CLUSTERING

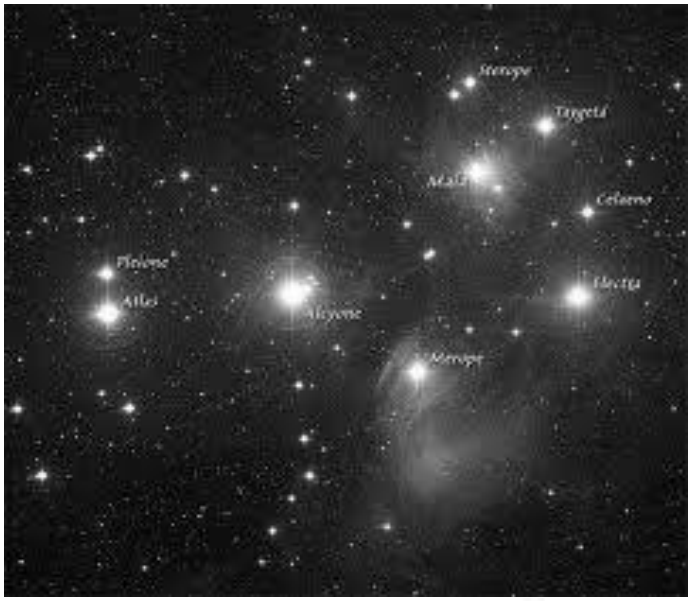
INTRODUZIONE

Introduzione

Raggruppare oggetti in modo tale che, oggetti appartenenti allo stesso gruppo siano simili o in relazione tra loro, oggetti appartenenti a gruppi differenti siano tra loro dissimili o non in relazione tra loro.

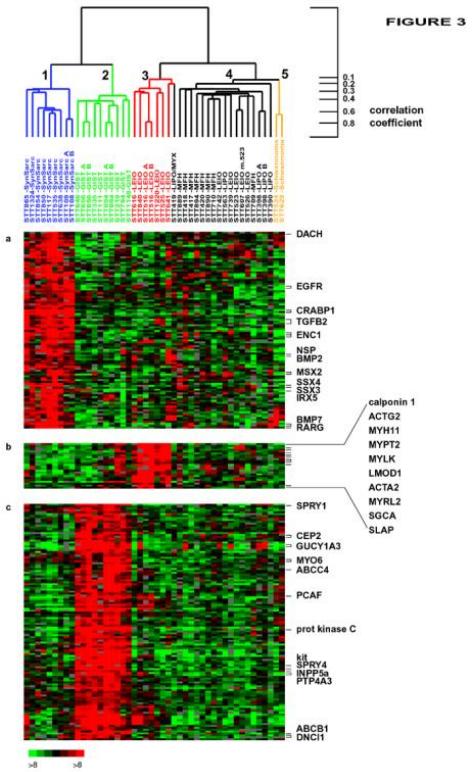
Introduzione

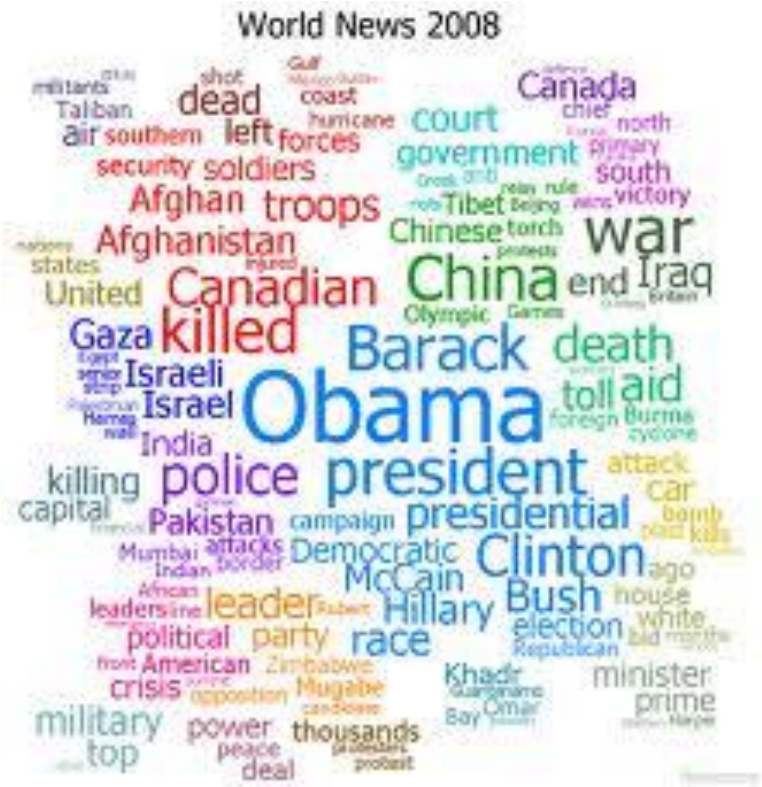
Il clustering si occupa della suddivisione delle osservazioni di un dataset in gruppi omogenei, riferiti con il termine di *cluster*.

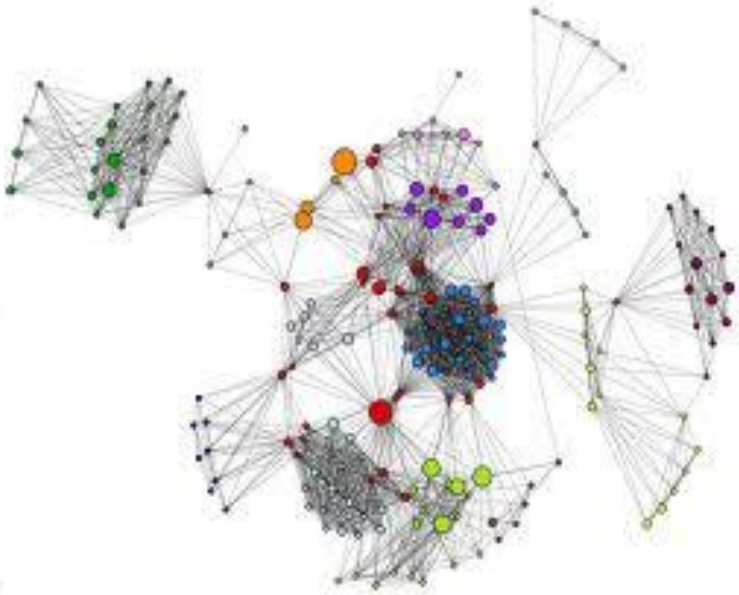


I cluster sono costruiti in modo tale che:

- *osservazioni simili stanno nello stesso cluster*
- *osservazioni dissimili stanno in cluster diversi*







Introduzione

I cluster possono fornire un'interpretazione significativa del fenomeno analizzato (il rappruppamento dei clienti in base a comportamento d'acquisto può evidenziare segmenti di mercato da porre sotto monitoraggio o per i quali valutare operazioni di marketing).

I cluster possono essere propedeutici a successive fasi di data mining, sviluppare modelli di classificazione specifici per ogni cluster.

I cluster possono aiutare a evidenziare osservazioni anomale (outlier) e pertanto il clustering può favorirci nella ripulitura del dataset e nella conseguente riduzione della dimensionalità.

I *metodi di clustering* devono soddisfare i seguenti requisiti:

- ***Flessibilità**: algoritmi applicabili sia a dataset con attributi numerici che a dataset con attributi categorici, la metrica euclidea induce cluster sferici ma fatica con geometrie complesse.*
- ***Robustezza**: stabilità dei cluster generati in funzione di variazioni contenute dei valori degli attributi, robustezza rispetto al rumore presente nei dati.*
- ***Efficienza**: numero di osservazioni tipicamente elevato, costruire cluster in modo efficiente per garantire tempi contenuti.*

Introduzione: tipologie di modelli

Metodi di clustering suddivisi in tipologie base secondo la logica di costruzione dei raggruppamenti:

- ***Metodi di partizione:*** *suddividono il dataset in un numero fisso e noto "K" di cluster non vuoti. Inducono gruppi di forma sferica o convessa, applicabili a dataset di medie dimensioni.*
- ***Metodi gerarchici:*** *ricavano molteplici suddivisioni in cluster, sfruttano la struttura ad albero e utilizzano valori di soglia differenti all'interno di ogni cluster e soglie di disomogeneità tra cluster distinti.*
- ***Metodi basati sulla densità:*** *sfruttano la densità locale per un intorno delle osservazioni, ragionano su un diametro specificato in modo tale che esso contenga un numero di osservazioni non inferiore ad una certa soglia fornita dal decision maker, identificano cluster con forma non convessa e sono in grado di isolare gli outlier.*
- ***Metodi a griglia:*** *discretizzazione preventiva dello spazio delle osservazioni ottenendo celle, offrono riduzioni significative dei tempi di calcolo a scapito di minore accuratezza.*

Introduzione: tipologie di modelli

Una seconda ripartizione riguarda la modalità di assegnazione delle osservazioni ai singoli cluster.

- ***Attribuzione esclusiva:*** *ogni osservazione è assegnata ad un solo cluster.*
- ***Attribuzione soft:*** *ogni osservazione può appartenere a più cluster con diversi gradi di appartenenza.*
- ***Attribuzione completa:*** *ogni osservazione viene assegnata ad almeno un cluster.*
- ***Attribuzione parziale:*** *alcune osservazioni possono essere non assegnate ad alcun cluster, molto utili per identificare presenza di outlier nel dataset.*

La maggior parte dei metodi di clustering ha natura euristica, genera cluster di buona qualità ma è difficile parlare di "*ottimalità*". Un metodo esaustivo per le suddivisioni di "*m*" osservazioni in "*K*" cluster richiede di esaminare un *numero di soluzioni* pari a

Stirling numbers of the second kind

$$\frac{1}{K!}\sum_{h=0}^{K-1}(-1)^h\binom{K}{h}(K-h)^m$$

Introduzione: misure di affinità

I modelli di clustering si basano su una *misura di similarità tra gli oggetti* (osservazioni). In molti casi è possibile ricavare una misura di similarità adottando un'opportuna nozione di distanza tra osservazioni.

Assegnato un dataset "*D*" costituito da "*m*" osservazioni, è possibile rappresentare ogni osservazione tramite un vettore "*n*"-dimensionale, dove "*n*" rappresenta il numero di attributi misurati per ogni oggetto.

Possiamo rappresentare il dataset tramite una matrice rettangolare **X**=[m×n] e computare la matrice quadrata simmetrica di dimensioni [m×m] dove ogni elemento *(i,k)* rappresenta la distanza tra l'oggetto "*i*"-mo e quello "*k*"-mo.

$$\text{Dist} = [d_{ik}] = \begin{bmatrix} 0 & d_{12} & \dots & d_{1(m-1)} & d_{1m} \\ & 0 & \dots & d_{2(m-1)} & d_{2m} \\ & & \dots & \dots & \dots \\ & & & 0 & d_{(m-1)m} \\ & & & & 0 \end{bmatrix} \quad d_{ik} = \text{dist}(x_i, x_k) = \text{dist}(x_k, x_i) \quad i,k = 1,\dots,m$$

CLUSTERING

VALUTAZIONE MODELLI

Validità dei Cluster

Per i metodi di apprendimento supervisionato, classificazione e regressione, valutare l'accuratezza predittiva è parte integrante e centrale del processo di costruzione, sviluppo e validazione di un modello. La valutazione si basa su precisi indicatori numerici.

La stessa cosa non accade così frequentemente nel caso di metodi di apprendimento non supervisionato, la mancanza di una esplicita variabile target rende la valutazione un processo meno diretto e poco intuitivo.

È tuttavia importante prendere in considerazione misure di adeguatezza e significatività per gli algoritmi di clustering:

- *accertarsi che i cluster generati corrispondano ad effettiva regolarità dei dati.*
- *applicare diversi algoritmi di clustering e confrontarne i risultati.*
- *valutare se il numero di cluster identificati è stabile rispetto ai diversi algoritmi.*

Valutazione Modelli

Un indicatore che *combina coesione e separazione* è rappresentato dal ***coefficiente di silhouette***.

Data un'osservazione $\underline{x}_i$, il coefficiente di silhouette viene computato in tre passi.

Calcolo del coefficiente di silhouette

1. Calcolare la distanza media di "$\underline{x}_i$" da tutte le osservazioni appartenenti al suo stesso cluster, sia tale valore "u_i".
2. Per ogni cluster C_f diverso da quello di appartenenza dell'osservazione "$\underline{x}_i$" si calcoli la distanza media tra "$\underline{x}_i$" e tutte le osservazioni di C_f, la si indichi con "w_{if}". Si determini la minima delle distanze "w_{if}" al variare del cluster C_f e la si indichi con "v_i".
3. Si definisce il *coefficiente di silhouette di* "$\underline{x}_i$" come:

$$silh(\underline{x}_i) = \frac{v_i - u_i}{max(v_i, u_i)}$$

Valutazione Modelli

Il ***coefficiente di silhouette*** varia nell'intervallo [-1,+1], valori positivi prossimi a uno sono indice di clusterizzazione ideale.

Il ***coefficiente di silhouette complessivo*** viene computato come media dei coefficienti delle singole osservazioni del dataset "*D*".

CLUSTERING

METODI DI PARTIZIONE

Metodi di Partizione: K-medie

Riceve in ingresso il dataset "*D*", un parametro "*K*" e una funzione di distanza.

Centroide del cluster C_h, $h=1, \ldots, K$, è il punto z_h con coordinate

$$z_{hj} = \frac{\sum_{x_i \in C_h} x_{ij}}{|C_h|}$$

ovvero pari al valore atteso di ciascun attributo computato sulle sole osservazioni assegnate al cluster C_h.

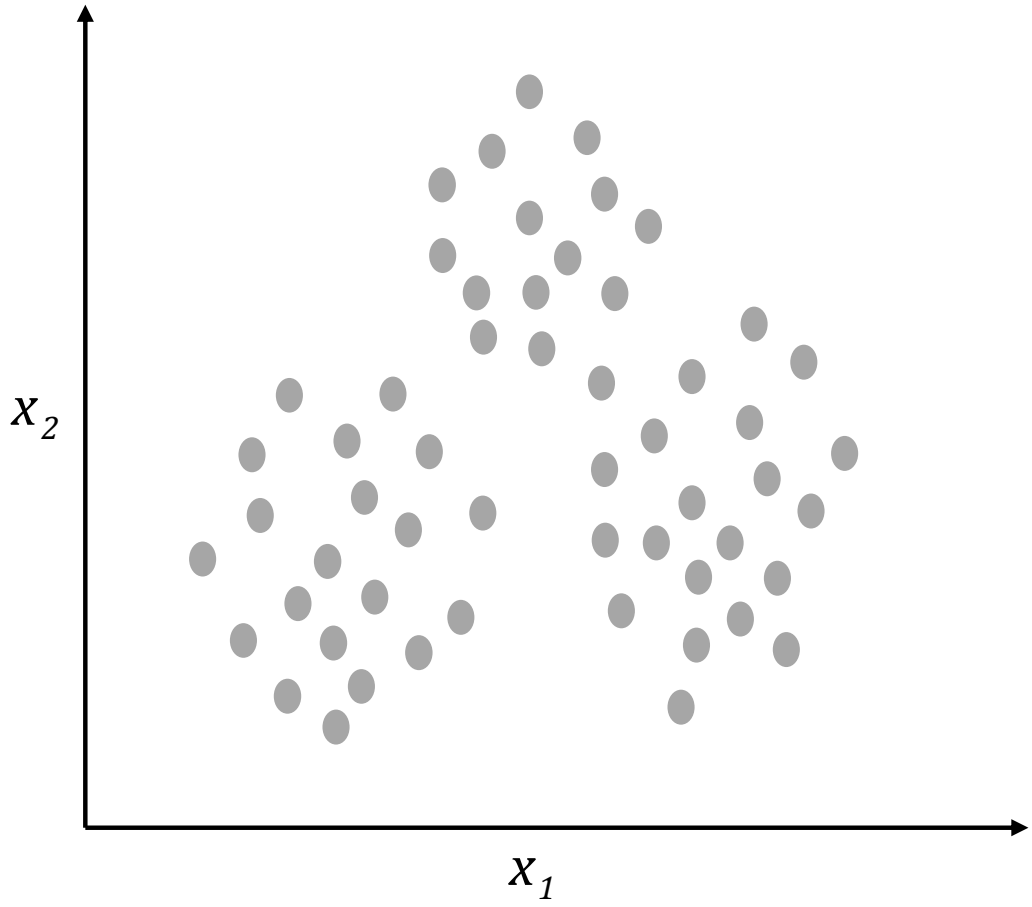
Algoritmo K-medie

1. Scegliere "*K*" osservazioni del dataset "*D*", siano esse i centroidi iniziali dei "*K*" cluster.
2. Assegnare ogni osservazione del dataset "*D*" ad uno dei "*K*" cluster, si assegna l'osservazione al cluster per cui viene minimizzata la distanza dal centroide.
3. Se nessuna osservazione è stata assegnata ad un cluster differente da quello a cui era assegnata all'iterazione precedente l'algoritmo termina.
4. Calcolare per ogni cluster il nuovo centroide, si torni al **Passo 2**.

Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

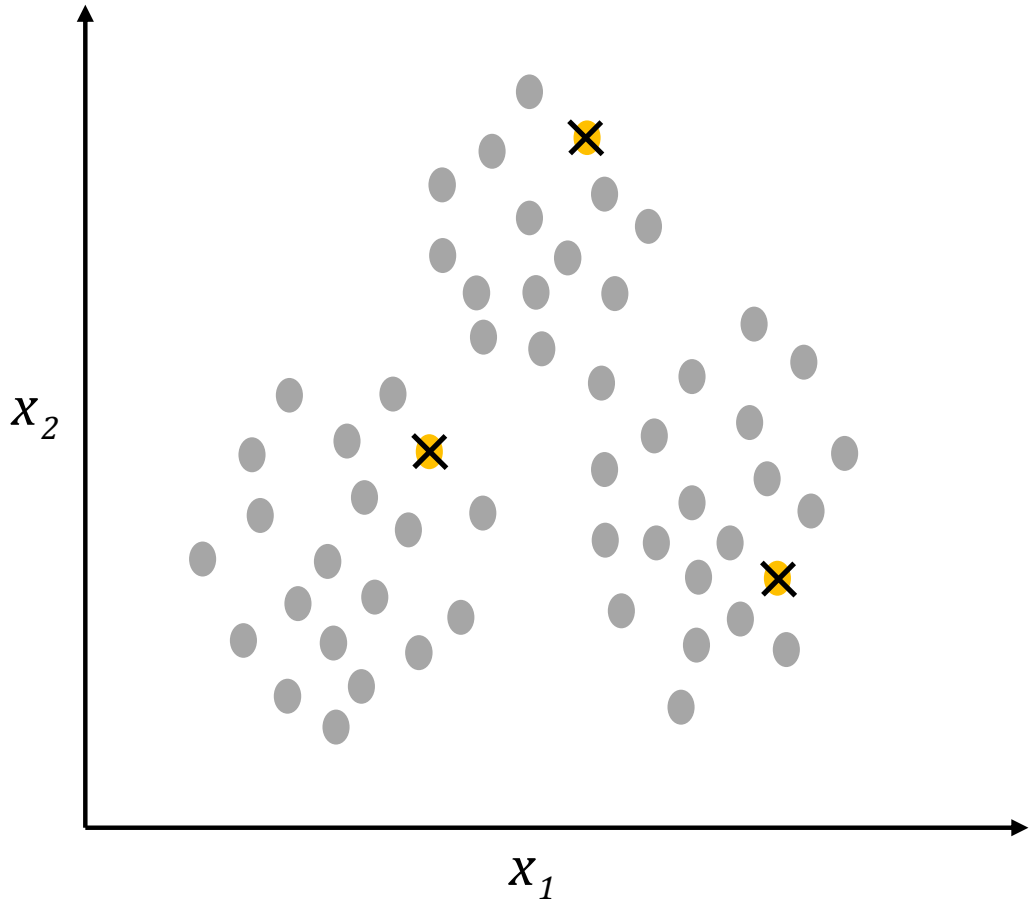
Il dataset rappresentato graficamente sotto



Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

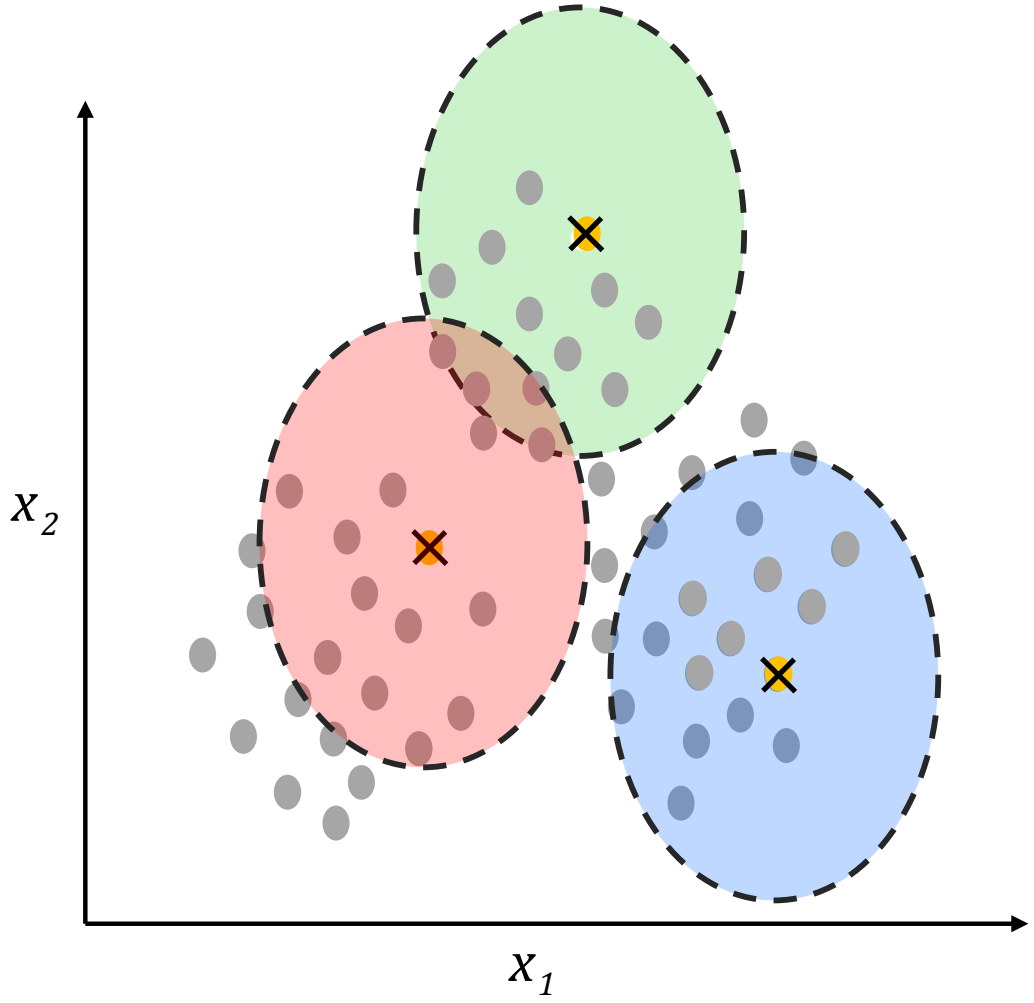
Scegliamo *K=3* osservazioni che fungeranno da centroidi iniziali.



Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

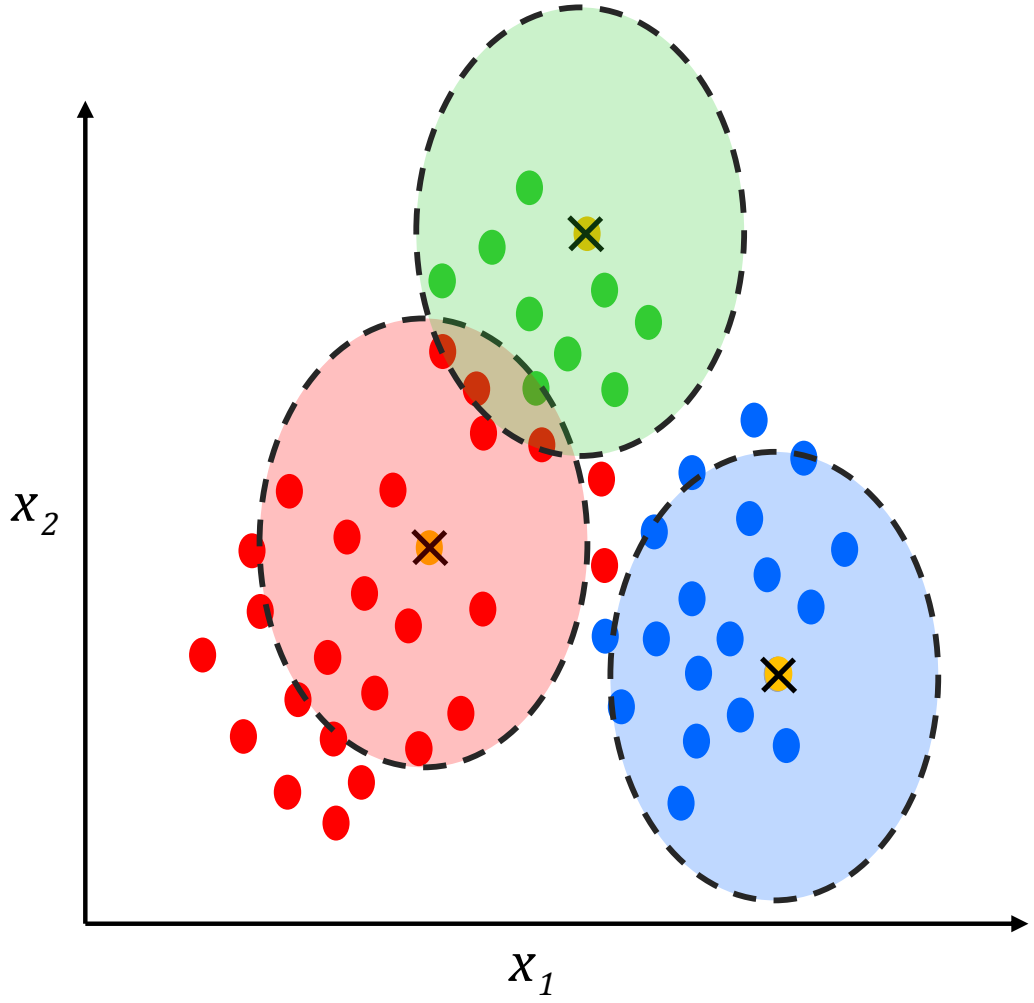
Assegniamo le osservazioni ai cluster identificati dai *3* centroidi



Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

Assegniamo le osservazioni ai cluster identificati dai *3* centroidi

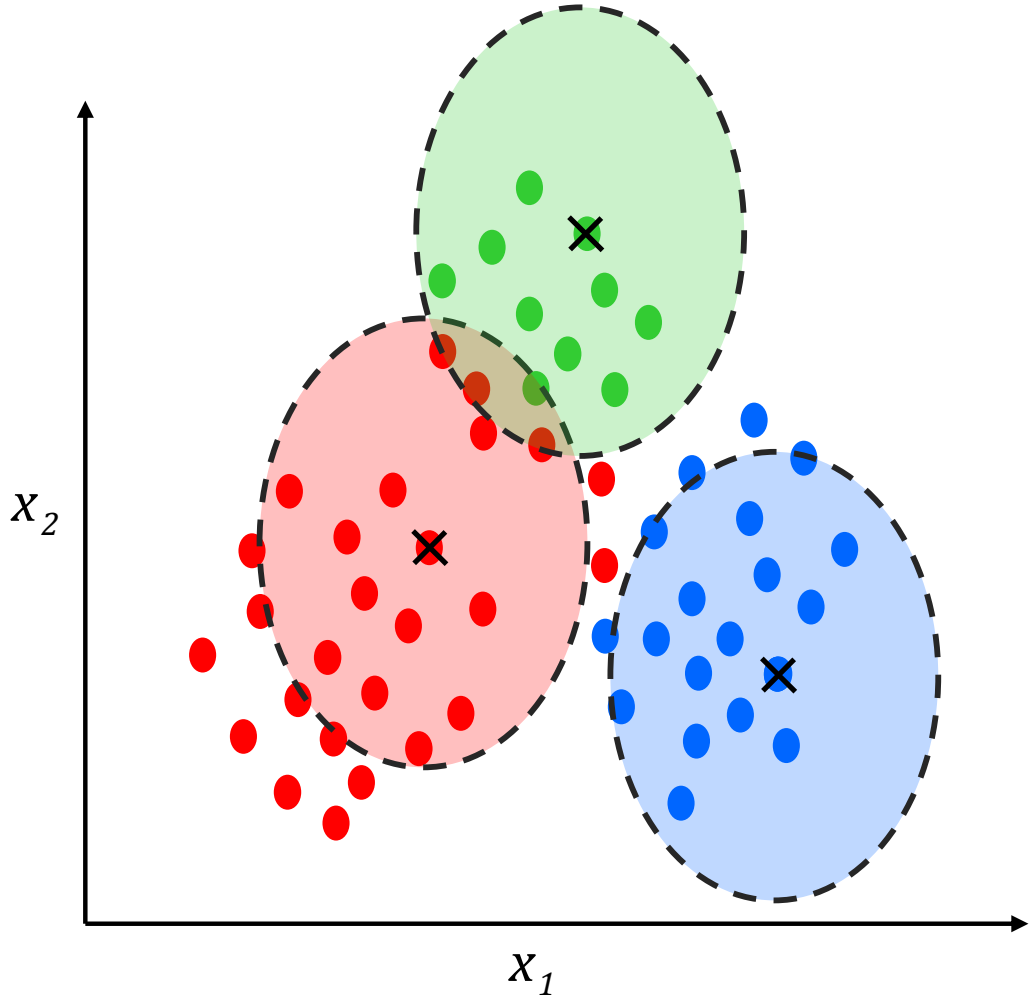


Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

Ricalcoliamo i centroidi dei cluster

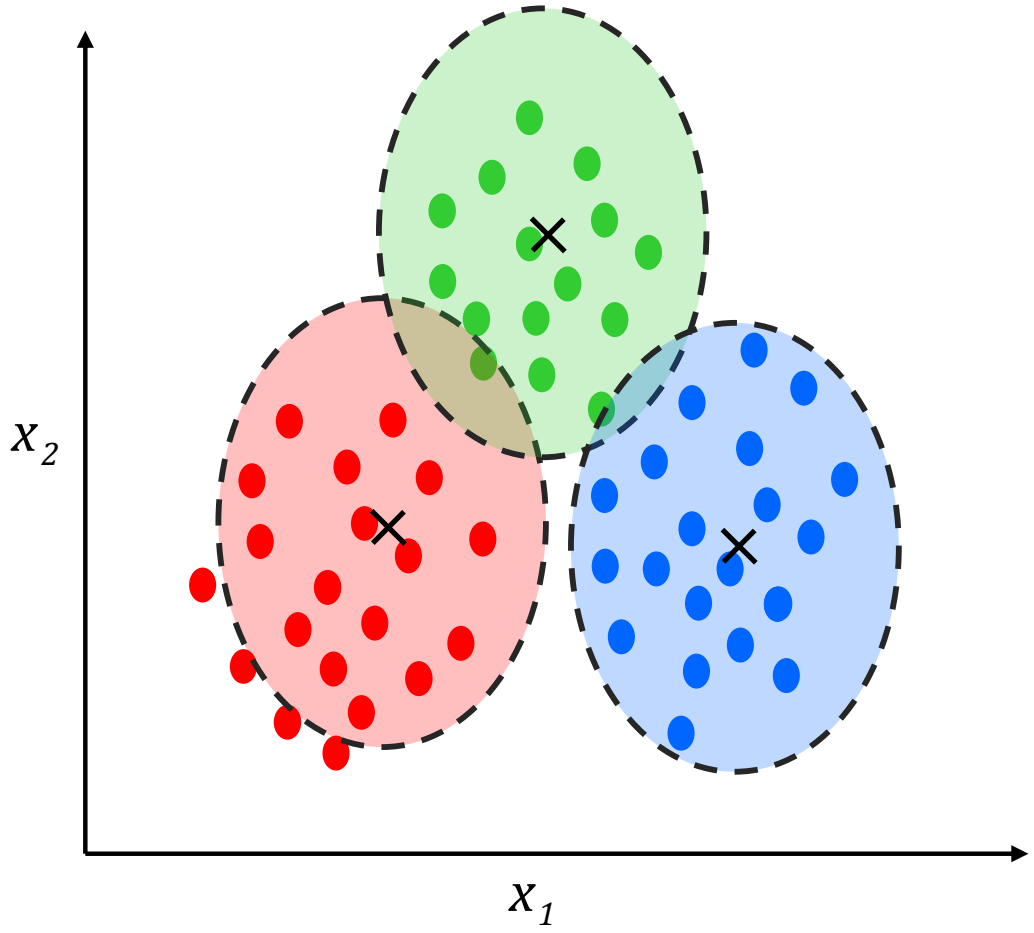
$$z_{hj} = \frac{\sum_{x_i \in C_h} x_{ij}}{|C_h|}$$



Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

Ri-assegnamo le osservazioni ai cluster identificati dai nuovi centroidi

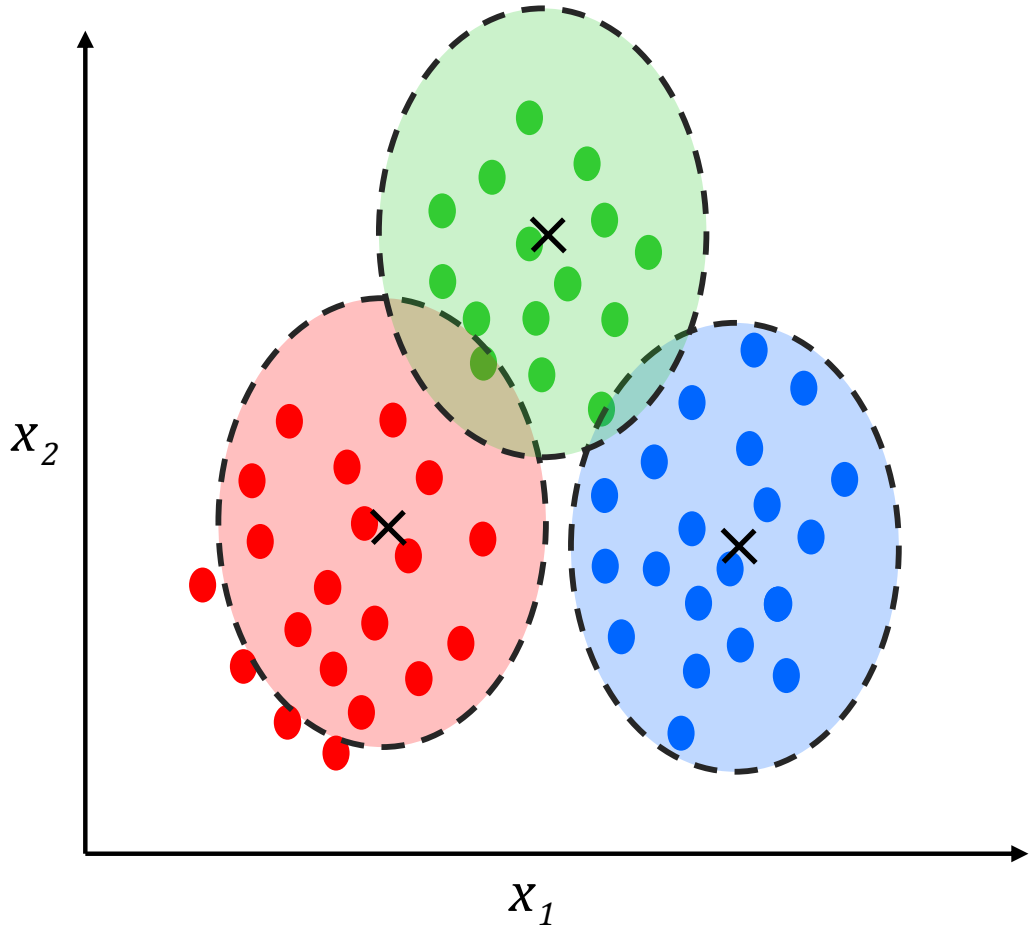


Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

Calcoliamo i nuovi centroidi

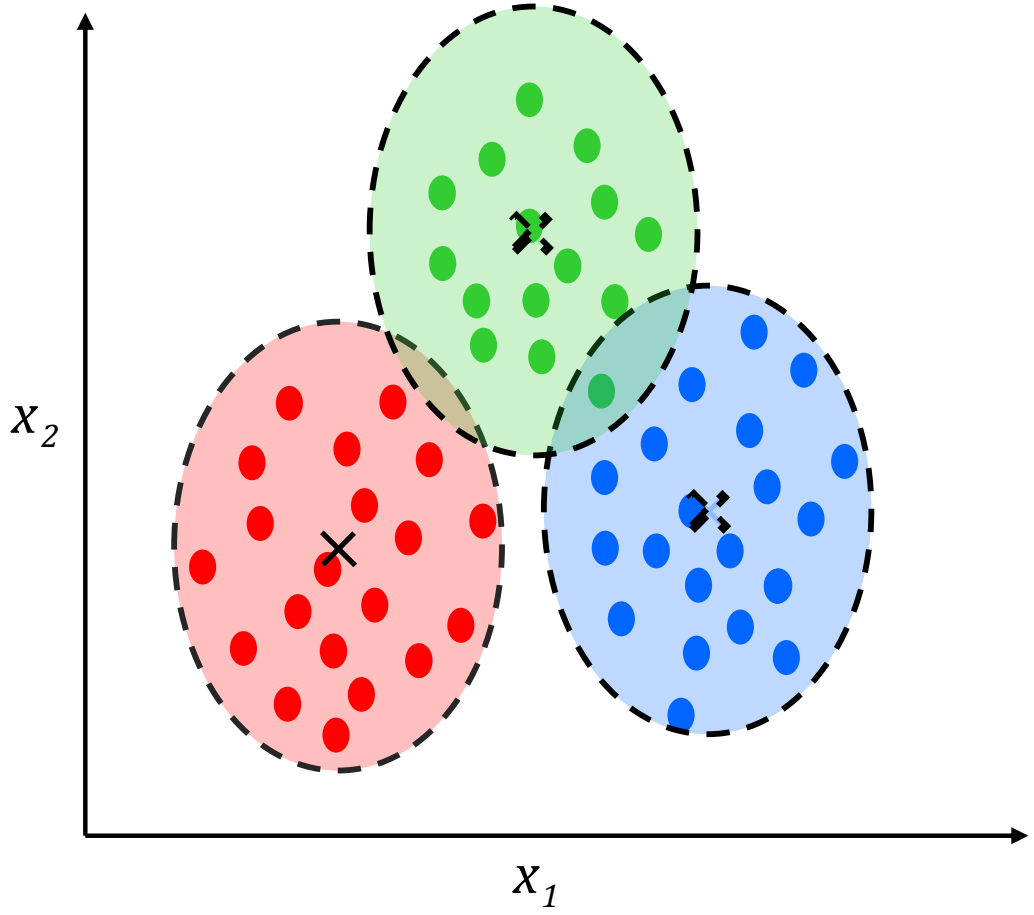
$$z_{hj} = \frac{\sum_{x_i \in C_h} x_{ij}}{|C_h|}$$



Metodi di Partizione: K-medie

Clustering del dataset costituito da punti nello spazio bi-dimensionale, scegliamo il valore "*K*" pari a tre (*K=3*) e la distanza Euclidea come misura di affinità.

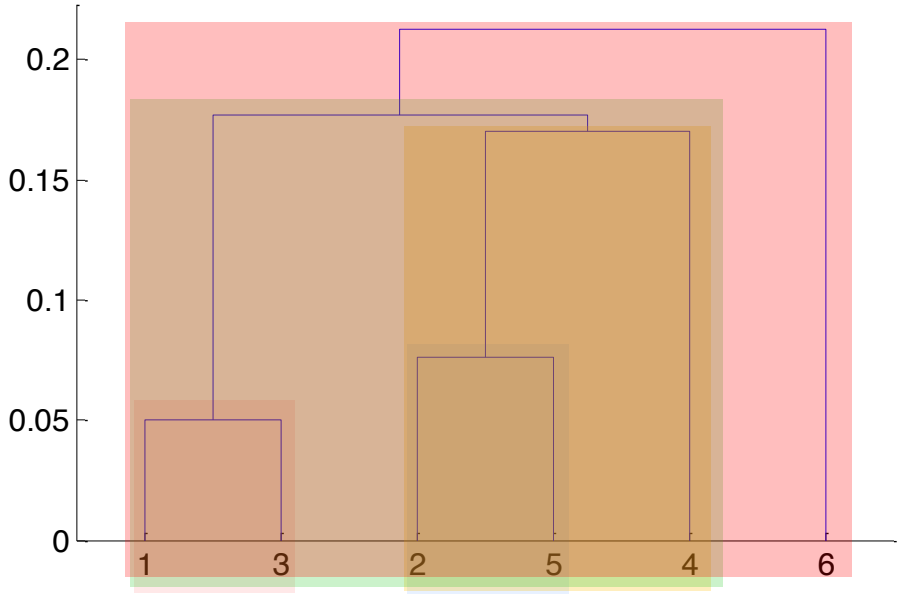
Non esistono osservazioni da ri-assegnare, l'algoritmo termina

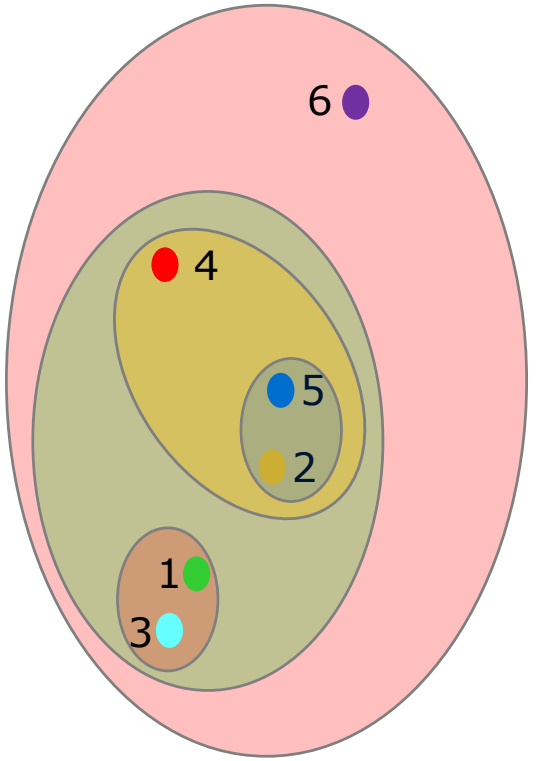


CLUSTERING

METODI GERARCHICI

Metodi Gerarchici: agglomerativi





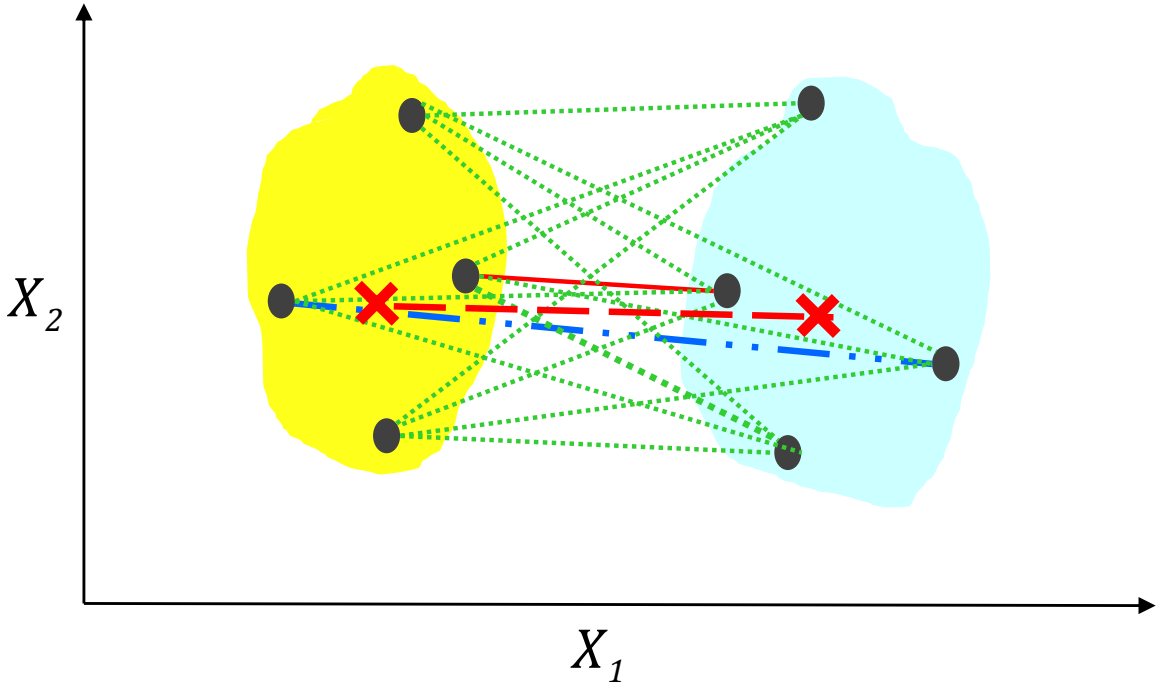
Scelta della misura di distanza

MIN ——

MAX — · · —

MEDIA ·····

Centroid — —



CLUSTERING

METODI GRAPH-BASED

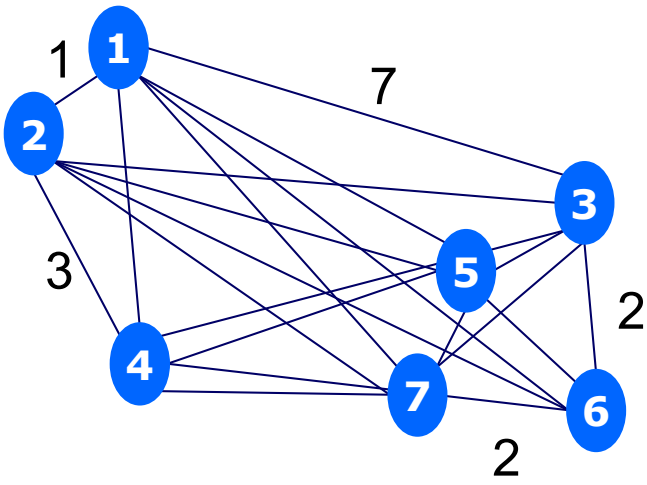
Metodi Graph-based: sparsificazione

Si utilizza la matrice delle distanze **Dist**, si ricava un *grafo denso* (*completo*) in cui

- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (*peso connessione*)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0



Metodi Graph-based: sparsificazione

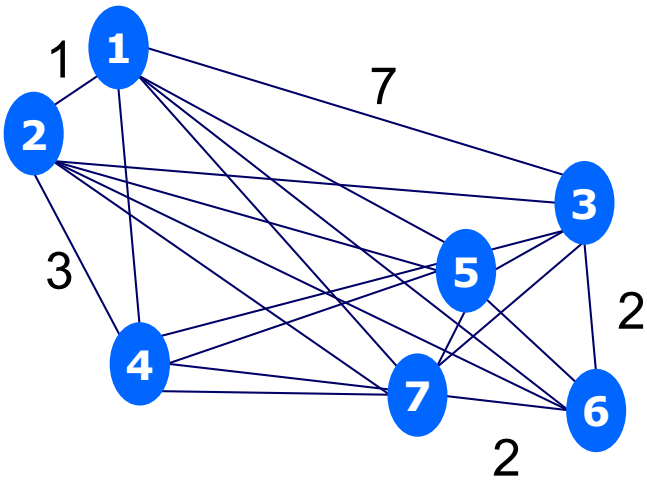
Si utilizza la matrice delle distanze **Dist**, si ricava un grafo denso (completo) in cui

- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (peso connessione)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0

soglia = 5



Metodi Graph-based: sparsificazione

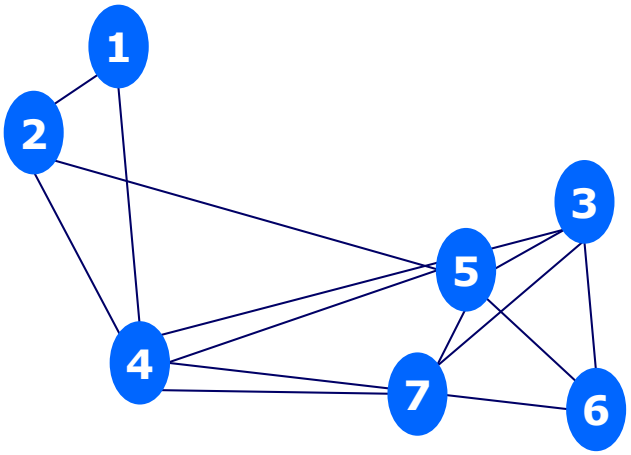
Si utilizza la matrice delle distanze **Dist**, si ricava un grafo denso (completo) in cui

- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (peso connessione)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0

soglia = 5



Metodi Graph-based: sparsificazione

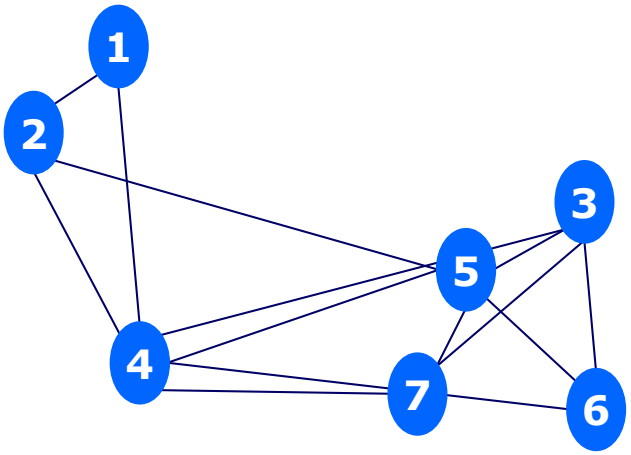
Si utilizza la matrice delle distanze **Dist**, si ricava un grafo denso (completo) in cui

- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (peso connessione)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0

soglia = 3



Metodi Graph-based: sparsificazione

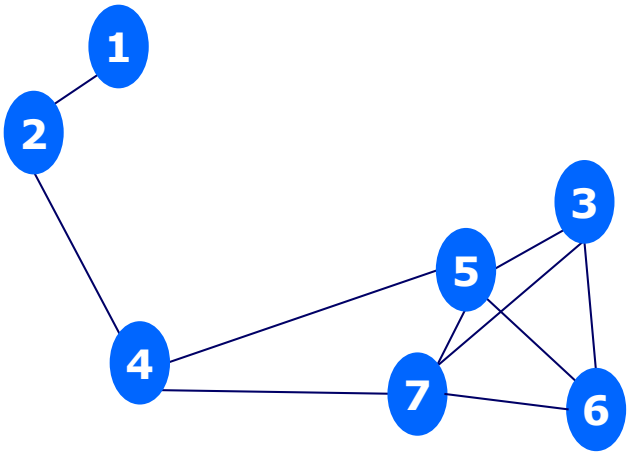
Si utilizza la matrice delle distanze **Dist**, si ricava un grafo denso (completo) in cui

- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (peso connessione)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0

soglia = 3



Metodi Graph-based: sparsificazione

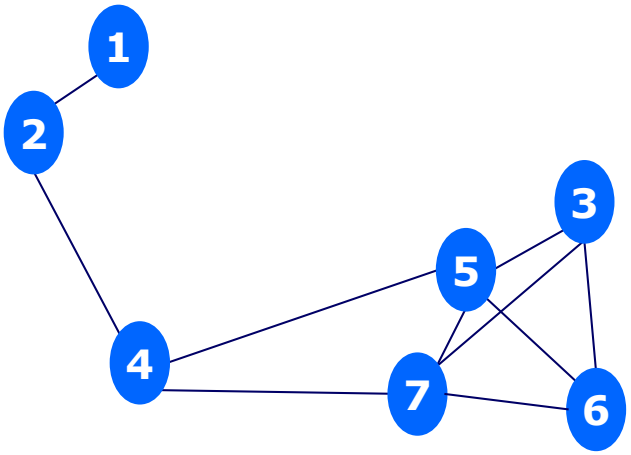
Si utilizza la matrice delle distanze **Dist**, si ricava un grafo denso (completo) in cui

- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (peso connessione)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0

soglia = 2



Metodi Graph-based: sparsificazione

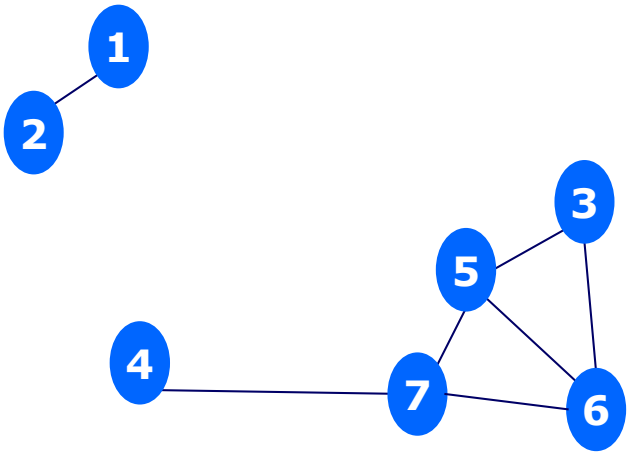
Si utilizza la matrice delle distanze **Dist**, si ricava un grafo denso (completo) in cui

- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (peso connessione)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0

soglia = 2



Metodi Graph-based: sparsificazione

Si utilizza la matrice delle distanze **Dist**, si ricava un grafo denso (completo) in cui

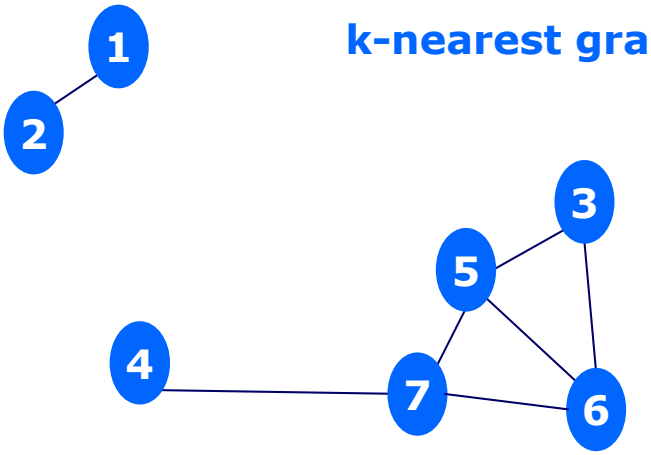
- **nodo** = osservazione
- **arco** = misura la distanza (**Dist**) tra la coppia di nodi che collega (peso connessione)

Di norma ogni osservazione sarà simile ad un numero limitato di altre osservazioni, si procede alla sparsificazione del grafo, si fissa il valore di una soglia e si modifica il valore del peso della connessione per quegli archi che superano il valore di soglia e lo si pone pari a zero.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	1	7	5	6	8	7
2		0	7	3	4	9	8
3			0	5	1	2	3
4				0	3	4	2
5					0	2	1
6						0	2
7							0

soglia = 2

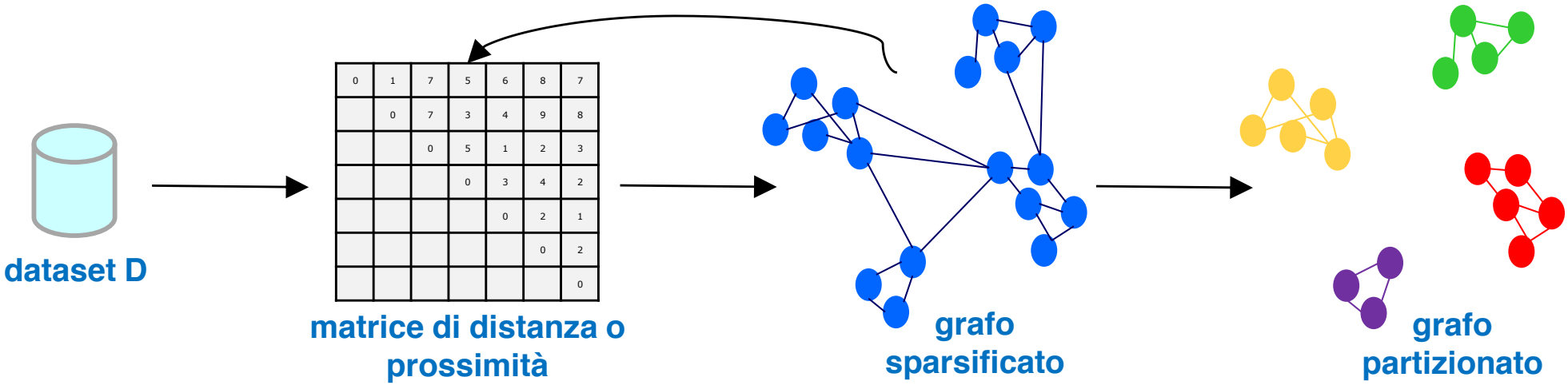
k-nearest graph



Metodi Graph-based: sparsificazione

Vantaggi

- *Riduzione numero di elementi non nulli della matrice delle distanze* **Dist**
- *Potenziale miglioramento del processo di clustering (nearest neighbor composto da elementi che tendono ad appartenere alla medesima classe). Ridotto l'impatto del rumore e degli outlier.*
- *Possibilità di utilizzare algoritmi di partizionamento dei grafi. Disponibilità di un corpo di contributi significativo (individuazione del partizionamento min-cut di grafi sparsi, computazione parallela, MapReduce)*



CLUSTERING

METODI PROTOTYPE-BASED

Metodi Prototype-based

Un cluster è visto come collezione di osservazioni, caratterizzato dal fatto che ogni osservazione è simile al prototipo che definisce il cluster al quale l'osservazione viene assegnata.

L'algoritmo delle *K-medie* è un esempio chiaro di metodo *prototype-based*, i centroidi delle osservazioni appartenenti ad ogni cluster vengono impiegati come prototipi dei cluster medesimi.

Presentiamo di seguito approcci di clustering che espandono tale concetto in una o più direzioni:

- *le osservazioni possono appartenere a più di un cluster*
- *le osservazioni sono campioni casuali provenienti da una distribuzione di probabilità*
- *i cluster sono vincolati ad avere relazioni fissate (SOMs)*

Metodi Prototype-based: SOMs

Kohonen (1982, 1984)

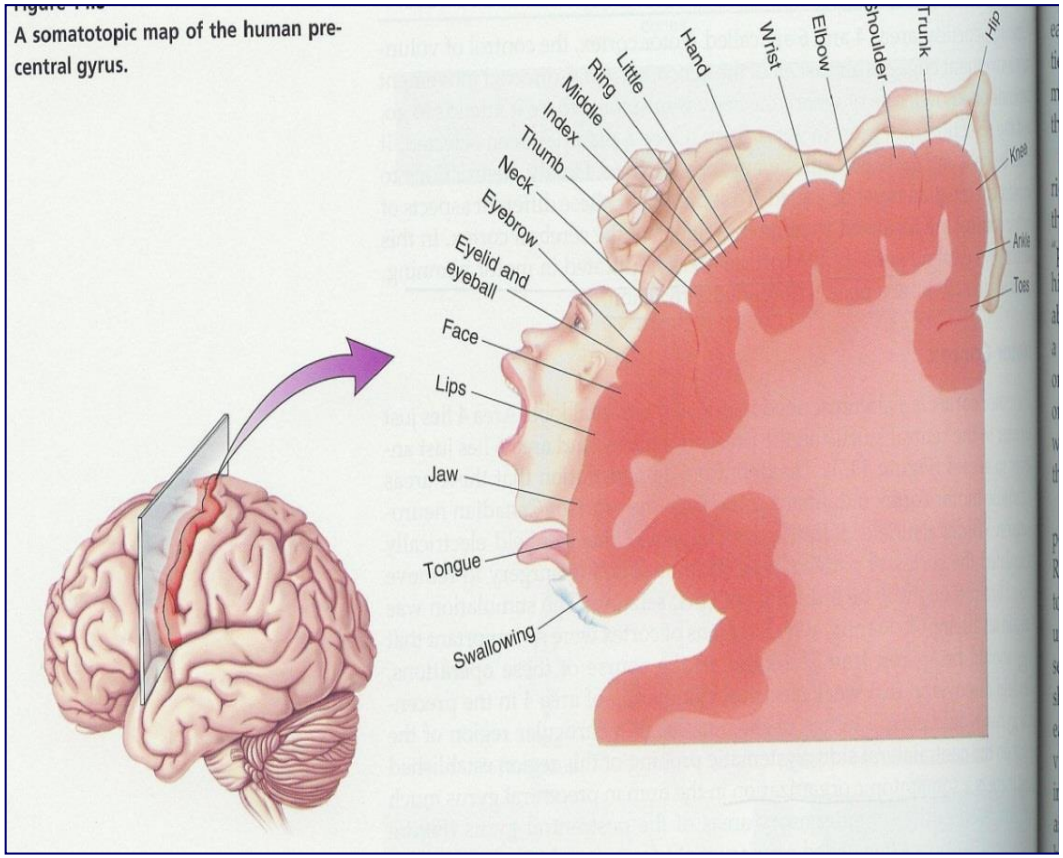
Self Organizing Maps (SOMs)

Osservazioni:

- **Nei sistemi biologici le cellule attivate da orientamenti simili tendono a posizionarsi in aree localizzate,**
- **Studi sui gatti tramite micro elettrodi,**
- **L'orientamento porta alla formazione di una mappa dove attivazioni simili sono vicine**

  -Topographic feature map

  -To train a network using competitive learning to create feature maps automatically



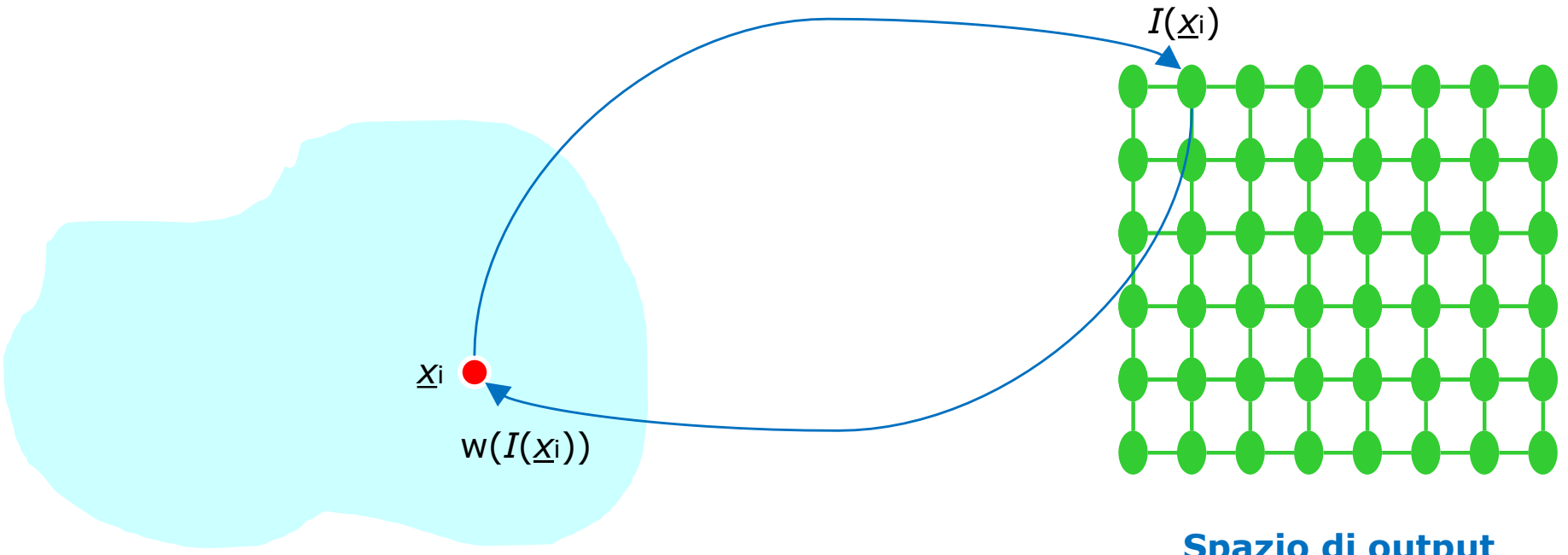
Self-Organizing Map (SOM)

ANN con un-supervised learning

- **Mappare uno spazio degli attributi molto grande in uno spazio di output bidimensionale (representation space)**

Come viene realizzato il mapping?

Ogni punto $\underline{x}_i$ dello spazio di input viene mappato su un elemento $I(\underline{x}_i)$ dello spazio di output (*representation space*). Ogni elemento dello spazio di output $I(\underline{x}_i)$ viene mappato sul corrispondente punto $w(I(\underline{x}_i))$ nello spazio di input.



Spazio di input

(continuo e a dimensionalità elevata)

Spazio di output

(discreto e a bassa dimensionalità)

Reti di Kohonen

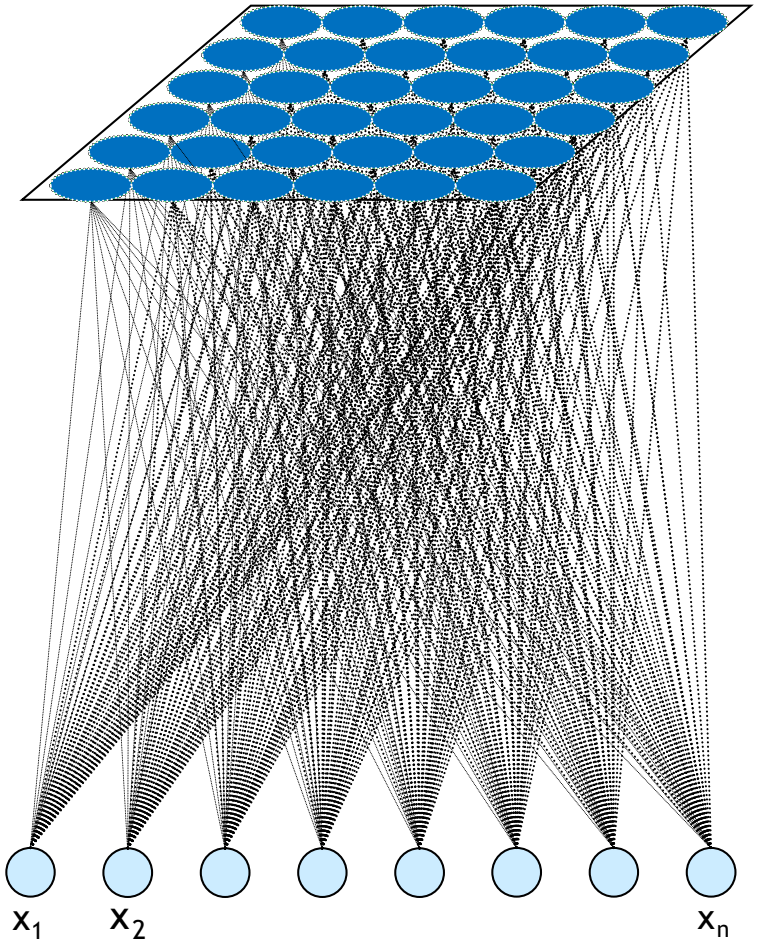
Struttura feedforward con un singolo *strato computazionale* organizzato su una griglia discreta, ogni neurone è completamente connesso con i nodi dello *strato di input*.

È possibile trattare anche modelli mono-dimensionali, strato computazionale (**spazio di output**) del tipo riportato sotto.

spazio di output

(spazio di output)

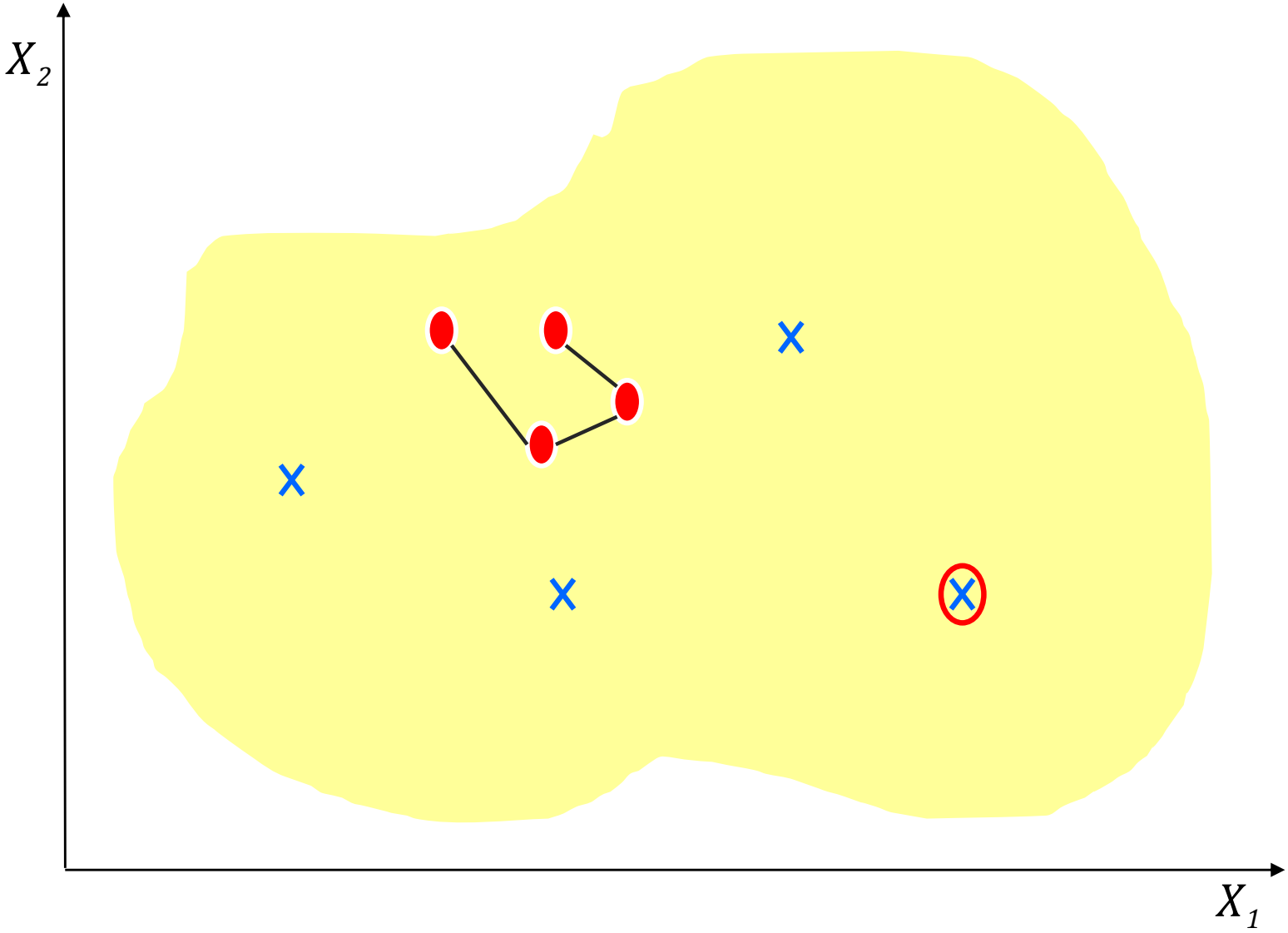
strato computazionale



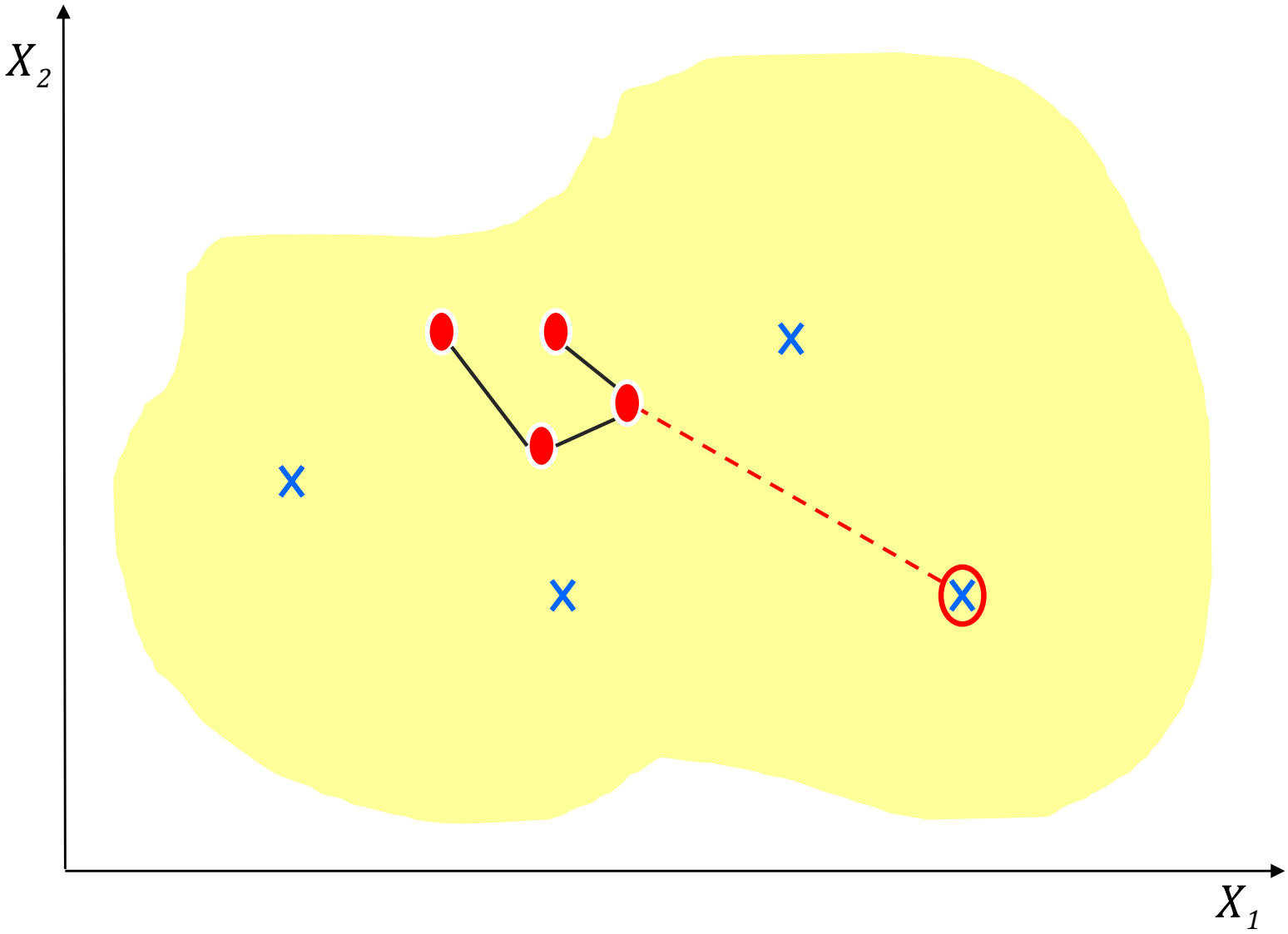
strato di input

(spazio di input)

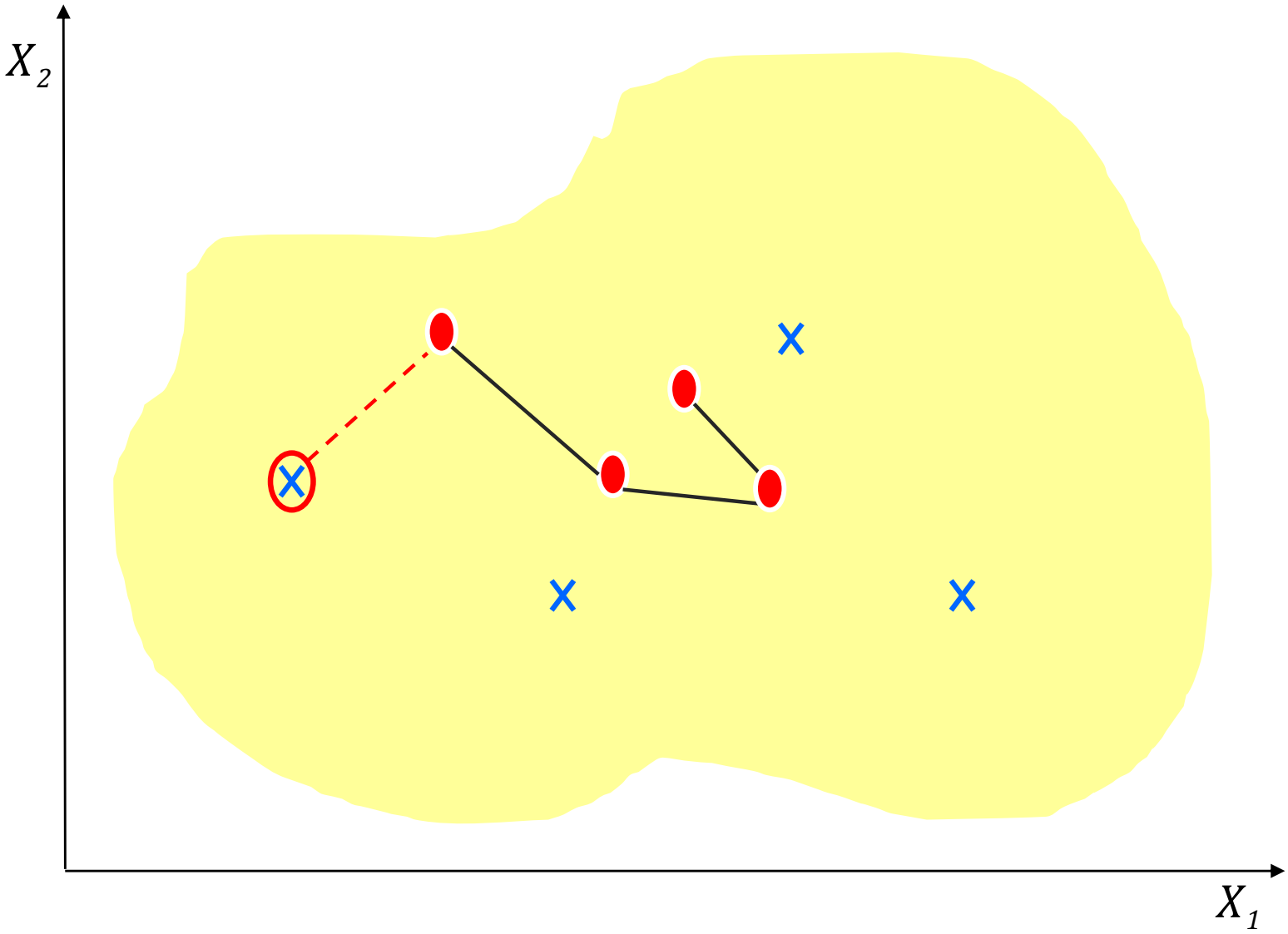
Metodi Prototype-based: SOMs



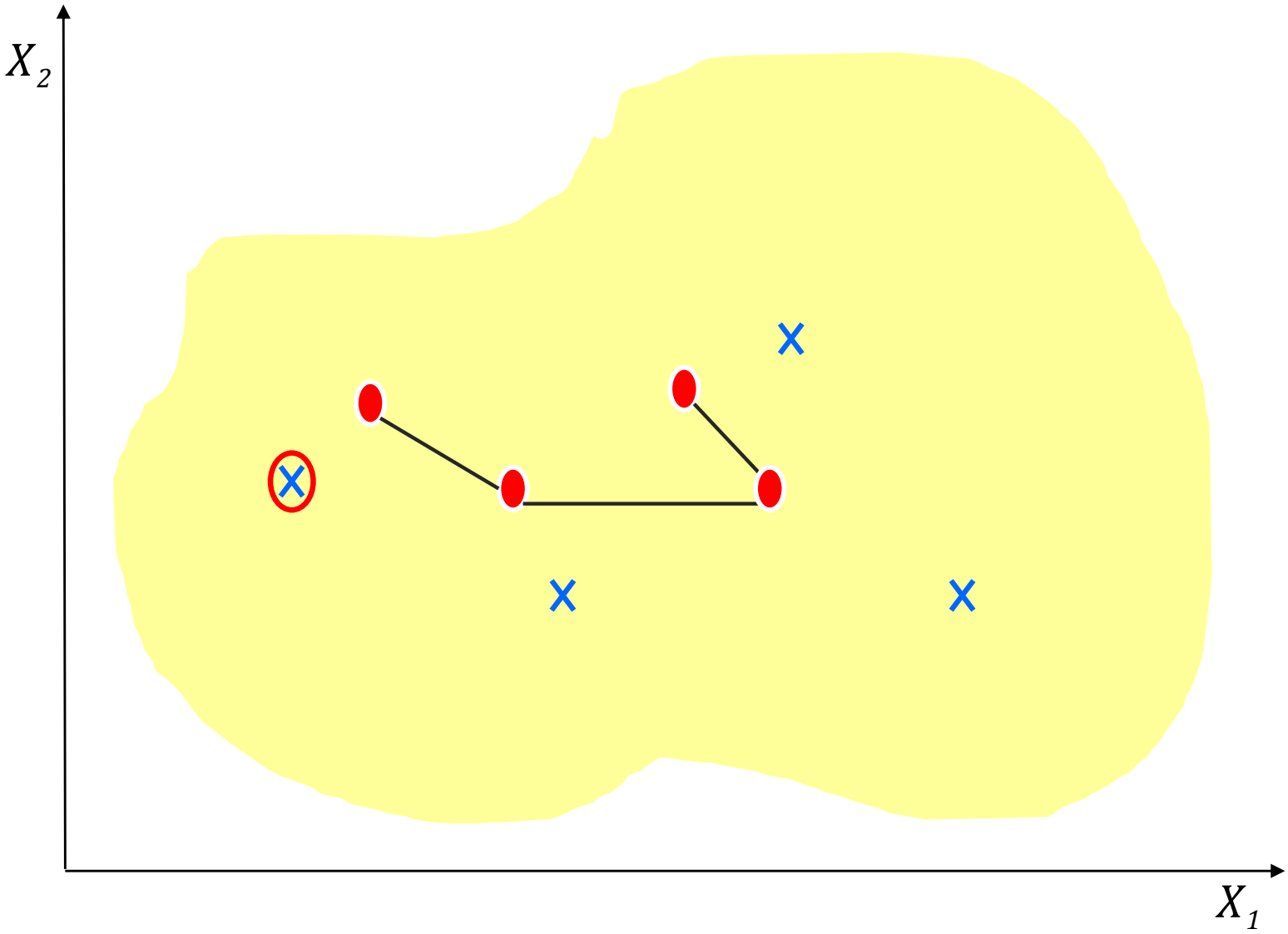
Metodi Prototype-based: SOMs



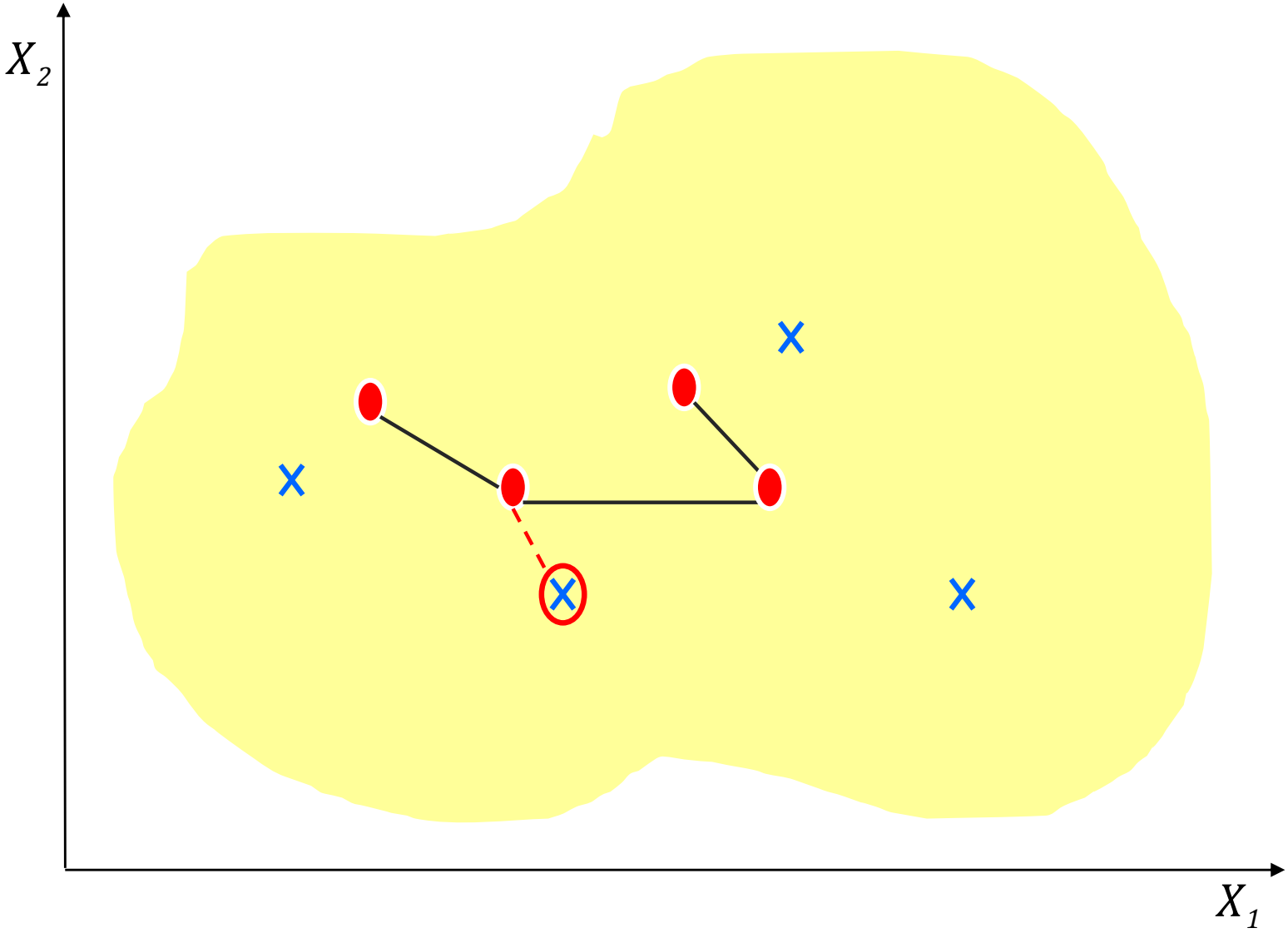
Metodi Prototype-based: SOMs



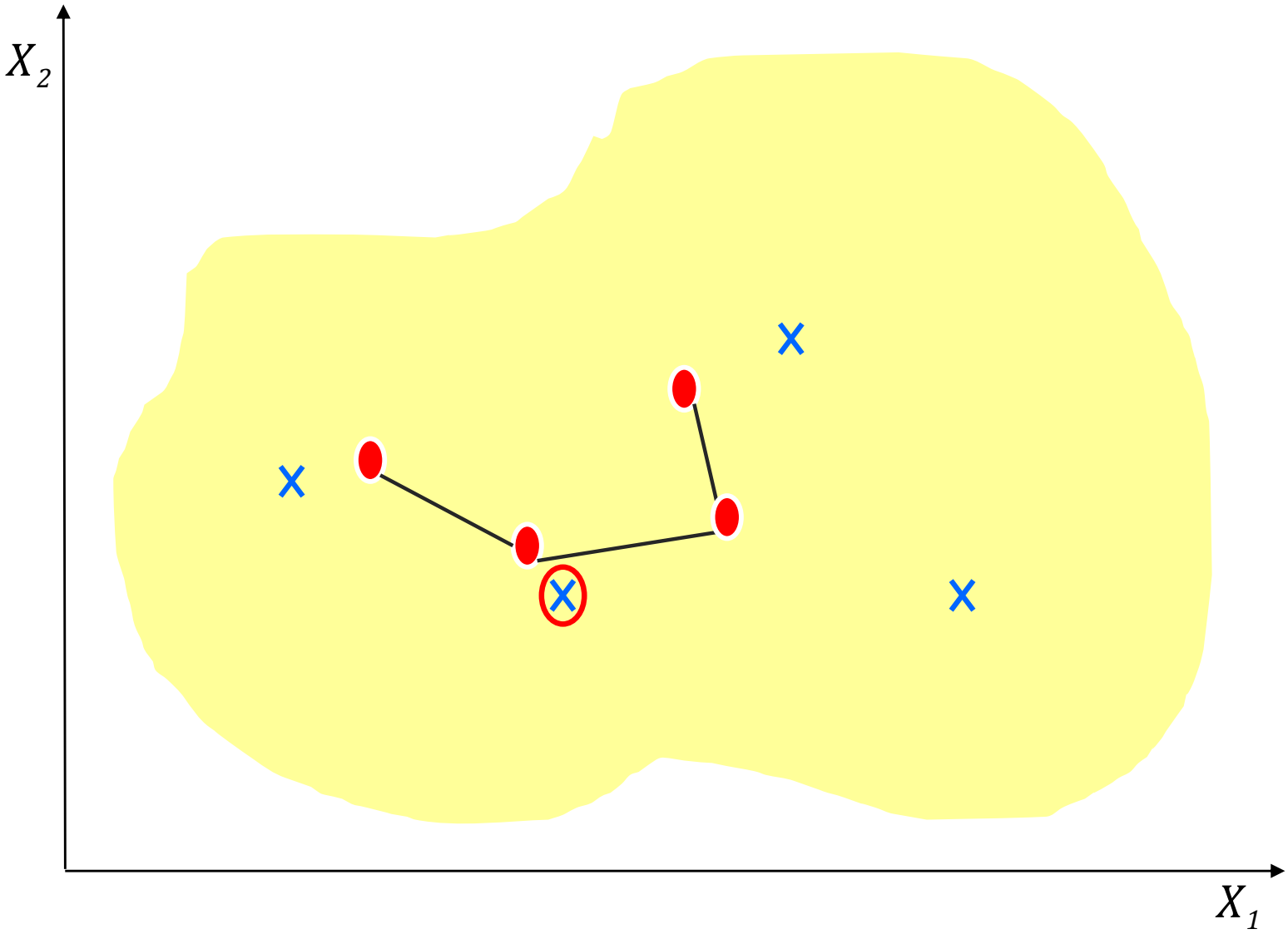
Metodi Prototype-based: SOMs



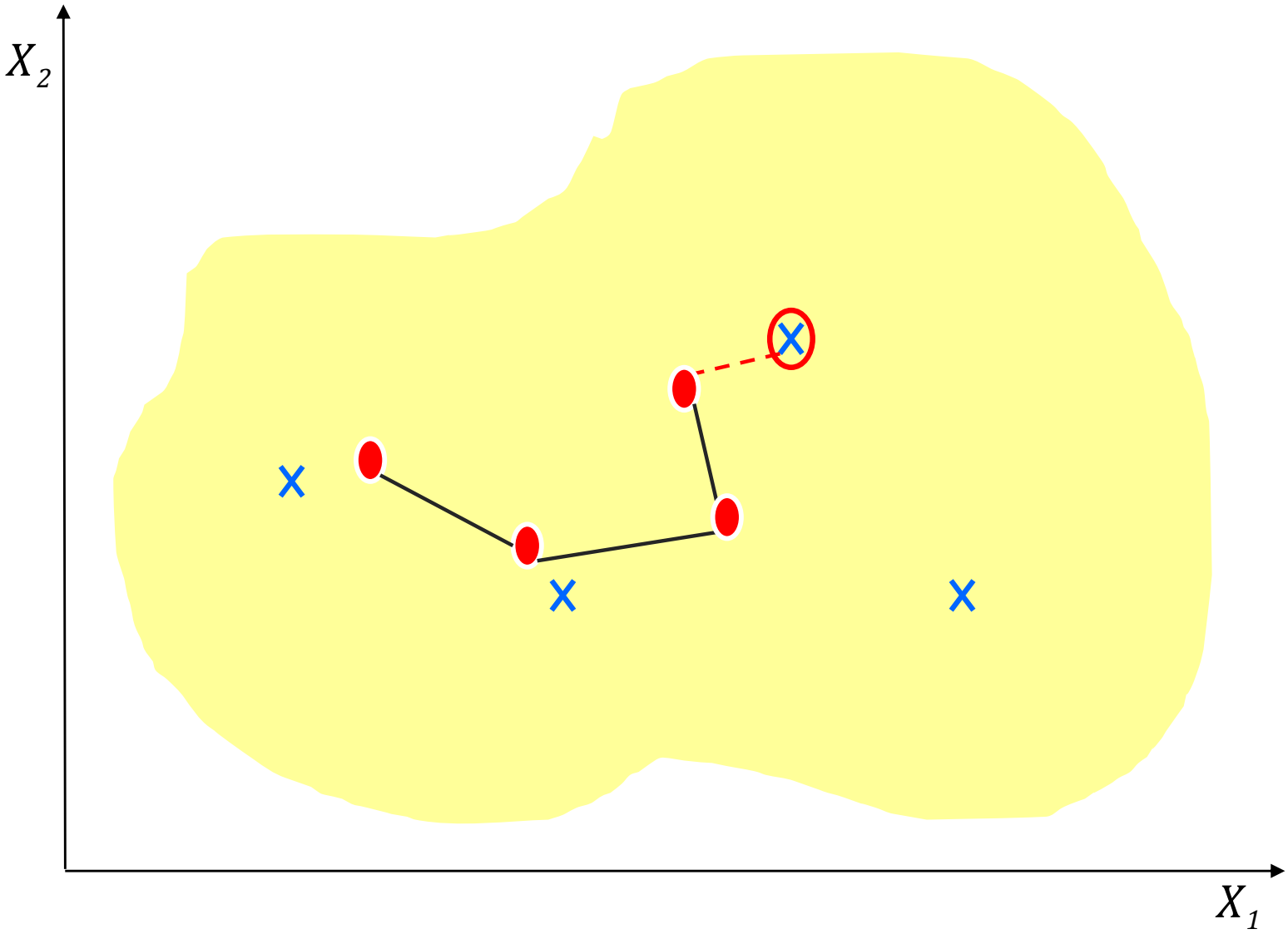
Metodi Prototype-based: SOMs



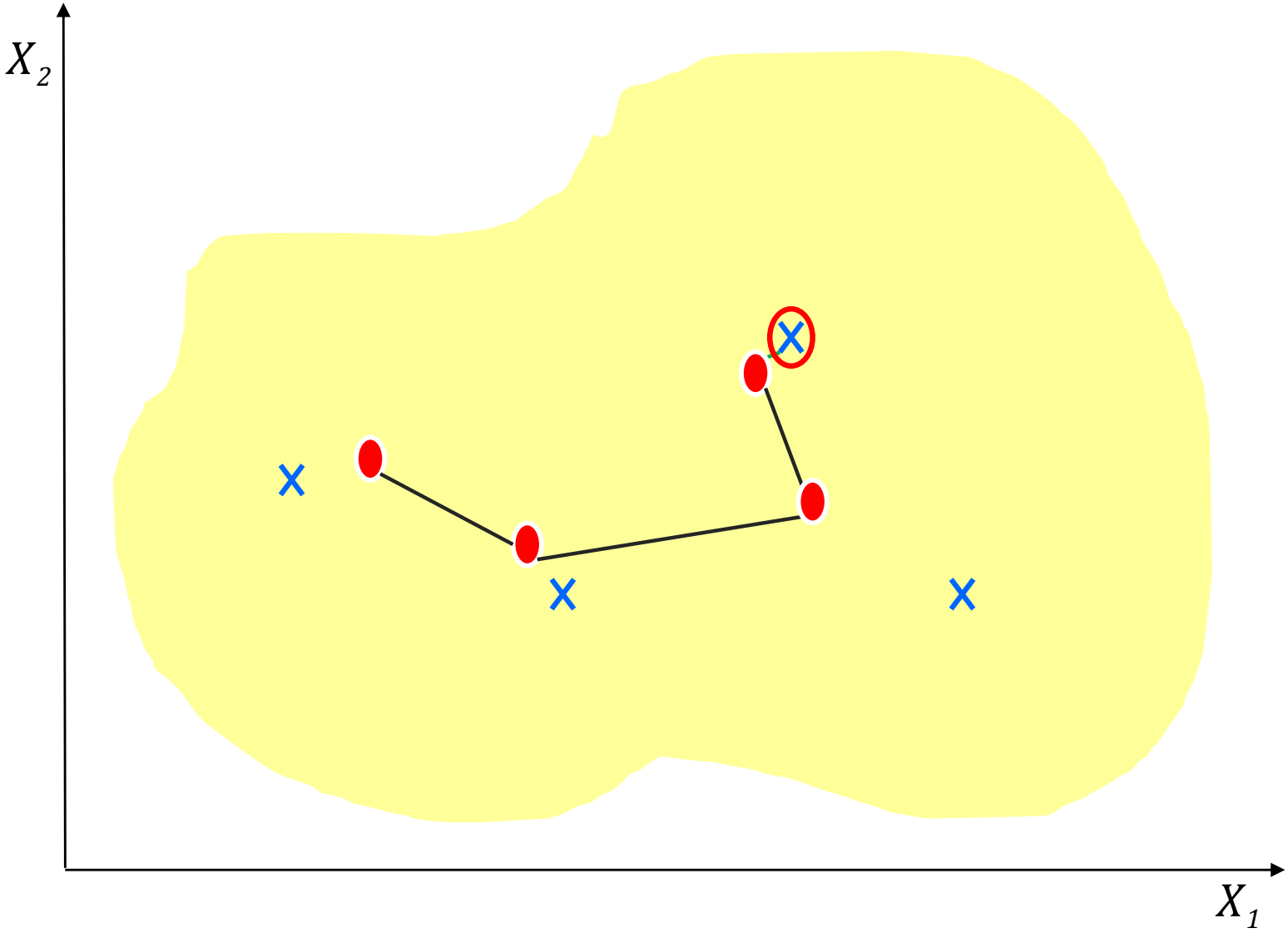
Metodi Prototype-based: SOMs



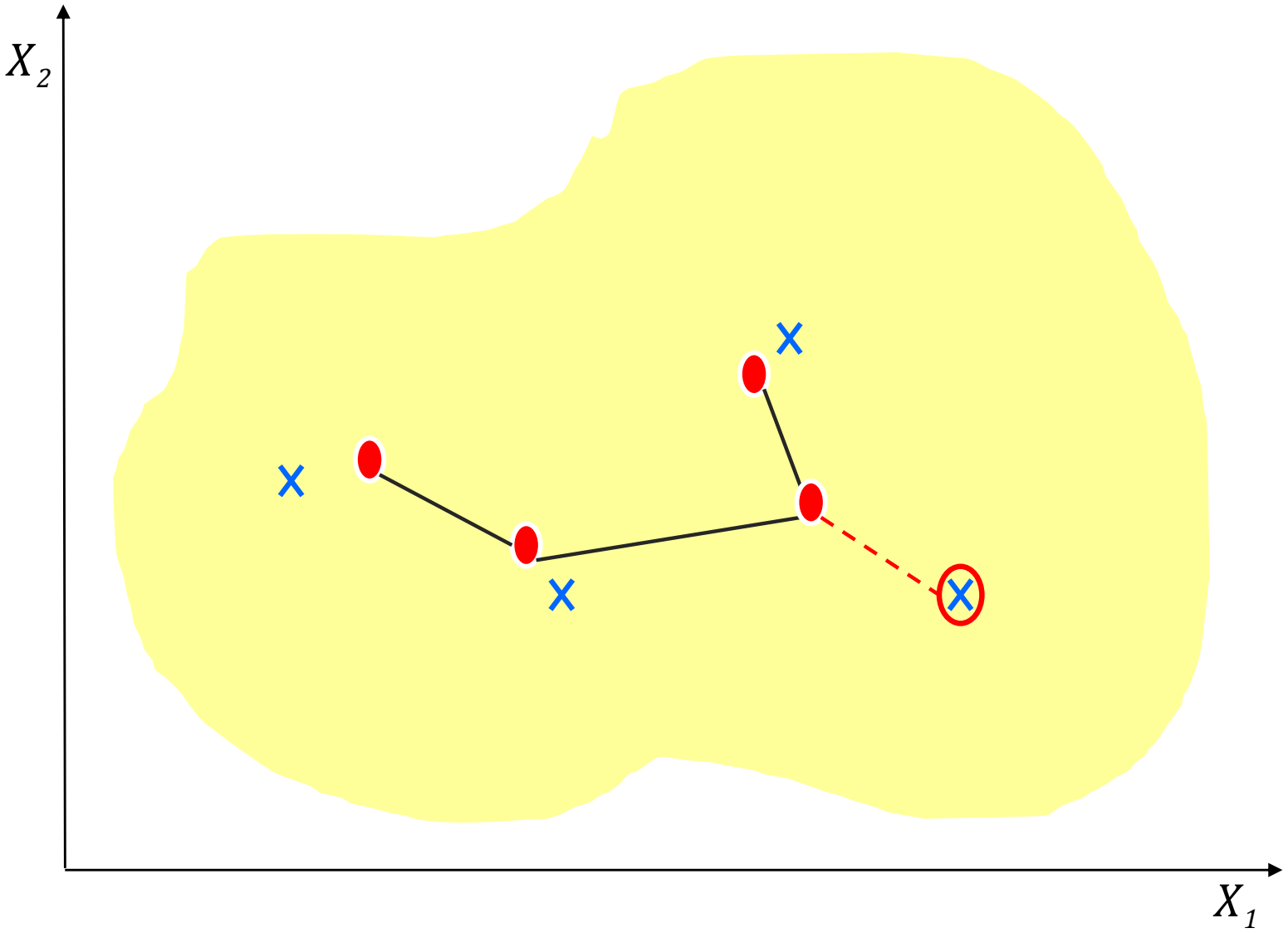
Metodi Prototype-based: SOMs



Metodi Prototype-based: SOMs



Metodi Prototype-based: SOMs



Metodi Prototype-based: SOMs