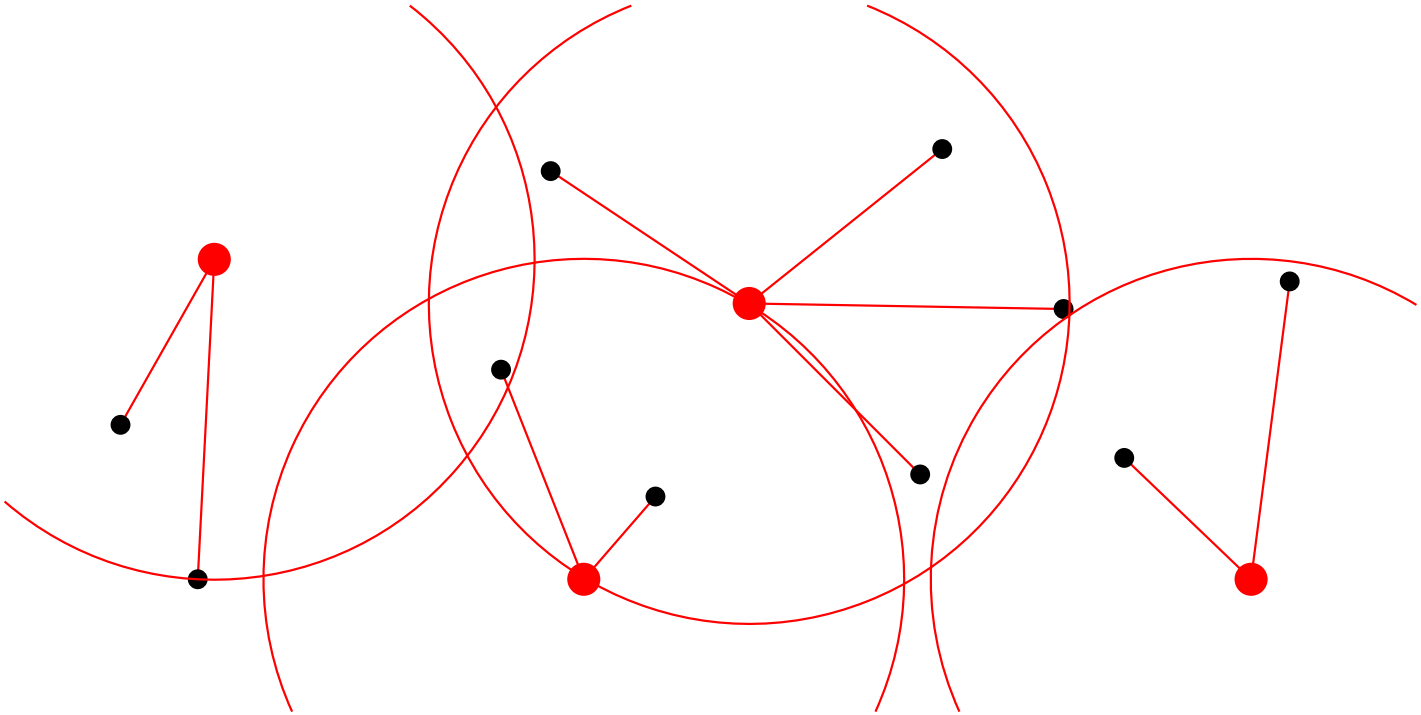


離散型 k-センター問題に対する近似アルゴリズム

k 個の要素を選択するまで，以下を繰り返す

- 2 個目からは，今まで選択した要素から最も遠い S の要素を選択する

$k = 4$ のときの例

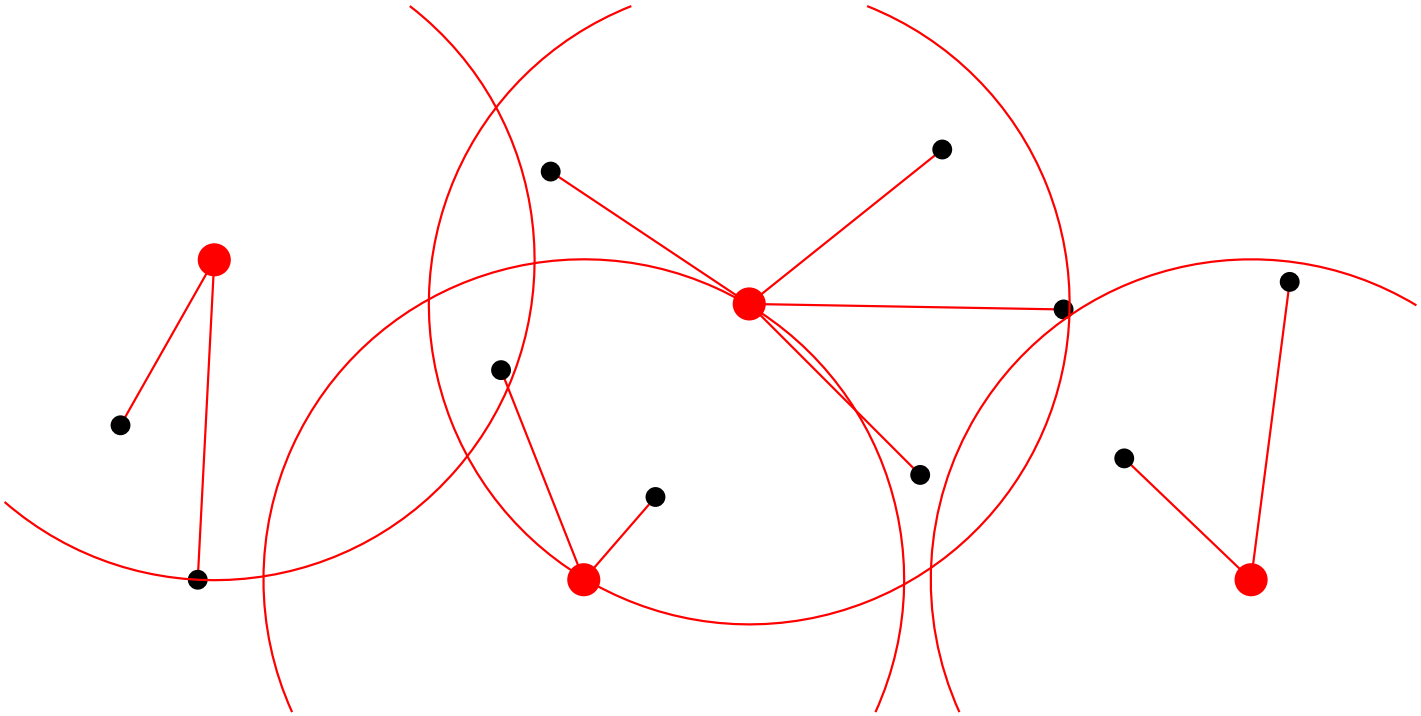


離散型 k-センター問題に対する近似アルゴリズム

つまり，$c_1, \ldots, c_{i-1}$ まで選択したとき，c_i は次のように選択する

- c_i は $\min_{j=1,\ldots,i-1} d(x, c_j)$ を最大化する $x \in S$

$k = 4$ のときの例



アルゴリズム：まとめ

入力 S, k

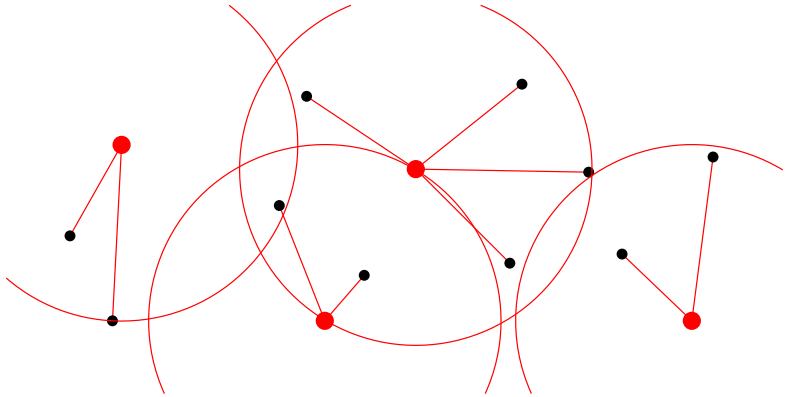
(1) S の任意の点を選び，c_1 とする

(2) $i = 2, \ldots, k$ に対して，以下を繰り返す

(2-1) $\min_{j=1,\ldots,i-1} d(x, c_j)$ を最大化する $x \in S$ を c_i とする

(3) $r = \max_{x \in S} \min_{i=1,\ldots,k} d(x, c_i)$ とする

(4) クラスタの中心を $c_1, \ldots, c_k$，クラスタ半径を r として終了



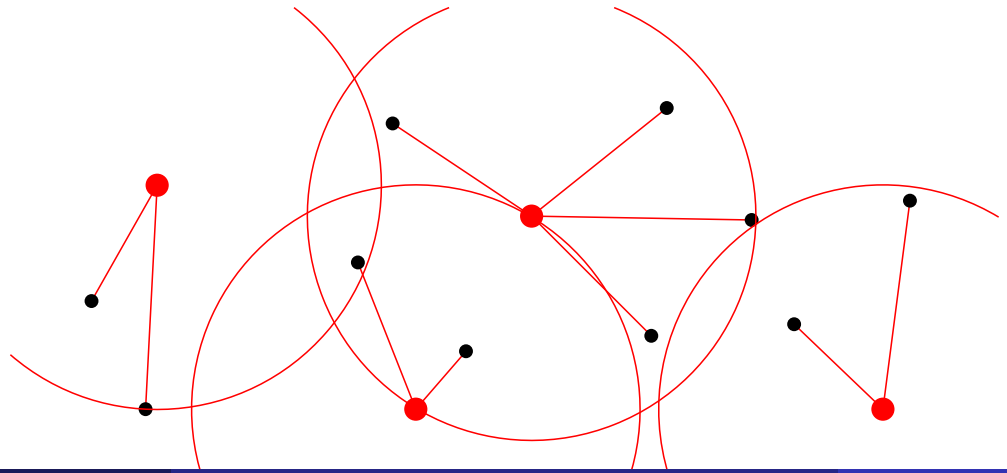
$O(k|S|)$ 時間で動作するように実装できる (それほど難しくない)

定理：アルゴリズムの近似比

このアルゴリズムは必ず 2 近似解を出力する

証明：

- アルゴリズムの出力する要素を $c_1, c_2, \ldots, c_k$ として，得られる半径を r とする
- 最適解において選択された要素を $c_1^*, c_2^*, \ldots, c_k^*$ として，最適解の半径を r^* とする
- 半径 r が要素 $x \in S$ と $c_i \in S$ によって達成されるとする すなわち，$d(x, c_i) = r$

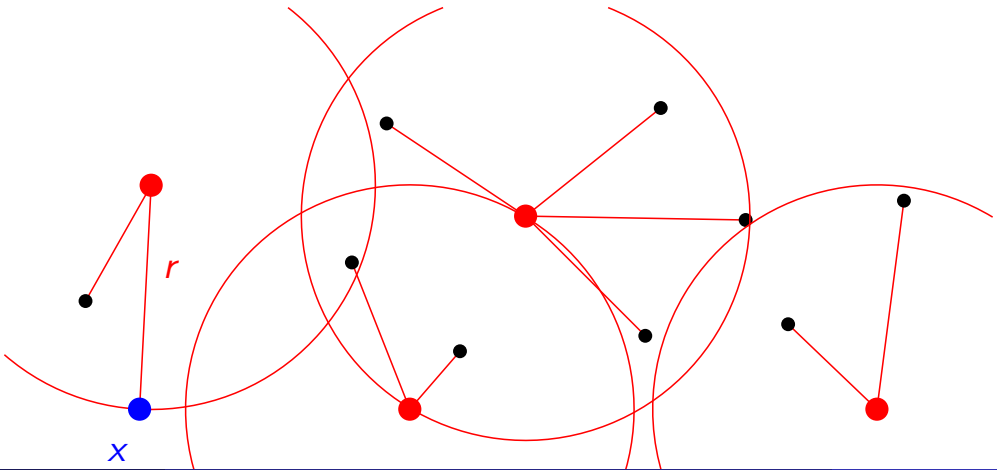


定理：アルゴリズムの近似比

このアルゴリズムは必ず 2 近似解を出力する

証明：

- アルゴリズムの出力する要素を $c_1, c_2, \ldots, c_k$ として，得られる半径を r とする
- 最適解において選択された要素を $c_1^*, c_2^*, \ldots, c_k^*$ として，最適解の半径を r^* とする
- 半径 r が要素 $x \in S$ と $c_i \in S$ によって達成されるとする すなわち，$d(x, c_i) = r$

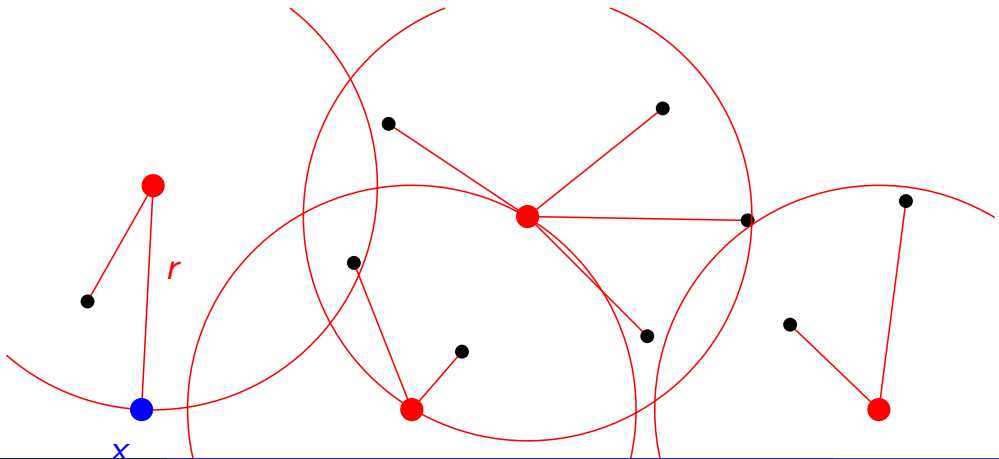


定理：アルゴリズムの近似比

このアルゴリズムは必ず 2 近似解を出力する

証明 (続)：

- このとき，$C = \{c_1, c_2, \ldots, c_k, x\}$ とすると，
  任意の $c, c' \in C$ に対して，$d(c, c') \geq r$
- 一方で，$|C| = k + 1$ であるから，ある j に対して，
  2 つの要素 $c, c' \in C$ が存在して，$d(c, c_j^*) \leq r^*, d(c', c_j^*) \leq r^*$
- 三角不等式を用いると，
  $r \leq d(c, c') \leq d(c, c_j^*) + d(c', c_j^*) \leq 2r^*$ □

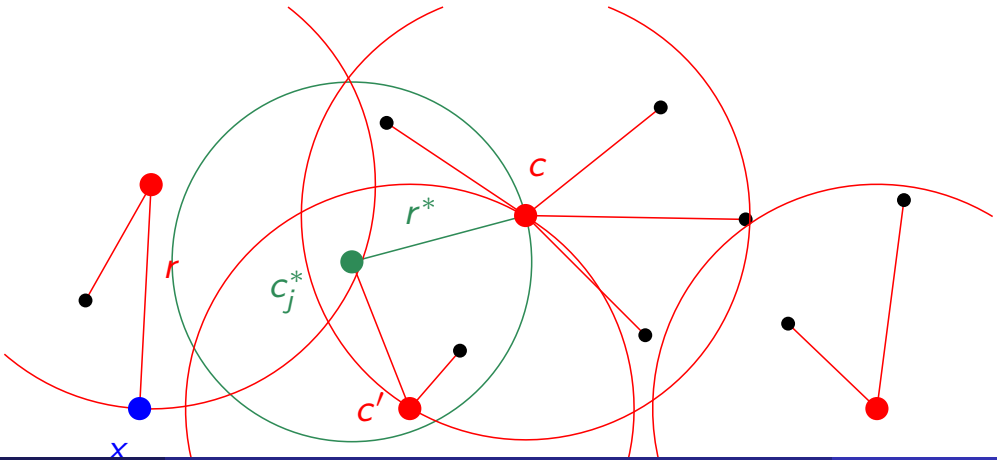


定理：アルゴリズムの近似比

このアルゴリズムは必ず 2 近似解を出力する

証明 (続)：

- このとき，$C = \{c_1, c_2, \ldots, c_k, x\}$ とすると，
  任意の $c, c' \in C$ に対して，$d(c, c') \geq r$
- 一方で，$|C| = k + 1$ であるから，ある j に対して，
  2 つの要素 $c, c' \in C$ が存在して，$d(c, c_j^*) \leq r^*, d(c', c_j^*) \leq r^*$
- 三角不等式を用いると，
  $r \leq d(c, c') \leq d(c, c_j^*) + d(c', c_j^*) \leq 2r^*$ □

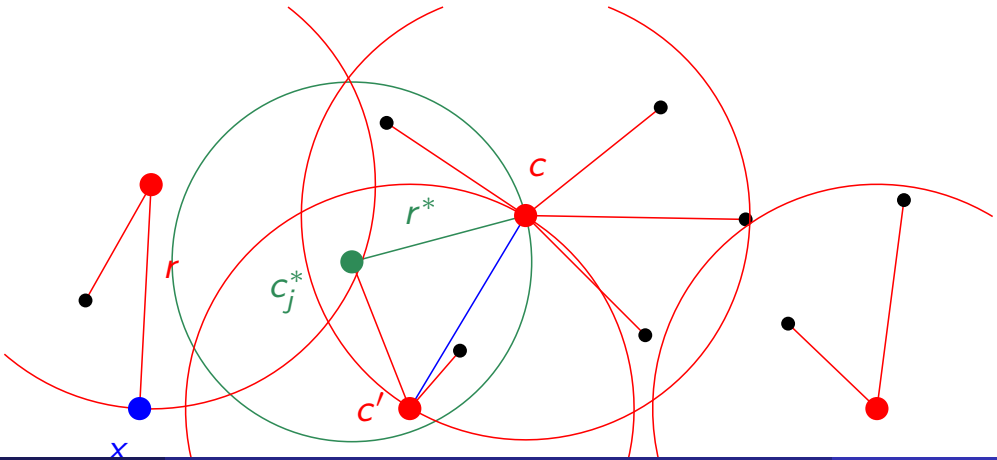


定理：アルゴリズムの近似比

このアルゴリズムは必ず 2 近似解を出力する

証明 (続)：

- このとき，$C = \{c_1, c_2, \ldots, c_k, x\}$ とすると，
  任意の $c, c' \in C$ に対して，$d(c, c') \geq r$
- 一方で，$|C| = k + 1$ であるから，ある j に対して，
  2 つの要素 $c, c' \in C$ が存在して，$d(c, c_j^*) \leq r^*, d(c', c_j^*) \leq r^*$
- 三角不等式を用いると，
  $r \leq d(c, c') \leq d(c, c_j^*) + d(c', c_j^*) \leq 2r^*$ □



目次

「アルゴリズムの限界」ということばの意味

- 特定のアルゴリズムに対する限界
- 任意のアルゴリズムに対する限界

ここでは,「特定のアルゴリズムに対する限界」を見る

目標

いま考えたアルゴリズムの近似比が 2 よりも小さくないことを証明する

証明の方針：アルゴリズムが 2 よりもよい近似比を持つ解を出力しないような問題例を構成する

近似アルゴリズムの限界：$k = 1$

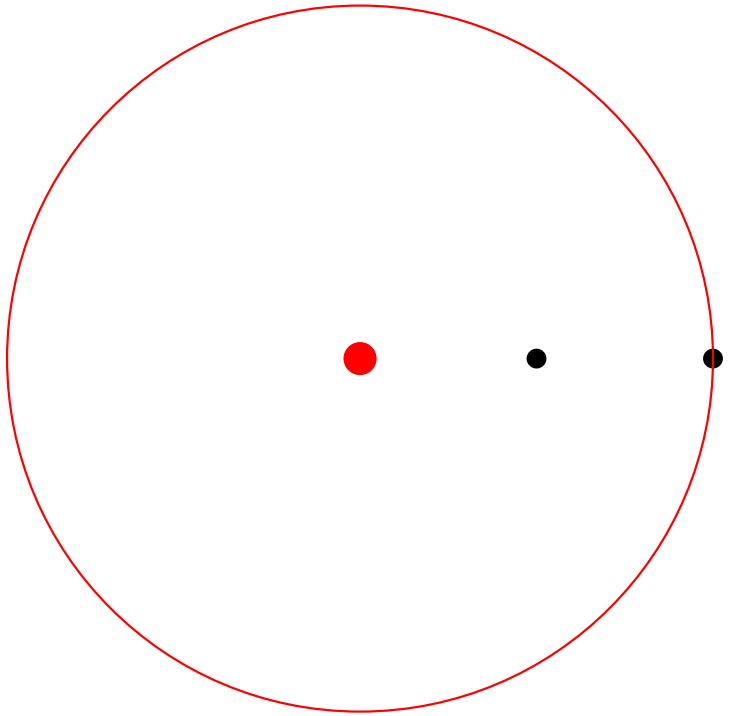
$k = 1$ のとき：

近似アルゴリズムの限界：$k = 1$

$k = 1$ のとき：

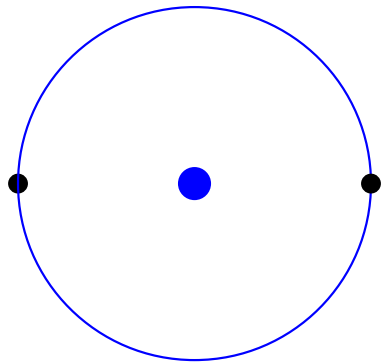
近似アルゴリズムの限界：$k = 1$

$k = 1$ のとき：



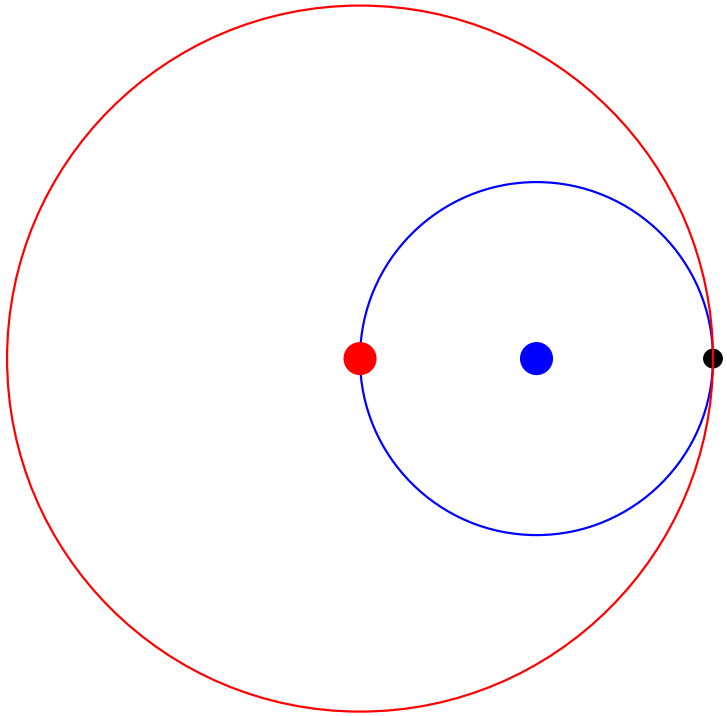
近似アルゴリズムの限界：$k = 1$

$k = 1$ のとき：



近似アルゴリズムの限界：$k = 1$

$k = 1$ のとき：



アルゴリズムの出力 $= 2$, 最適値 ≤ 1

近似アルゴリズムの限界：$k = 2$

$k = 2$ のとき：

近似アルゴリズムの限界：$k = 2$

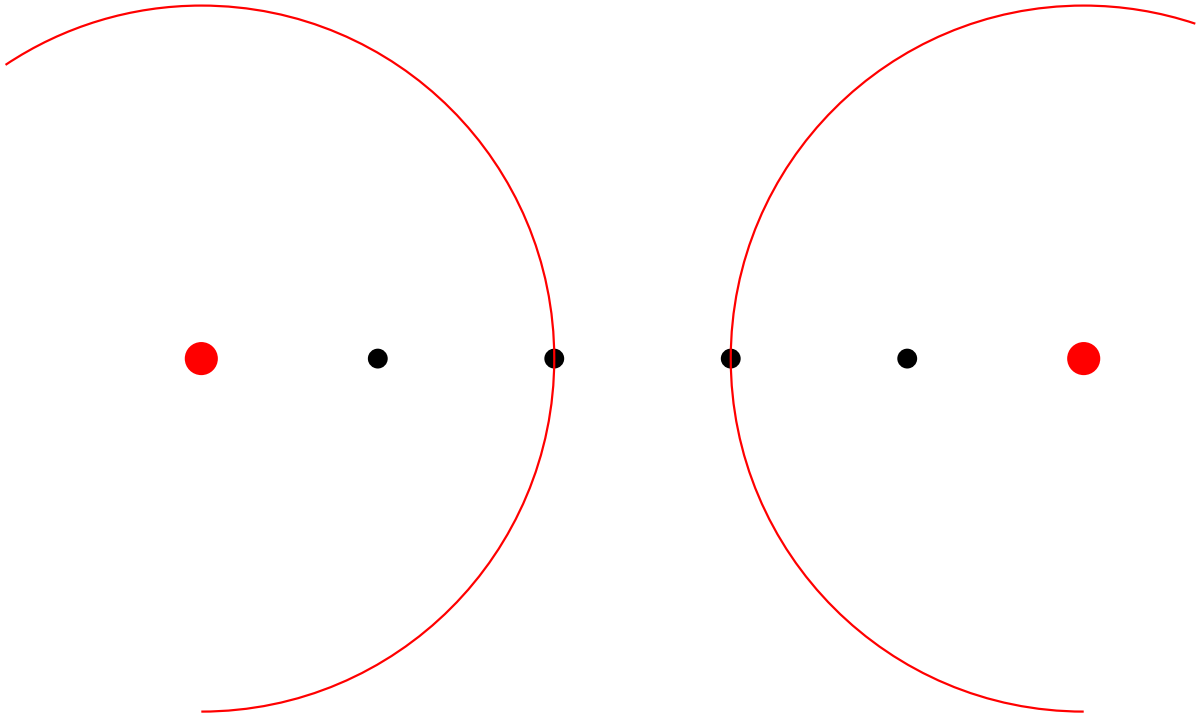
$k = 2$ のとき：

近似アルゴリズムの限界：$k = 2$

$k = 2$ のとき：

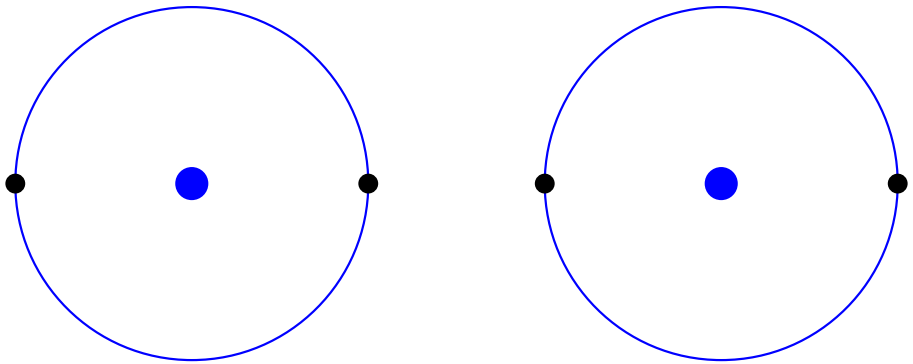
近似アルゴリズムの限界：$k = 2$

$k = 2$ のとき：



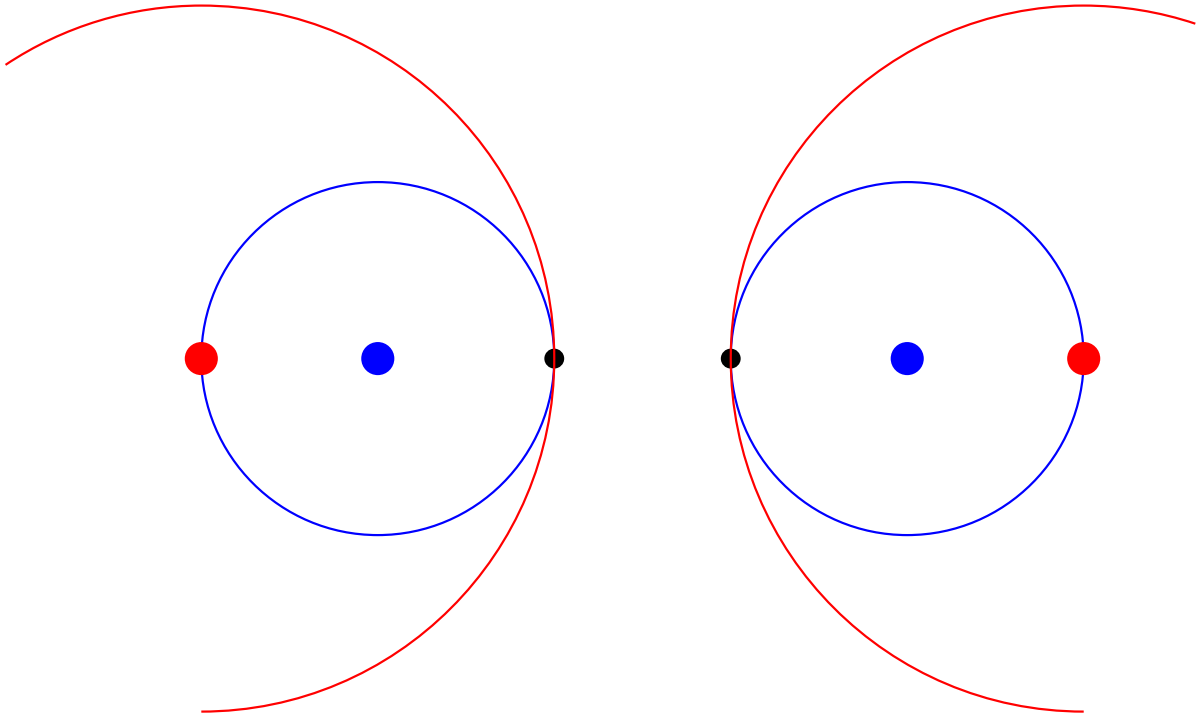
近似アルゴリズムの限界：$k = 2$

$k = 2$ のとき：



近似アルゴリズムの限界：$k = 2$

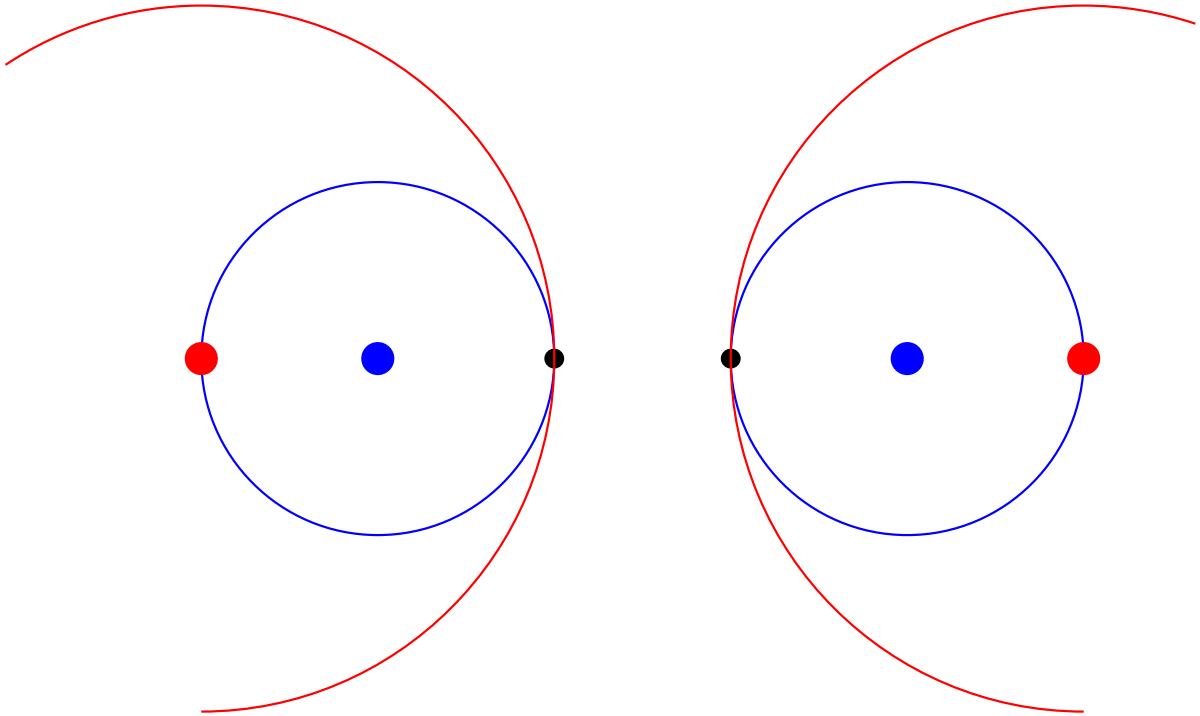
$k = 2$ のとき：



アルゴリズムの出力 $= 2$, 最適値 ≤ 1

近似アルゴリズムの限界：$k = 2$

$k = 2$ のとき：



アルゴリズムの出力 $= 2$, 最適値 ≤ 1

$k \geq 3$ のとき，演習問題

目次

クラスタリング

- クラスタリング：様々な最適化モデル
  - k-センター，k-メディアン，k-ミーンズ
- k-センター：近似アルゴリズム
- k-センター：近似アルゴリズムの限界

k-センター問題に対する近似アルゴリズムは
次の論文に基づく

- Teofilo F. Gonzalez, Clustering to Minimize the Maximum Intercluster Distance. Theor. Comput. Sci. **38** 293–306 (1985)

http://www.cs.ucsb.edu/~teo/

「アルゴリズムの限界」ということばの意味

- 特定のアルゴリズムに対する限界
- 任意のアルゴリズムに対する限界

先ほどは,「特定のアルゴリズムに対する限界」を見た

任意のアルゴリズムに対する限界

P $\neq$ NP という仮定の下で，任意の $\epsilon > 0$ に対して

- 多項式時間 $(2-\epsilon)$ 近似アルゴリズムは存在しない (Gonzalez '85)
- ユークリッド平面上に限っても，
  多項式時間 1.822 近似アルゴリズムは存在しない (Feder, Greene '88)

未解決問題

ユークリッド平面上の k-センター問題に対して
2 よりよい近似比を達成する多項式時間アルゴリズムはあるか？

残った時間の使い方

- 演習問題をやる
  - 相談推奨 (ひとりでやらない)
- 質問をする
  - 教員は巡回
- 退室時，小さな紙に感想など書いて提出する ← 重要
  - 内容は何でも OK
  - 匿名で OK

目次