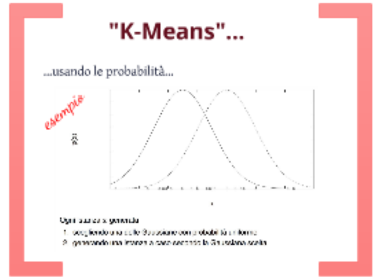
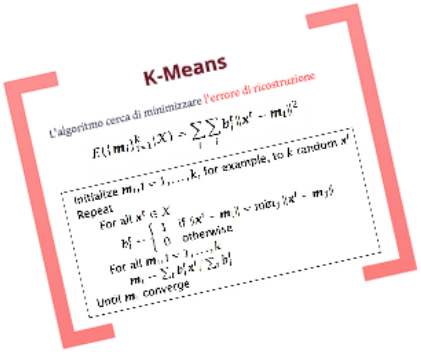
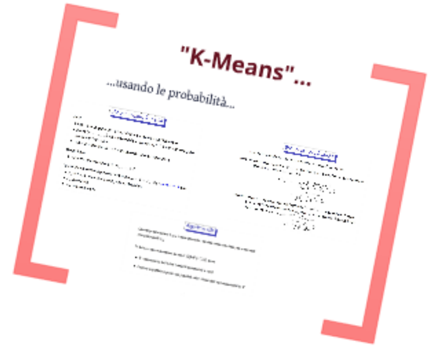
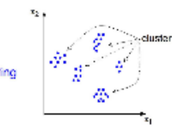






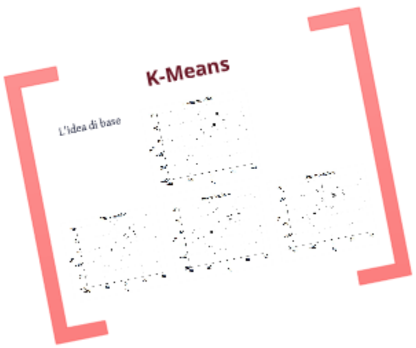
Apprendimento non Supervisionato

- dato in insieme di esempi $Tr = \{x^{(i)}\}$, estrarre regolarità e/o pattern (valide(i) su tutto il dominio di ingresso)
- non esiste nessun esperto (o maestro) che ci fornisca un aiuto
  - Clustering



  - Scoperta di Regole (Discovery)

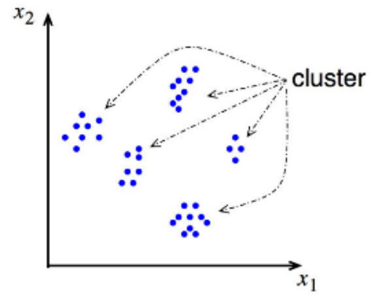
Esempio di applicazione: data mining su database strutturati



Clustering

Apprendimento non Supervisionato

- dato in insieme di esempi $Tr = \{x^{(i)}\}$, estrarre regolarità e/o pattern (valide(i) su tutto il dominio di ingresso)
- non esiste nessun esperto (o maestro) che ci fornisca un aiuto
  - Clustering

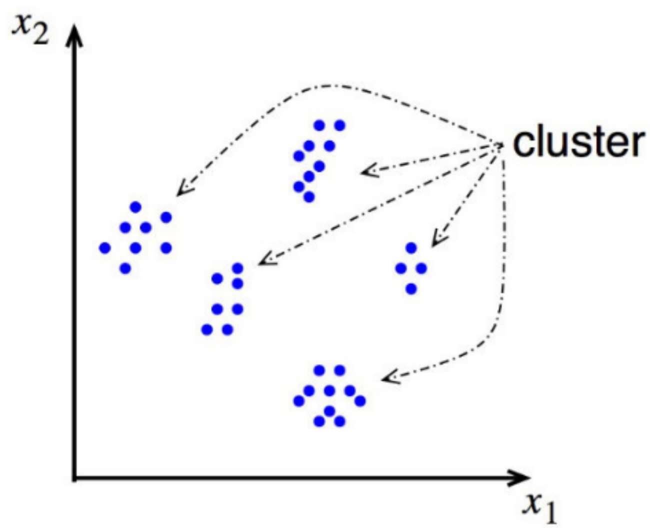


  - Scoperta di Regole (Discovery)

Esempio di applicazione: data mining su database strutturati

- dato in insieme di esempi $Tr = \{x^{(i)}\}$, estrarre regolarità e/o pattern (valide(i) su tutto il dominio di ingresso)
- non esiste nessun esperto (o maestro) che ci fornisca un aiuto

  - Clustering

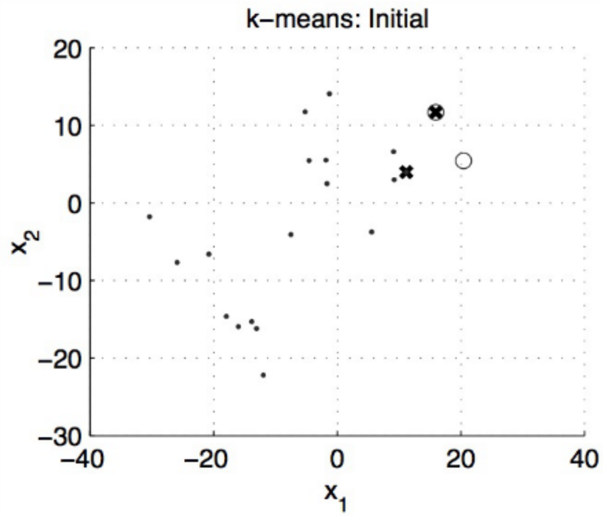


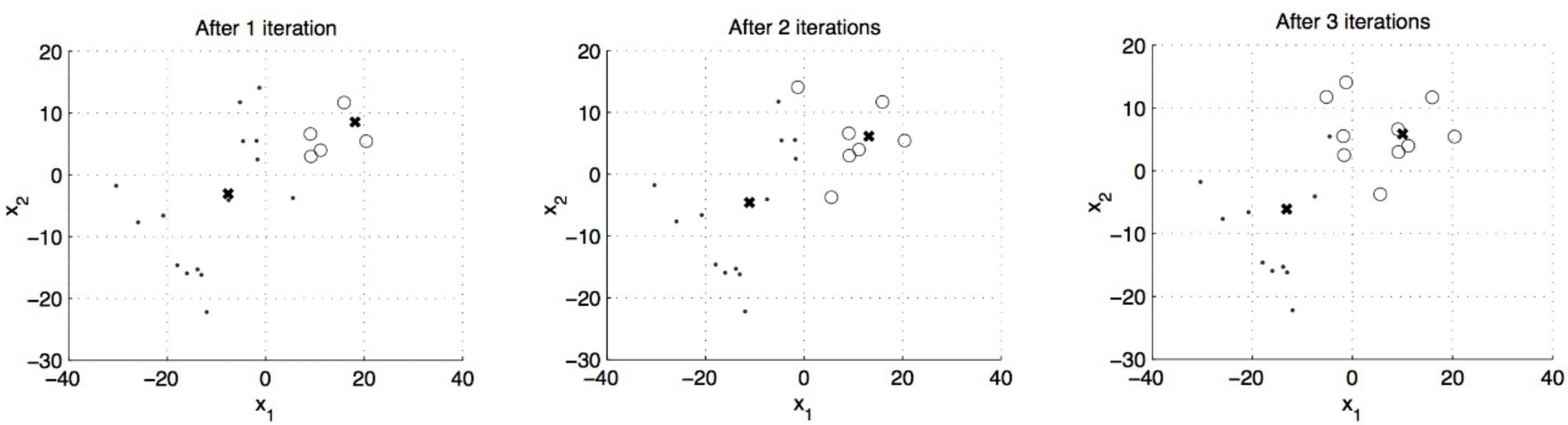
  - Scoperta di Regole (Discovery)

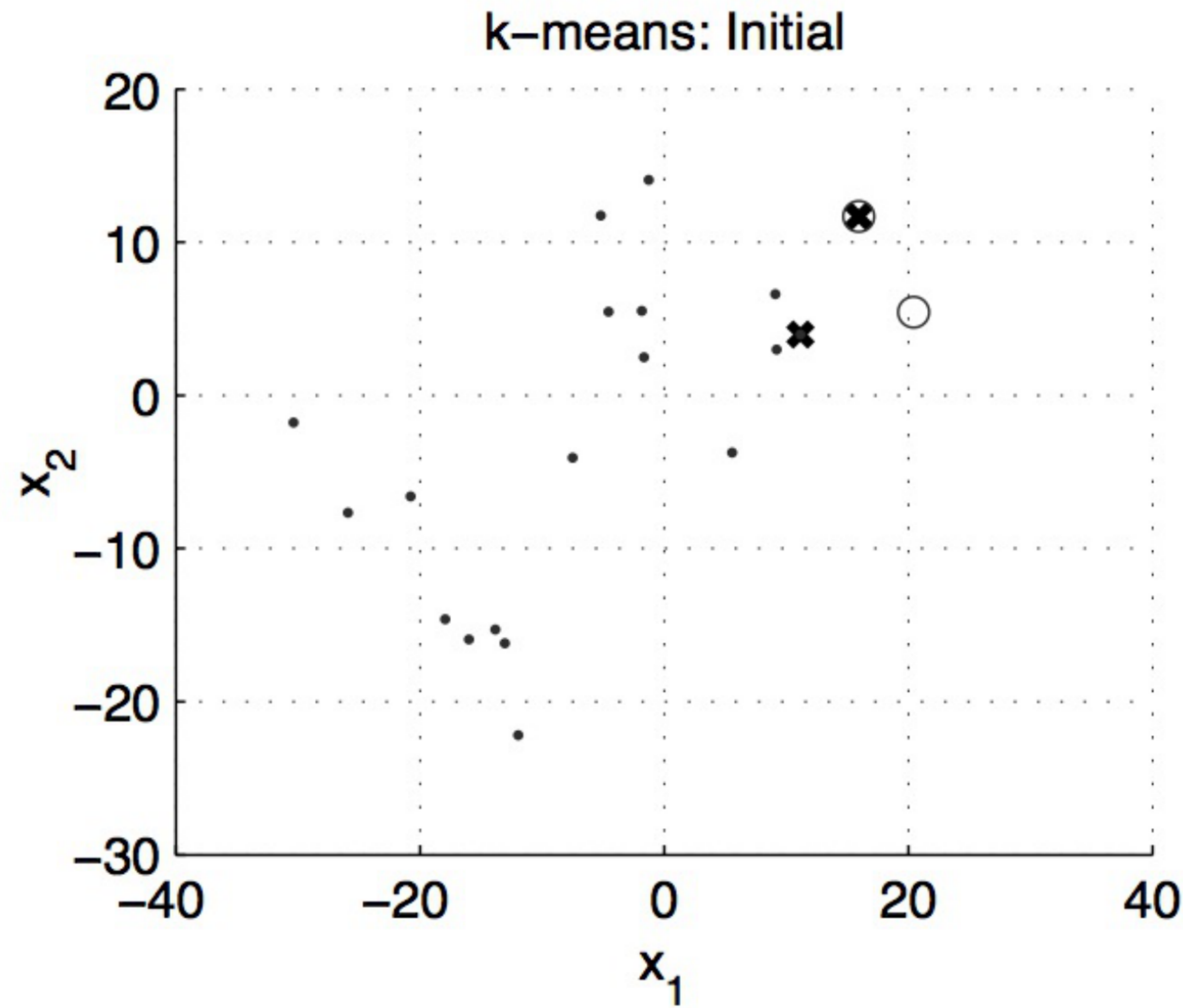
Esempio di applicazione: data mining su database strutturati

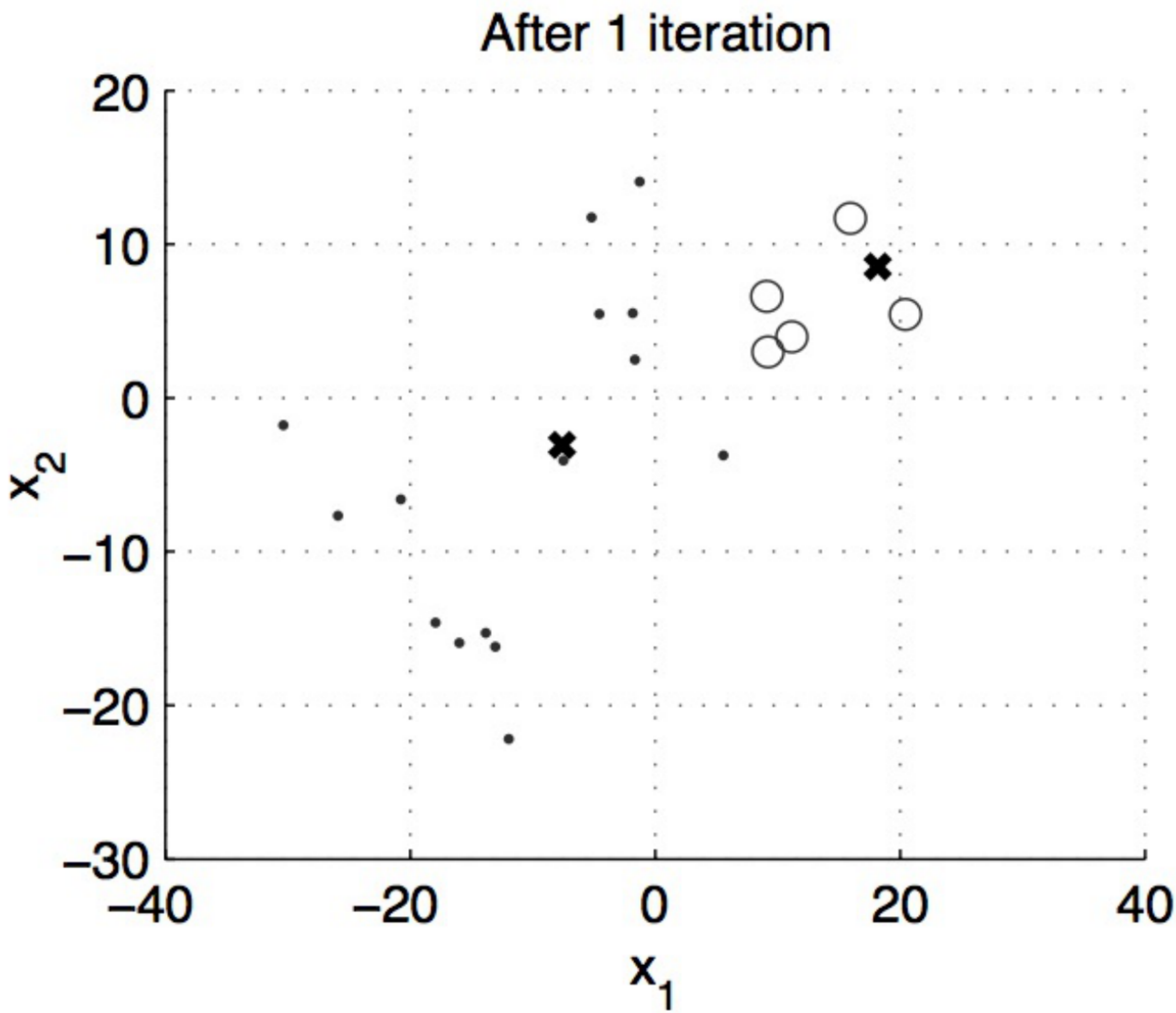
K-Means

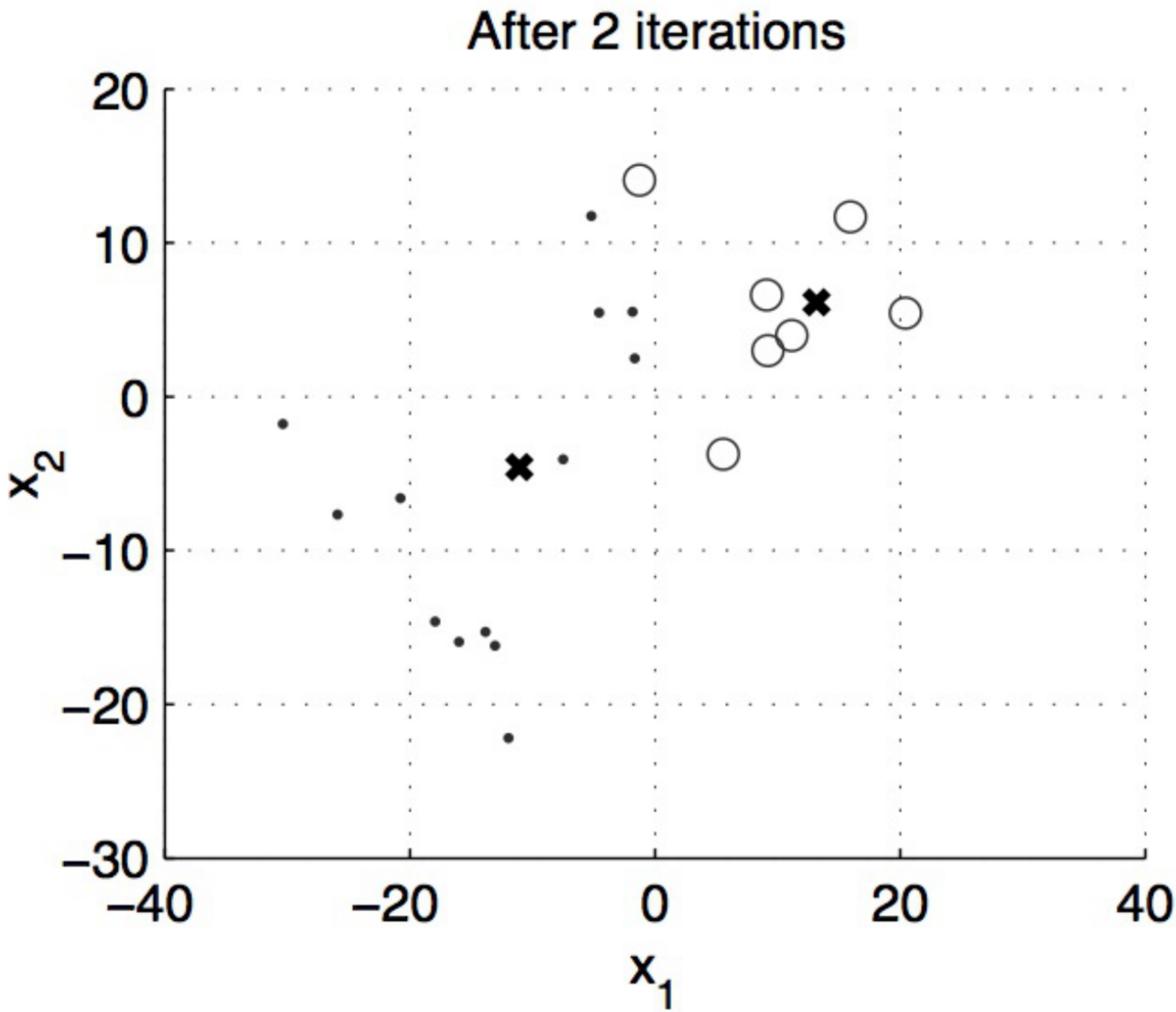
L'idea di base

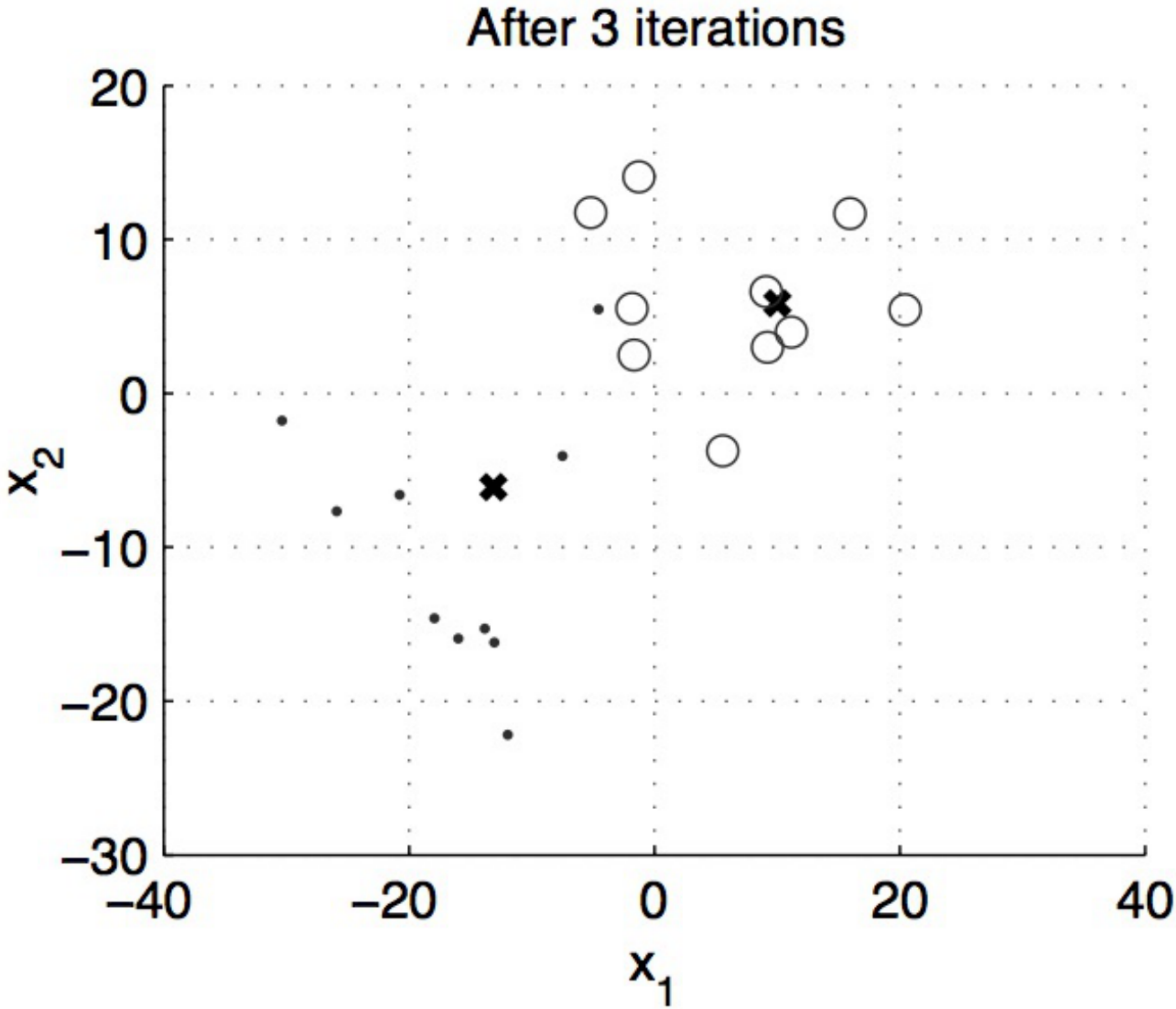





k−means: Initial
20
10
0
−10
−20
−30
−40
−20
0
20
40
x_1
x_2


After 1 iteration
20
10
0
−10
−20
−30
−40
−20
0
20
40
x_1
x_2


After 2 iterations
20
10
0
−10
−20
−30
−40
−20
0
20
40
x2
x1


After 3 iterations
20
10
0
−10
−20
−30
−40
−20
0
20
40
x_2
x_1

K-Means

L'algoritmo cerca di minimizzare l'errore di ricostruzione

$$E(\{\boldsymbol{m}_i\}_{i=1}^{k} | \mathcal{X}) = \sum_t \sum_i b_i^t \|\boldsymbol{x}^t - \boldsymbol{m}_i\|^2$$

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$

Repeat

  For all $\boldsymbol{x}^t \in \mathcal{X}$

$$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

  For all $\boldsymbol{m}_i, i = 1, \ldots, k$

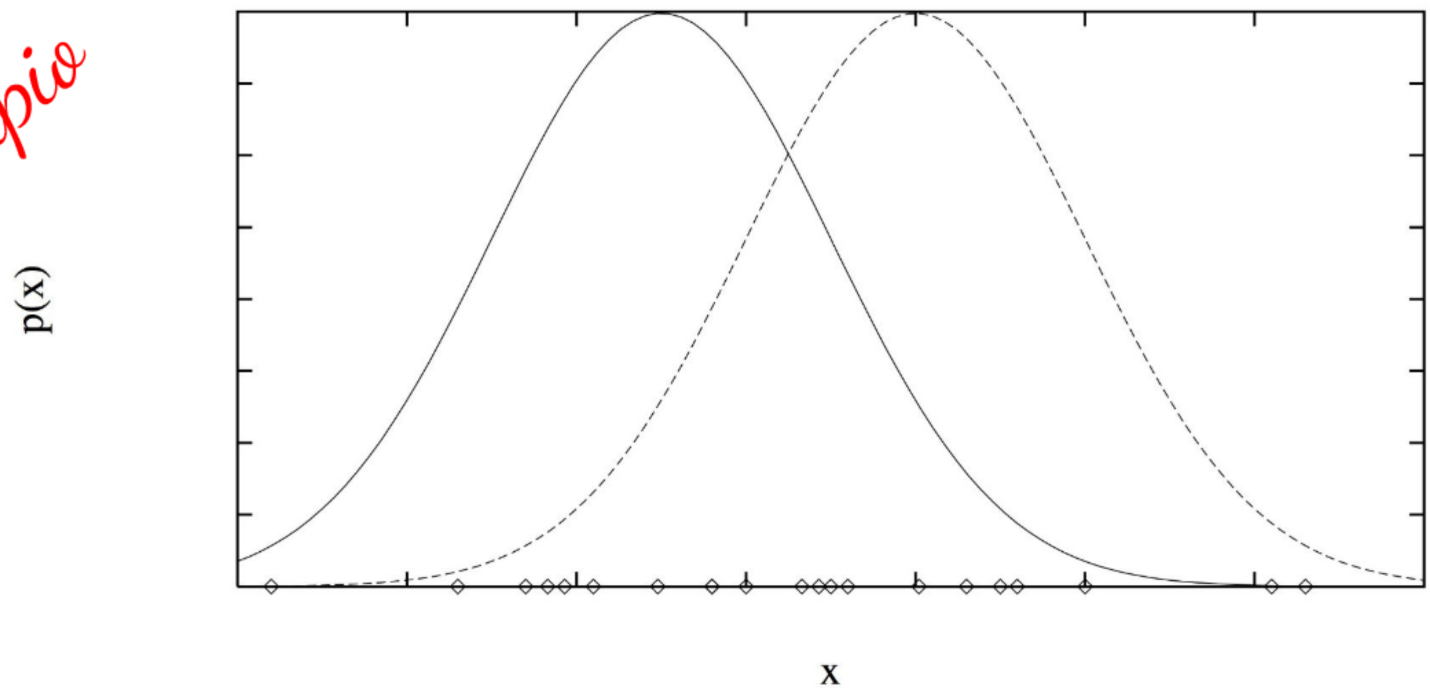
$$\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$$

Until $\boldsymbol{m}_i$ converge

"K-Means"...

...usando le probabilità...

esempio



Ogni istanza x generata

1. scegliendo una delle Gaussiane con probabilità uniforme
2. generando una istanza a caso secondo la Gaussiana scelta

"K-Means"...

...usando le probabilità...

EM per stimare k medie

Date:

- istanze da X generate da una mistura di k distribuzioni Gaussiane
- medie $\langle \mu_1, \ldots, \mu_k \rangle$ sconosciute delle k Gaussiane (σ^2 conosciuto ed uguale per tutte le Gaussiane)
- non si sa quale istanza x_i è stata generata da quale Gaussiana

Determinare:

- stime maximum likelihood di $\langle \mu_1, \ldots, \mu_k \rangle$

ogni istanza può essere pensata nella forma $y_i = \langle x_i, z_{i1}, z_{i2} \rangle$ (caso $k = 2$), dove

- z_{ij} è 1 se x_i è generata dalla j-esima Gaussiana
- x_i osservabile
- z_{ij} non osservabile

EM per stimare k medie

Algoritmo EM: scegliere a caso l'ipotesi iniziale $h = \langle \mu_1, \mu_2 \rangle$, poi ripetere

passo E: calcola il valore aspettato $E[z_{ij}]$ di ogni variabile non osservabile z_{ij}, assumendo che valga l'ipotesi corrente $h = \langle \mu_1, \mu_2 \rangle$

$$E[z_{ij}] = \frac{p(x = x_i | \mu = \mu_j)}{\sum_{n=1}^{2} p(x = x_i | \mu = \mu_n)} = \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}}{\sum_{n=1}^{2} e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}}$$

passo M: calcola la nuova ipotesi maximum likelihood $h' = \langle \mu'_1, \mu'_2 \rangle$, assumendo che il valore preso da ogni variabile non osservabile z_{ij} sia il suo valore aspettato $E[z_{ij}]$ (calcolato sopra). Rimpiazza $h = \langle \mu_1, \mu_2 \rangle$ con $h' = \langle \mu'_1, \mu'_2 \rangle$.

$$\mu_j \leftarrow \frac{\sum_{i=1}^{m} E[z_{ij}]\ x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

Algoritmo EM

Converge alla ipotesi h_{ML} locale (massimo locale) fornendo stime per le variabili non osservabili z_{ij}

Di fatto, trova un massimo locale di $E[\ln P(Y|h)]$, dove

- Y rappresenta tutti i dati (variabili osservabili e non)
- il valore aspettato è preso sui possibili valori di variabili non osservabili in Y

EM per stimare k medie

Date:

- istanze da X generate da una mistura di k distribuzioni Gaussiane
- medie $\langle \mu_1, \ldots, \mu_k \rangle$ sconosciute delle k Gaussiane (σ^2 conosciuto ed uguale per tutte le Gaussiane)
- non si sa quale istanza x_i è stata generata da quale Gaussiana

Determinare:

- stime maximum likelihood di $\langle \mu_1, \ldots, \mu_k \rangle$

ogni istanza può essere pensata nella forma $y_i = \langle x_i, z_{i1}, z_{i2} \rangle$ (caso $k = 2$), dove

- z_{ij} è 1 se x_i è generata dalla j-esima Gaussiana
- x_i osservabile
- z_{ij} non osservabile

EM per stimare k medie

Algoritmo EM: scegliere a caso l'ipotesi iniziale $h = \langle \mu_1, \mu_2 \rangle$, poi ripetere

passo E: calcola il valore aspettato $E[z_{ij}]$ di ogni variabile non osservabile z_{ij}, assumendo che valga l'ipotesi corrente $h = \langle \mu_1, \mu_2 \rangle$

$$
\begin{aligned}
E[z_{ij}] &= \frac{p(x = x_i | \mu = \mu_j)}{\sum_{n=1}^{2} p(x = x_i | \mu = \mu_n)} \\
&= \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}}{\sum_{n=1}^{2} e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}}
\end{aligned}
$$

passo M: calcola la nuova ipotesi maximum likelihood $h' = \langle \mu'_1, \mu'_2 \rangle$, assumendo che il valore preso da ogni variabile non osservabile z_{ij} sia il suo valore aspettato $E[z_{ij}]$ (calcolato sopra). Rimpiazza $h = \langle \mu_1, \mu_2 \rangle$ con $h' = \langle \mu'_1, \mu'_2 \rangle$.

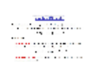
$$
\mu_j \leftarrow \frac{\sum_{i=1}^{m} E[z_{ij}]\ x_i}{\sum_{i=1}^{m} E[z_{ij}]}
$$

Algoritmo EM

Converge alla ipotesi h_{ML} locale (massimo locale) fornendo stime per le variabili non osservabili z_{ij}

Di fatto, trova un massimo locale di $E[\ln P(Y|h)]$, dove

- Y rappresenta tutti i dati (variabili osservabili e non)
- il valore aspettato è preso sui possibili valori di variabili non osservabili in Y



Problema EM in generale

Dati:

- dati osservati $X = \{x_1, \ldots, x_m\}$
- dati non osservabili $Z = \{z_1, \ldots, z_m\}$
- distribuzione di probabilità parametrizzata $P(Y|h)$, dove
  - $Y = \{y_1, \ldots, y_m\}$ è tutto l'insieme dei dati $y_i = x_i \cup z_i$
  - h sono i parametri

Determinare:

- h che massimizza (localmente) $E[\ln P(Y|h)]$

Metodo EM Generale

Definire la funzione di verosimiglianza (likelihood) $Q(h'|h)$ che calcola $Y = X \cup Z$ usando i dati osservati X ed i parametri correnti h per stimare Z

$$Q(h'|h) \leftarrow E[\ln P(Y|h')|h, X]$$

Algoritmo EM:

passo di stima (E): calcolare $Q(h'|h)$ usando l'ipotesi corrente h ed i dati osservati X per stimare la distribuzione di probabilità su Y

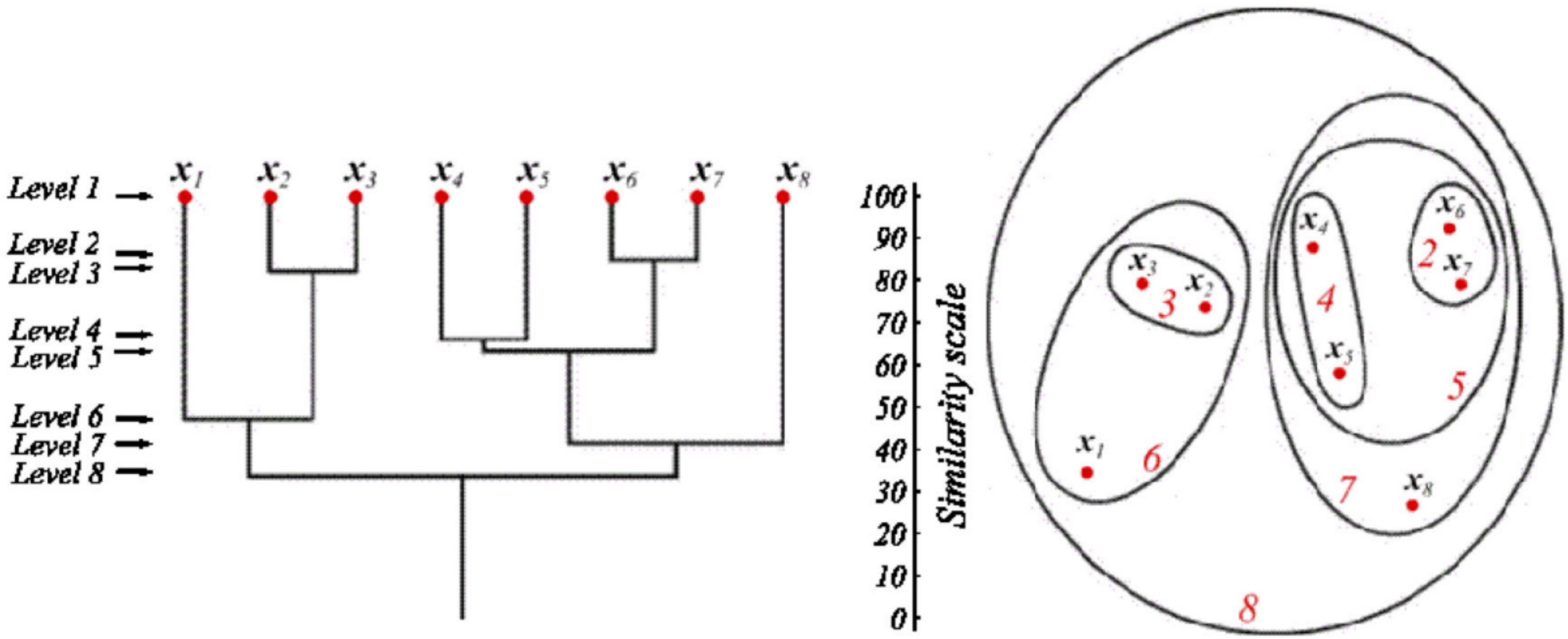
$$Q(h'|h) \leftarrow E[\ln P(Y|h')|h, X]$$

passo di massimizzazione (M): rimpiazza l'ipotesi h tramite l'ipotesi h' che massimizza la funzione Q

$$h \leftarrow \arg\max_{h'} Q(h'|h)$$

Altro esempio di clustering

Hierarchical Clustering



1. si selezionano i 2 vettori più vicini (es., secondo la distanza euclidea) e si costruisce un sottoalbero con figli dati dai due vettori e padre il centroide (media) dei due vettori;

2. i due vettori vengono sostituiti nell'insieme dei vettori correnti dal centroide, e si torna al passo 1.