7 - Clustering problems

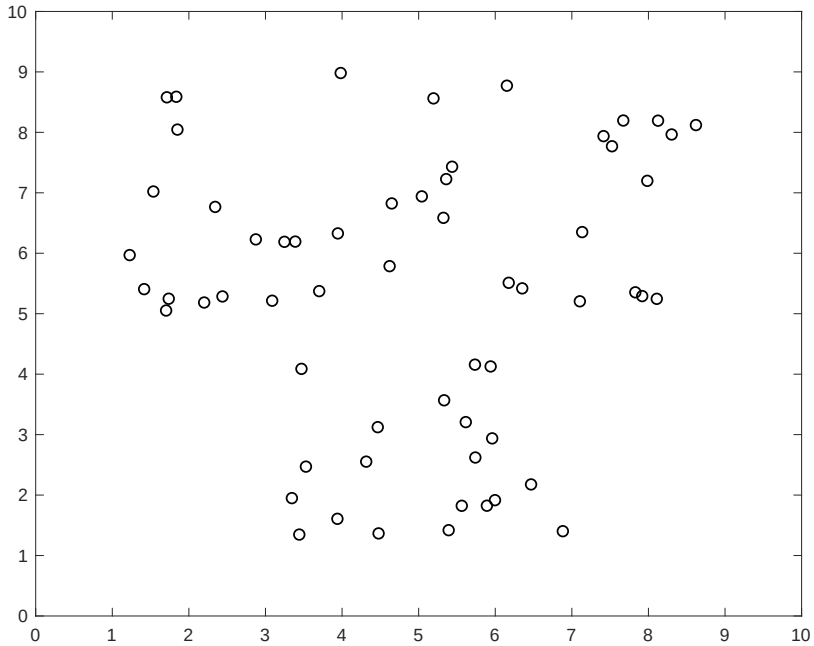
G. Mastroeni and M. Passacantando

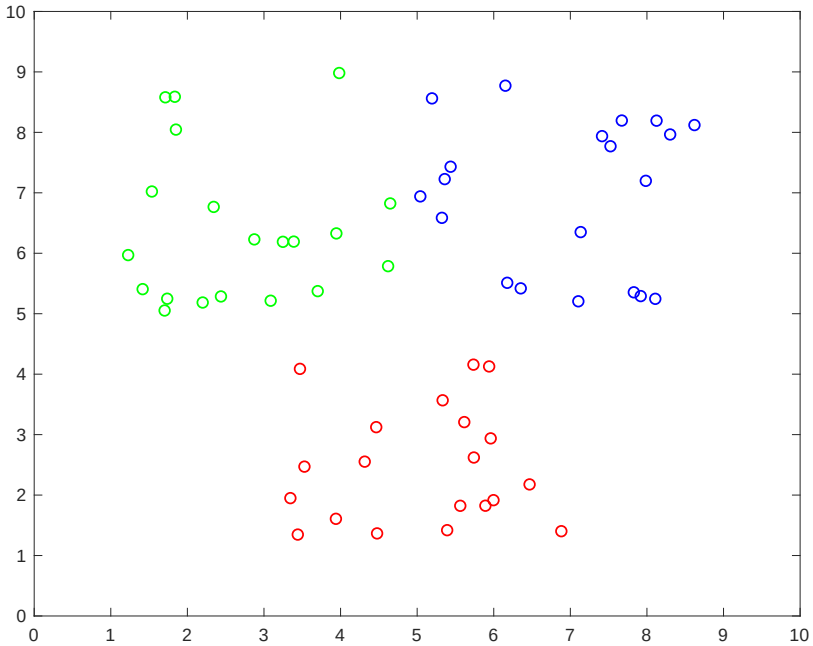
Department of Computer Science, University of Pisa

Optimization Methods and Game Theory
Master of Science in Artificial Intelligence and Data Engineering
University of Pisa – A.Y. 2023/24

Definition

Given a set S of patterns and an integer number k, a clustering problem consists in finding a partition of S in k subsets $S_1, \ldots, S_k$ (clusters) that are homogeneous and well separated.





Clustering problem is of interest in unsupervisioned machine learning.

The optimization model

- Assume that patterns are vectors $p_1, \ldots, p_\ell \in \mathbb{R}^n$.
- Consider a distance $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ between vectors in $\mathbb{R}^n$.
  In what follows we will consider the square distance $d(x, y) = \|x - y\|_2^2$ or $d(x, y) = \|x - y\|_1$.
- For each cluster S_j introduce a centroid $x_j \in \mathbb{R}^n$ (unknown).
- Define clusters so that each pattern is associated to the closest centroid.

We aim to find k centroids in order to minimize the sum of the distances between each pattern and the closest centroid.

The model

$$\begin{cases} \min \displaystyle\sum_{i=1}^{\ell} \min_{j=1,\ldots,k} d(p_i, x_j) \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1, \ldots, k \end{cases}$$

The optimization model with $\|\cdot\|_2$

Consider the square distance $d(x, y) = \|x - y\|_2^2$.
The optimization problem to solve is

$$\begin{cases} \min \sum_{i=1}^{\ell} \min_{j=1,\dots,k} \|p_i - x_j\|_2^2 \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1, \dots, k \end{cases}$$

If $k = 1$ (one cluster), then it is a convex quadratic programming problem without constraints:

$$\begin{cases} \min \sum_{i=1}^{\ell} \|p_i - x\|_2^2 = \min \sum_{i=1}^{\ell} (x - p_i)^{\mathsf{T}}(x - p_i) \\ x \in \mathbb{R}^n \end{cases} \tag{1}$$

The global optimum is the stationary point:

$$2\ell x - 2\sum_{i=1}^{\ell} p_i = 0 \quad \Longleftrightarrow \quad x = \frac{\sum_{i=1}^{\ell} p_i}{\ell} \quad \text{(mean or baricenter)}$$

The optimization model with $\|\cdot\|_2$

If $k > 1$ (at least two clusters), then the problem is nonconvex and nondifferentiable:

$$\begin{cases} \min\limits_{x} \sum\limits_{i=1}^{\ell} \min\limits_{j=1,\dots,k} \|p_i - x_j\|_2^2 \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1, \dots, k \end{cases} \tag{2}$$

We observe that for fixed p_i and x_j, $j = 1, .., k$,

$$\min_{j=1,\dots,k} \|p_i - x_j\|_2^2 = \begin{cases} \min \sum\limits_{j=1}^{k} \alpha_{ij} \|p_i - x_j\|_2^2 \\ \sum\limits_{j=1}^{k} \alpha_{ij} = 1 \\ \alpha_{ij} \geq 0 \qquad \forall\ \ j = 1, \dots, k \end{cases} \tag{3}$$

Remark

It is enough to notice that $\min\limits_{j=1,\dots,k} \{a_j\} = \min \left\{ \sum\limits_{j=1}^{k} \alpha_j a_j : \ \sum\limits_{j=1}^{k} \alpha_j = 1,\ \alpha \geq 0 \right\}$.

An optimal solution of (3) is given by

$$\alpha_{ij}^* = \begin{cases} 1 & \text{if } \|p_i - x_j\|_2 = \min\limits_{h=1,\dots,k} \|p_i - x_h\|_2 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Remark

Observe that $\alpha_{ij}^* = 1$ if pattern i is assigned to cluster j.

Theorem

Problem (2) is equivalent to the following nonconvex differentiable problem:

$$\begin{cases} \min\limits_{x,\alpha} \; f(x,\alpha) := \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{k} \alpha_{ij} \|p_i - x_j\|_2^2 \\ \sum\limits_{j=1}^{k} \alpha_{ij} = 1 \quad \forall\, i = 1,\dots,\ell \\ \alpha_{ij} \geq 0 \qquad \forall\, i = 1,\dots,\ell,\; j = 1,\dots,k \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1,\dots,k. \end{cases} \tag{5}$$

The k-means algorithm is based on the following properties of problem (5):

- If x_j are fixed, then (5) is decomposable into ℓ simple LP problems of the form (3) : for any $i = 1, \ldots, \ell$, the optimal solution is

$$\alpha_{ij}^* = \begin{cases} 1 & \text{if } j \text{ is the first index s.t. } \|p_i - x_j\|_2 = \min\limits_{h=1,\ldots,k} \|p_i - x_h\|_2 \\ & (x_j \text{ is the first closest centroid to } p_i), \\ 0 & \text{otherwise.} \end{cases}$$

- If α_{ij} are fixed, then (5) is decomposable into k convex QP problems (in the unknown x_j) similar to (1), i.e.,

$$\begin{cases} \min \sum\limits_{i=1}^{\ell} \alpha_{ij} \|p_i - x_j\|_2^2 = \min \sum\limits_{i=1}^{\ell} \alpha_{ij} (x_j - p_i)^{\mathsf{T}} (x_j - p_i) \\ x_j \in \mathbb{R}^n \end{cases} \tag{6}$$

For any $j = 1, \ldots, k$, the optimal solution of (6) is

$$x_j^* = \frac{\sum\limits_{i=1}^{\ell} \alpha_{ij} p_i}{\sum\limits_{i=1}^{\ell} \alpha_{ij}} \qquad \text{(mean of patterns)}.$$

Clustering problem – k-means algorithm

The k-means algorithm consists in an alternating minimization of $f(x,\alpha) = \sum_{i=1}^{\ell}\sum_{j=1}^{k} \alpha_{ij}\|p_i - x_j\|_2^2$ with respect to the two blocks of variables x and α.

0. (Inizialization) Set $t = 0$, choose centroids $x_1^0, \ldots, x_k^0 \in \mathbb{R}^n$ and assign patterns to clusters: for any $i = 1, \ldots, \ell$

$$\alpha_{ij}^0 = \begin{cases} 1 & \text{if } j \text{ is the first index s.t. } \|p_i - x_j^0\|_2 = \min\limits_{h=1,\ldots,k} \|p_i - x_h^0\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

1. (Update centroids) For each $j = 1, \ldots, k$ compute the mean

$$x_j^{t+1} = \left(\sum_{i=1}^{\ell} \alpha_{ij}^t p_i\right) / \left(\sum_{i=1}^{\ell} \alpha_{ij}^t\right).$$

2. (Update clusters) For any $i = 1, \ldots, \ell$ compute

$$\alpha_{ij}^{t+1} = \begin{cases} 1 & \text{if } j \text{ is the first index s.t. } \|p_i - x_j^{t+1}\|_2 = \min\limits_{h=1,\ldots,k} \|p_i - x_h^{t+1}\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

3. (Stopping criterion) If $f(x^{t+1}, \alpha^{t+1}) = f(x^t, \alpha^t)$ then STOP
else $t = t + 1$, go to Step 1.

Theorem

The k-means algorithm stops after a finite number of iterations at a solution (x^*, α^*) of the KKT system of problem (5) such that

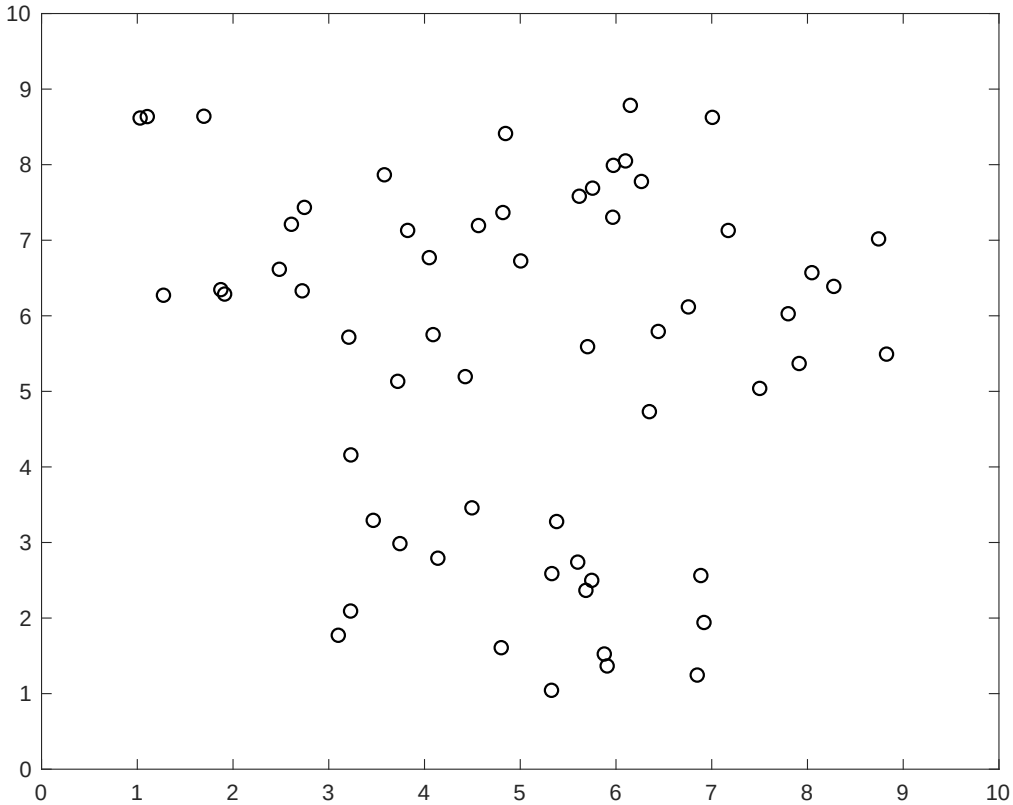
$$f(x^*, \alpha^*) \leq f(x^*, \alpha), \qquad \forall\ \alpha \geq 0 \ \text{ s.t. } \sum_{j=1}^{k} \alpha_{ij} = 1 \quad \forall\ i = 1, \ldots, \ell,$$

$$f(x^*, \alpha^*) \leq f(x, \alpha^*), \qquad \forall\ x \in \mathbb{R}^{kn}.$$

Remark. The k-means algorithm does not guarantee to find a global optimum.

Exercise 7.1. Consider the k-means algorithm, with $k = 3$, for the following set of patterns

$$\begin{pmatrix} 1.2734 & 6.2721 \\ 2.7453 & 7.4345 \\ 1.6954 & 8.6408 \\ 1.1044 & 8.6364 \\ 4.8187 & 7.3664 \\ 2.7224 & 6.3303 \\ 4.8462 & 8.4123 \\ 4.0497 & 6.7696 \\ 1.0294 & 8.6174 \\ 3.7202 & 5.1327 \\ 3.8238 & 7.1297 \\ 3.5805 & 7.8660 \\ 3.2092 & 5.7172 \\ 1.8724 & 6.3461 \\ 4.0895 & 5.7509 \\ 1.9121 & 6.2877 \\ 2.4835 & 6.6154 \\ 4.5637 & 7.1943 \\ 4.4255 & 5.1950 \\ 2.6097 & 7.2109 \end{pmatrix} \quad \begin{pmatrix} 6.0992 & 8.0496 \\ 5.9660 & 7.3042 \\ 5.9726 & 7.9907 \\ 5.6166 & 7.5821 \\ 8.8257 & 5.4929 \\ 8.7426 & 7.0176 \\ 8.2749 & 6.3890 \\ 7.9130 & 5.3686 \\ 5.7032 & 5.5914 \\ 6.4415 & 5.7927 \\ 5.7552 & 7.6891 \\ 5.0048 & 6.7260 \\ 6.2657 & 7.7776 \\ 7.7985 & 6.0271 \\ 7.5010 & 5.0390 \\ 7.1722 & 7.1291 \\ 6.7561 & 6.1176 \\ 6.1497 & 8.7849 \\ 7.0066 & 8.6258 \\ 8.0462 & 6.5707 \end{pmatrix} \quad \begin{pmatrix} 3.0994 & 1.7722 \\ 5.6857 & 2.3666 \\ 6.3487 & 4.7316 \\ 6.8860 & 2.5627 \\ 3.2277 & 2.0929 \\ 4.8013 & 1.6078 \\ 5.3299 & 2.5884 \\ 5.7466 & 2.4989 \\ 5.8777 & 1.5245 \\ 5.6002 & 2.7402 \\ 5.9077 & 1.3661 \\ 4.4954 & 3.4585 \\ 5.3263 & 1.0439 \\ 3.4645 & 3.2930 \\ 3.2306 & 4.1589 \\ 6.9191 & 1.9415 \\ 4.1393 & 2.7921 \\ 5.3799 & 3.2774 \\ 6.8486 & 1.2456 \\ 3.7431 & 2.9852 \end{pmatrix}$$



a) Run the algorithm starting from centroids $x_1 = (5, 7)$, $x_2 = (6, 3)$, $x_3 = (4, 3)$.

b) Run the algorithm starting from centroids $x_1 = (5, 7)$, $x_2 = (6, 3)$, $x_3 = (4, 4)$.

c) Is it possible to improve the solutions obtained in a) and b)?

Matlab commands

```
data=[...];

k=3;   % number of clusters

InitialCentroids=[5,7;6,3;4,3];

[x,cluster,v] = kmeans1(data,k,InitialCentroids)

% plot centroids
plot(x(1,1),x(1,2),'b*',x(2,1),x(2,2),'r*',x(3,1),x(3,2),'g*');

hold on

% plot clusters

c1 = data(cluster==1,:);

c2 = data(cluster==2,:);

c3 = data(cluster==3,:);

plot(c1(:,1),c1(:,2),'bo',c2(:,1),c2(:,2),'ro',c3(:,1),c3(:,2),'go');
```

```
function [x,cluster,v] = kmeans1(data,k,InitialCentroids)

l = size(data,1); % number of patterns

x = InitialCentroids; % initialize centroids

% initialize clusters

cluster = zeros(l,1);

for i = 1 : l
  d = inf;
   for j = 1 : k
    if norm(data(i,:)-x(j,:)) < d
   d = norm(data(i,:)-x(j,:));
   cluster(i) = j;
    end
  end
end

% compute the objective function value

vold = 0;
for i = 1 : l
  vold = vold + norm(data(i,:)-x(cluster(i),:))^ 2 ;
end


while true
```

```
% update centroids
for j = 1 : k
  ind = find(cluster == j);
   if isempty(ind)==0
    x(j,:) = mean(data(ind,:),1);
   end
end

% update clusters
for i = 1 : l
  d = inf;
   for j = 1 : k
    if norm(data(i,:)-x(j,:)) < d
    d = norm(data(i,:)-x(j,:));
    cluster(i) = j;
    end
   end
end

% update objective function
v = 0;
for i = 1 : l
  v = v + norm(data(i,:)-x(cluster(i),:))^ 2 ;
end
```

```
% stopping criterion

if vold - v < 1e-5
  break
else
  vold = v;
end

end

end
```

Clustering problem – optimization model with $\|\cdot\|_1$

Consider now the distance $d(x,y) = \|x-y\|_1$.
The optimization problem to solve is

$$\begin{cases} \min \sum_{i=1}^{\ell} \min_{j=1,\dots,k} \|p_i - x_j\|_1 \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1,\dots,k \end{cases}$$

If $k = 1$ (one cluster), then it is a convex problem decomposable into n convex problems of one variable:

$$\begin{cases} \min \sum_{i=1}^{\ell} \|p_i - x\|_1 = \min \sum_{i=1}^{\ell} \sum_{h=1}^{n} |x_h - (p_i)_h| = \min \sum_{h=1}^{n} \underbrace{\sum_{i=1}^{\ell} |x_h - (p_i)_h|}_{f_h(x_h)} \\ x \in \mathbb{R}^n \end{cases} \tag{7}$$

Clustering problem – optimization model with $\|\cdot\|_1$

Given ℓ real numbers $a_1 < a_2 < \cdots < a_\ell$, what is the optimal solution of

$$\begin{cases} \min \sum_{i=1}^{\ell} |x - a_i| = f(x) \\ x \in \mathbb{R} \end{cases} \qquad ?$$

The objective function is convex and piecewise linear:

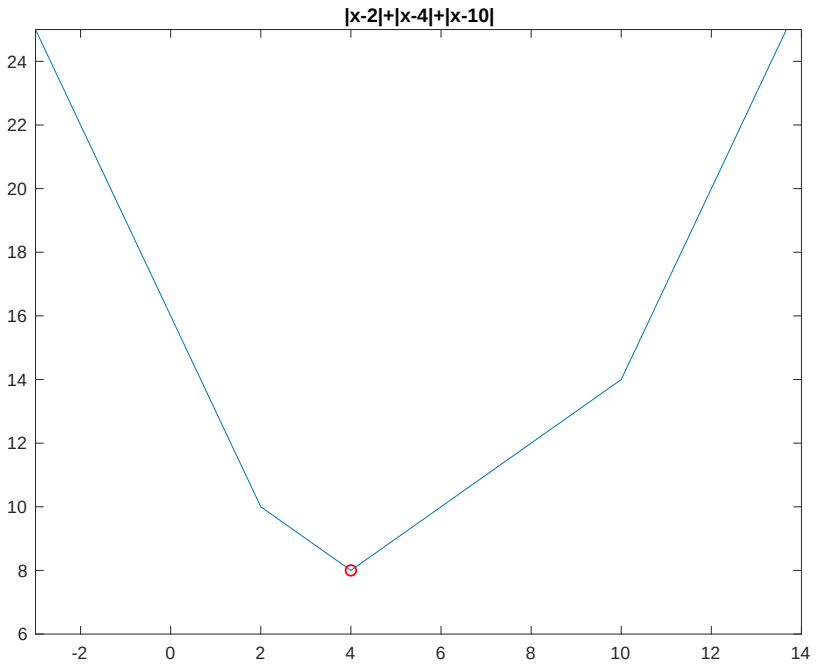
$$f(x) = \begin{cases} -\ell x + \sum_{i=1}^{\ell} a_i & \text{if } x < a_1 \\ (2-\ell)x + \sum_{i=2}^{\ell} a_i - a_1 & \text{if } x \in [a_1, a_2] \\ \dots & \dots \\ (2r-\ell)x + \sum_{i=r+1}^{\ell} a_i - \sum_{i=1}^{r} a_1 & \text{if } x \in [a_r, a_{r+1}] \\ \dots & \dots \\ (\ell-2)x + a_\ell - \sum_{i=1}^{\ell-1} a_i & \text{if } x \in [a_{\ell-1}, a_\ell] \\ \ell x - \sum_{i=1}^{\ell} a_i & \text{if } x > a_\ell \end{cases}$$

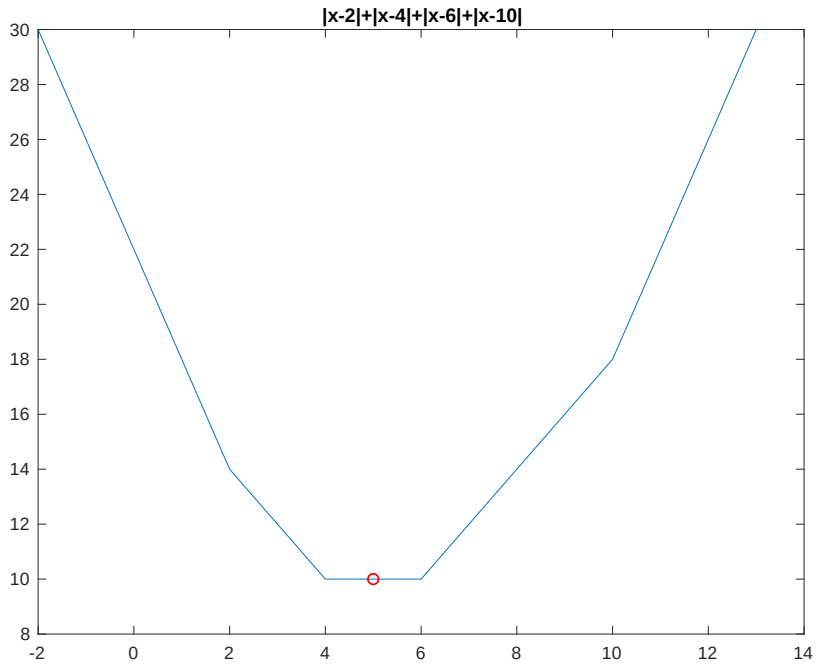
The global optimum is $\text{median}(a_1, \dots, a_\ell) = \begin{cases} a_{(\ell+1)/2} & \text{if } \ell \text{ is odd,} \\ \dfrac{a_{\ell/2} + a_{1+\ell/2}}{2} & \text{if } \ell \text{ is even.} \end{cases}$

Example

(a) $f(x) = |x-2| + |x-4| + |x-10|, \quad \ell = 3;$

(b) $f(x) = |x-2| + |x-4| + |x-6| + |x-10| \quad \ell = 4;$





The global optimum is $\text{median}(a_1, \ldots, a_\ell) = \begin{cases} a_{(\ell+1)/2} & \text{if } \ell = 3, \\ \dfrac{a_{\ell/2} + a_{1+\ell/2}}{2} & \text{if } \ell = 4\ . \end{cases}$

Clustering problem – optimization model with $\|\cdot\|_1$

If $k > 1$ (at least two clusters), then the problem is nonconvex and nonsmooth:

$$\begin{cases} \min\limits_{x} \sum\limits_{i=1}^{\ell} \min\limits_{j=1,\dots,k} \|p_i - x_j\|_1 \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1, \dots, k \end{cases} \tag{8}$$

Theorem

Problem (8) is equivalent to the following problem:

$$\begin{cases} \min\limits_{x,\alpha} \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{k} \alpha_{ij} \|p_i - x_j\|_1 \\ \sum\limits_{j=1}^{k} \alpha_{ij} = 1 \quad \forall\, i = 1, \dots, \ell \\ \alpha_{ij} \geq 0 \qquad \forall\, i = 1, \dots, \ell,\ j = 1, \dots, k \\ x_j \in \mathbb{R}^n \qquad \forall\, j = 1, \dots, k. \end{cases} \tag{9}$$

Note that

$$f(x,\alpha) := \sum_{i=1}^{\ell}\sum_{j=1}^{k} \alpha_{ij}\|p_i - x_j\|_1 = \sum_{i=1}^{\ell}\sum_{j=1}^{k}\sum_{h=1}^{n} \alpha_{ij}\, u_{ijh}$$

where we set

$$u_{ijh} = |(x_j)_h - (p_i)_h| = \max\{(x_j)_h - (p_i)_h, (p_i)_h - (x_j)_h\}$$

Consequently, we have the following result.

Theorem

Problem (9) is equivalent to the nonconvex differentiable (bilinear) problem:

$$\begin{cases} \min\limits_{x,\alpha,u} \sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{k}\sum\limits_{h=1}^{n} \alpha_{ij}\, u_{ijh} \\ u_{ijh} \geq (p_i)_h - (x_j)_h \qquad \forall\ i = 1,\dots,\ell,\ j = 1,\dots,k,\ h = 1,\dots,n \\ u_{ijh} \geq (x_j)_h - (p_i)_h \qquad \forall\ i = 1,\dots,\ell,\ j = 1,\dots,k,\ h = 1,\dots,n \\ \sum\limits_{j=1}^{k} \alpha_{ij} = 1 \qquad \forall\ i = 1,\dots,\ell \\ \alpha_{ij} \geq 0 \qquad \forall\ i = 1,\dots,\ell,\ j = 1,\dots,k \\ x_j \in \mathbb{R}^n \qquad \forall\ j = 1,\dots,k. \end{cases} \tag{10}$$

Clustering problem – k-median algorithm

The k-median algorithm is based on the following properties of problem (9):

- If x_j are fixed, then (9) is decomposable into ℓ simple LP problems: for any $i = 1, \ldots, \ell$, the optimal solution is

$$\alpha_{ij}^* = \begin{cases} 1 & \text{if } j \text{ is the first index s.t. } \|p_i - x_j\|_1 = \min\limits_{h=1,\ldots,k} \|p_i - x_h\|_1 \\ & \quad (x_j \text{ is the first closest centroid to } p_i), \\ 0 & \text{otherwise.} \end{cases}$$

- If $\alpha_{ij} \in \{0, 1\}$ are fixed, then (9) is decomposable into k simple convex problems similar to (7), i.e.,

$$\begin{cases} \min \sum\limits_{i=1}^{\ell} \alpha_{ij} \|p_i - x_j\|_1 = \min \sum\limits_{i=1}^{\ell} \sum\limits_{h=1}^{n} \alpha_{ij} |(x_j)_h - (p_i)_h| \\ x_j \in \mathbb{R}^n \end{cases} \tag{11}$$

For any $j = 1, \ldots, k$, the optimal solution is

$$x_j^* = \text{median}(p_i : \ \alpha_{ij} = 1).$$

Clustering problem – k-median algorithm

The k-median algorithm consists in an alternating minimization of $f(x,\alpha)=\sum_{i=1}^{\ell}\sum_{j=1}^{k}\alpha_{ij}\|p_i-x_j\|_1$ with respect to the two blocks of variables x and α.

0. (Inizialization) Set $t=0$, choose centroids $x_1^0,\dots,x_k^0\in\mathbb{R}^n$ and assign patterns to clusters: for any $i=1,\dots,\ell$

$$\alpha_{ij}^0=\begin{cases}1 & \text{if } j \text{ is the first index s.t. } \|p_i-x_j^0\|_1=\min\limits_{h=1,\dots,k}\|p_i-x_h^0\|_1\\ 0 & \text{otherwise.}\end{cases}$$

1. (Update centroids) For each $j=1,\dots,k$ compute

$$x_j^{t+1}=\text{median}(p_i:\ \alpha_{ij}^t=1).$$

2. (Update clusters) For any $i=1,\dots,\ell$ compute

$$\alpha_{ij}^{t+1}=\begin{cases}1 & \text{if } j \text{ is the first index s.t. } \|p_i-x_j^{t+1}\|_1=\min\limits_{h=1,\dots,k}\|p_i-x_h^{t+1}\|_1\\ 0 & \text{otherwise.}\end{cases}$$

3. (Stopping criterion) If $f(x^{t+1},\alpha^{t+1})=f(x^t,\alpha^t)$ then STOP
else $t=t+1$, go to Step 1.

Theorem

The k-median algorithm stops after a finite number of iterations at a stationary point (x^*, α^*) of problem (8) such that

$$f(x^*, \alpha^*) \leq f(x^*, \alpha), \qquad \forall\ \alpha \geq 0 \ \text{ s.t. } \sum_{j=1}^{k} \alpha_{ij} = 1 \quad \forall\ i = 1, \ldots, \ell,$$

$$f(x^*, \alpha^*) \leq f(x, \alpha^*), \qquad \forall\ x \in \mathbb{R}^{kn}.$$

Remark.

The k-median algorithm does not guarantee to find a global optimum.

Exercise 7.2. Consider the k-median algorithm, with $k = 3$, for the set of patterns given in Exercise 7.1.

a) Run the algorithm starting from centroids $x_1 = (5, 7)$, $x_2 = (6, 3)$, $x_3 = (4, 3)$.

b) Run the algorithm starting from centroids $x_1 = (5, 7)$, $x_2 = (6, 3)$, $x_3 = (4, 4)$.

c) Is it possible to improve the solutions obtained in a) and b)?

Matlab commands

Notice that the Matlab implementation of the k-median algorithm is analogous to the one of the k-means algorithm.
The changes with respect to the Matlab solution of Exercise 7.1 are given in blue.

```
data=[...];

k=3;   % number of clusters

InitialCentroids=[5,7;6,3;4,3];

[x,cluster,v] = kmedian2(data,k,InitialCentroids)

% plot centroids
plot(x(1,1),x(1,2),'b*',x(2,1),x(2,2),'r*',x(3,1),x(3,2),'g*');

hold on

% plot clusters

c1 = data(cluster==1,:);
c2 = data(cluster==2,:);
c3 = data(cluster==3,:);

plot(c1(:,1),c1(:,2),'bo',c2(:,1),c2(:,2),'ro',c3(:,1),c3(:,2),'go');
```

```
function [x,cluster,v] = kmedian2(data,k,InitialCentroids)

l = size(data,1); % number of patterns

x = InitialCentroids; % initialize centroids

% initialize clusters

cluster = zeros(l,1);

for i = 1 : l
  d = inf;
   for j = 1 : k
    if norm(data(i,:)-x(j,:),1) < d
   d = norm(data(i,:)-x(j,:),1);
   cluster(i) = j;
    end
  end
end

% compute the objective function value

vold = 0;
for i = 1 : l
  vold = vold + norm(data(i,:)-x(cluster(i),:),1) ;
end


while true
```

```
% update centroids
for j = 1 : k
  ind = find(cluster == j);
   if isempty(ind)==0
    x(j,:) = median(data(ind,:),1);
   end
end

% update clusters
for i = 1 : l
  d = inf;
   for j = 1 : k
    if norm(data(i,:)-x(j,:),1) < d
    d = norm(data(i,:)-x(j,:),1);
    cluster(i) = j;
    end
   end
end

% update objective function
v = 0;
for i = 1 : l
  v = v + norm(data(i,:)-x(cluster(i),:),1) ;
end
```

```
% stopping criterion

if vold - v < 1e-5
  break
else
  vold = v;
end

end

end
```