Fondamenti di Machine Learning

Il Clustering

Clustering e Classificazione

presentato da
Giuseppe Gullo

PROFESSION AI

Il Clustering

E’ un task dell’apprendimento non supervisionato,
l’obiettivo è raggruppare insieme osservazioni
in base a caratteristiche simili.

Il Clustering

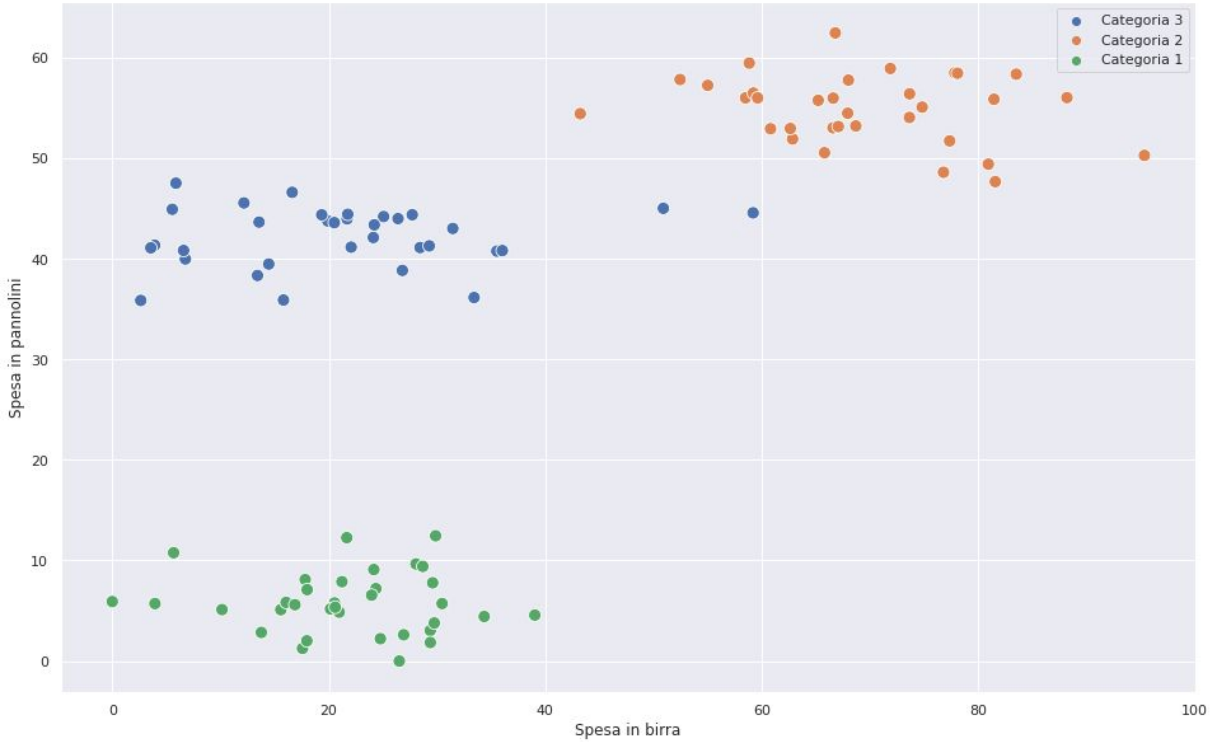
E’ un problema simile alla classificazione,
con la differenza che non abbiamo una classe o un label di esempio.

Classificare i clienti di un supermercato

Spesa mensile in birra	Spesa mensile in pannolini	Categoria
13.41	38.34	2
52.43	57.82	1
88.17	2.22	0
20.53	56.02	1
73.63	5.77	0
6.75	56.40	1
…	…	…
17.97	2.01	0

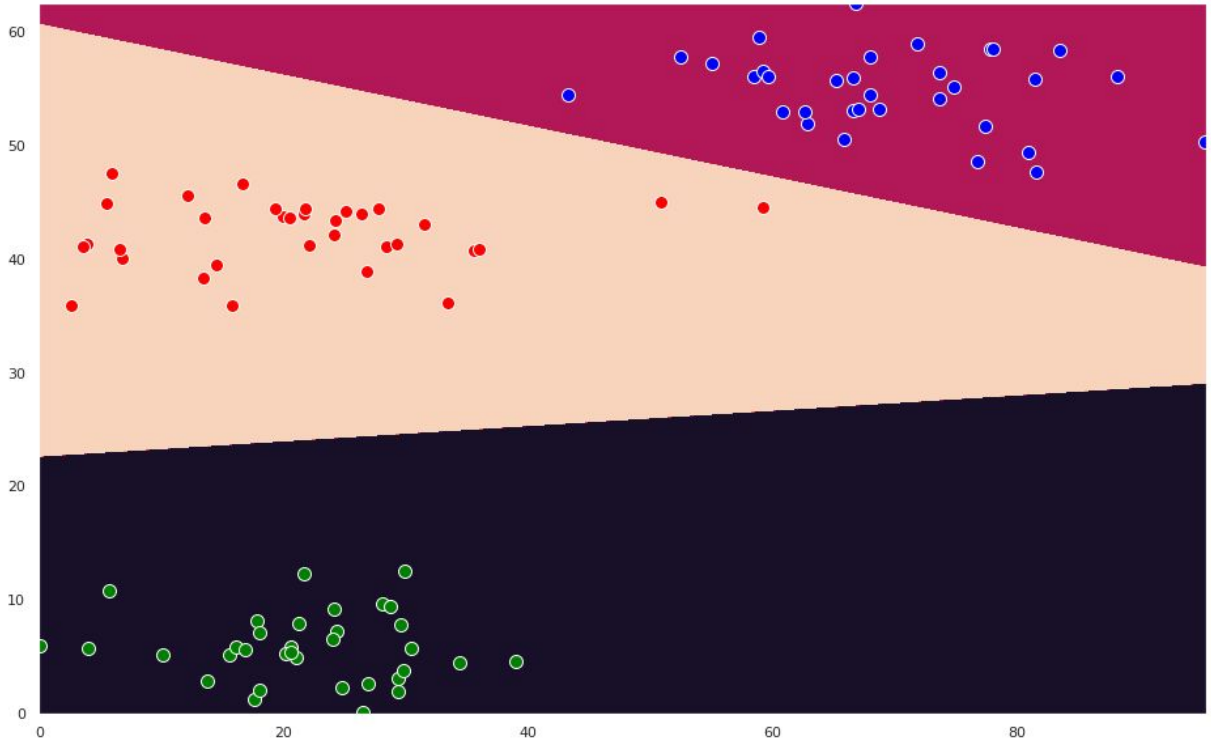
Classificare i clienti di un supermercato

Avendo la variabile target possiamo addestrare un modello di classificazione multiclasse



Classificare i clienti di un supermercato

Avendo la variabile target possiamo addestrare un modello di classificazione multiclasse

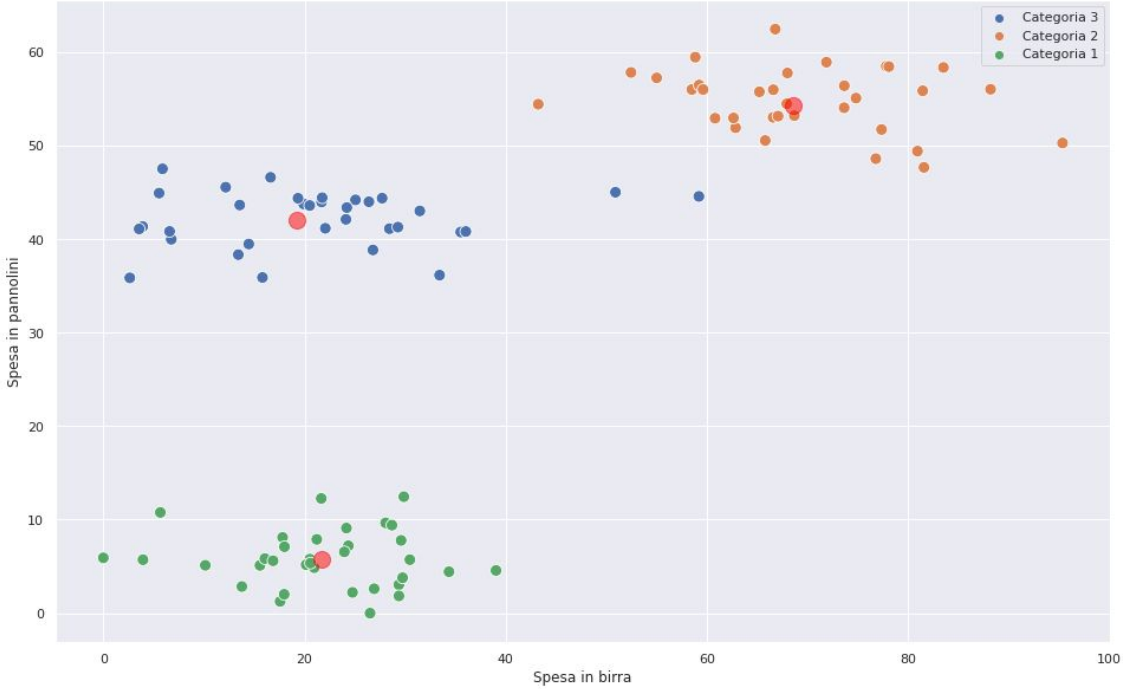


Clusterizzare i clienti di un supermercato

Spesa mensile in birra	Spesa mensile in pannolini
13.41	38.34
52.43	57.82
88.17	2.22
20.53	56.02
73.63	5.77
6.75	56.40
…	…
17.97	2.01

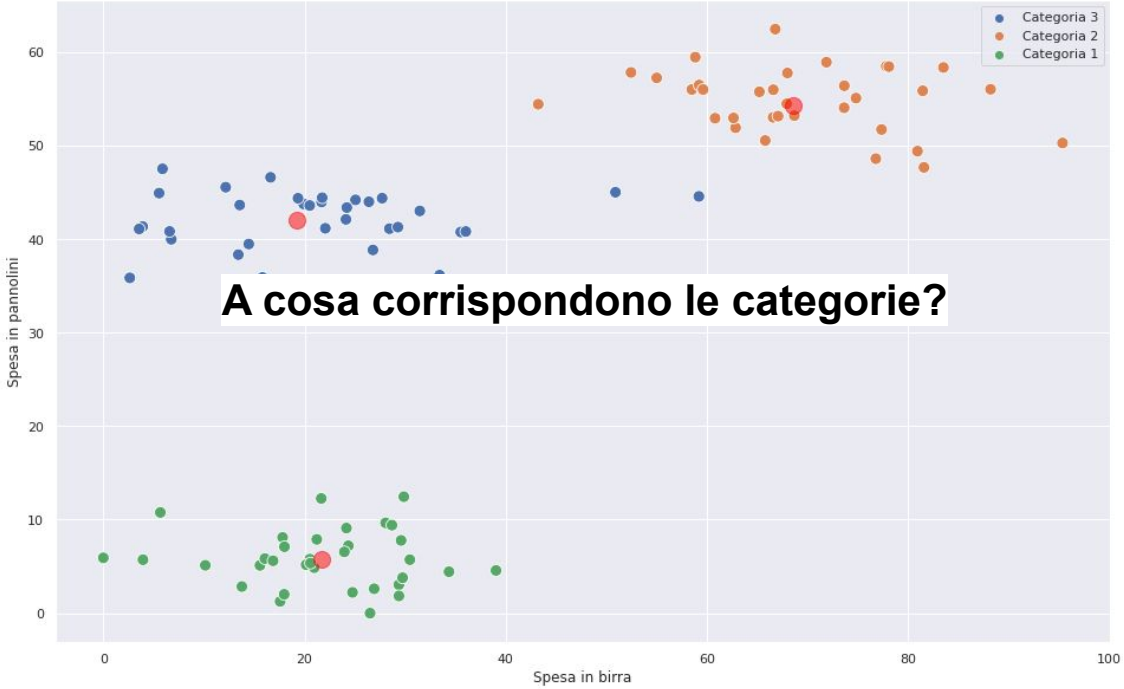
Clusterizzare i clienti di un supermercato

Un modello di clustering ci permette di raggruppare insieme le osservazioni in base alle loro caratteristiche



Clusterizzare i clienti di un supermercato

Un modello di clustering ci permette di raggruppare insieme le osservazioni in base alle loro caratteristiche



A cosa corrispondono i cluster?

A cosa corrispondono i cluster?

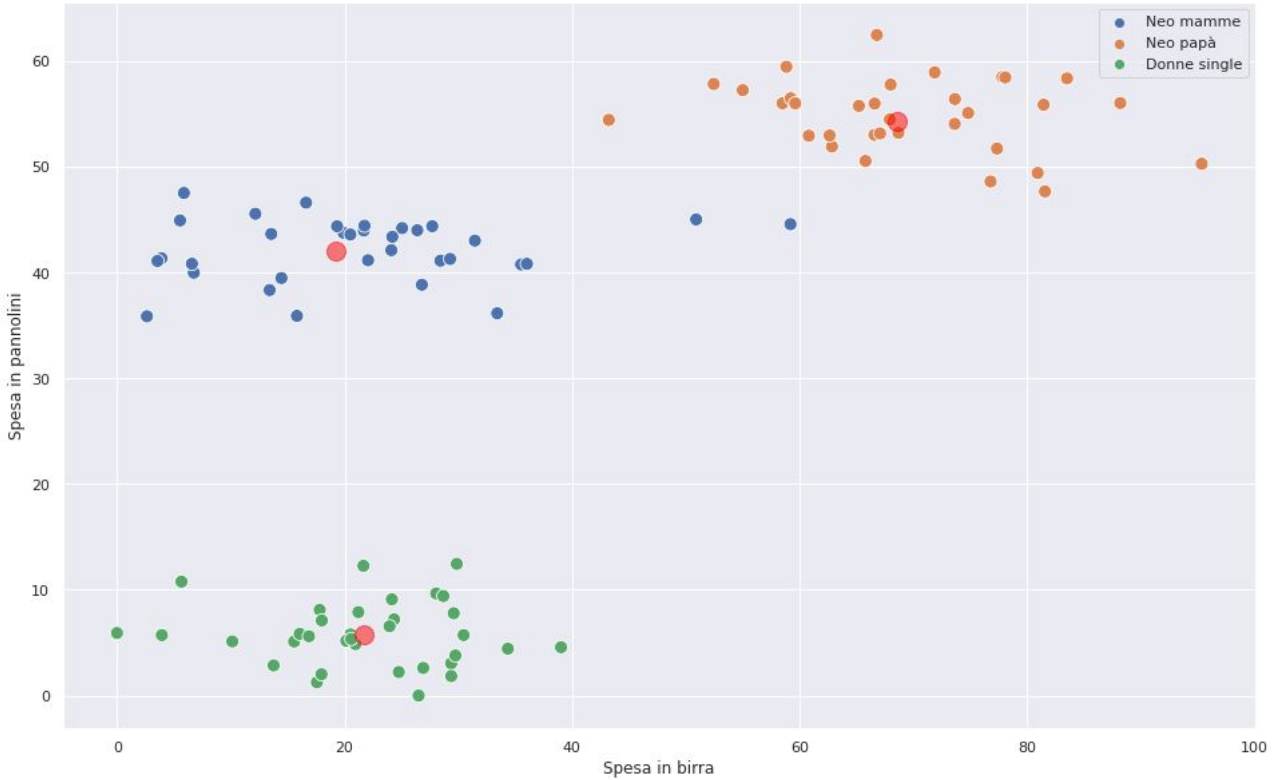
Il significato dei cluster va esplorato con analisi che vanno oltre la creazione di un modello di machine learning.

A cosa corrispondono i cluster?

Il significato dei cluster va esplorato con analisi che vanno oltre la creazione di un modello di machine learning.

Non sempre è necessario o ha senso farlo.

Clusterizzare i clienti di un supermercato



Fondamenti di Machine Learning

Il Clustering

L'algoritmo K-means

presentato da
Giuseppe Gullo

PROFESSION AI

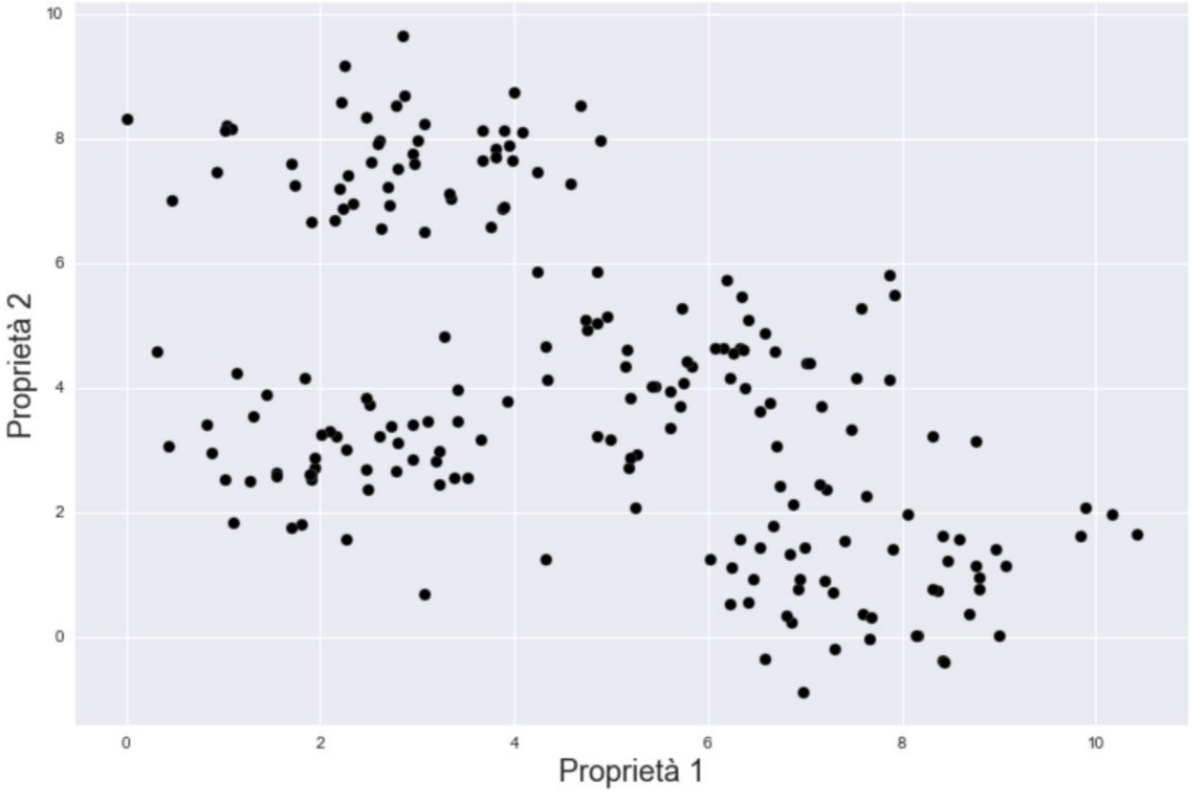
L’algoritmo K-means

E’ un algoritmo di clustering che ci permette di raggruppare le osservazioni in base alle loro caratteristiche.

K-means passo per passo

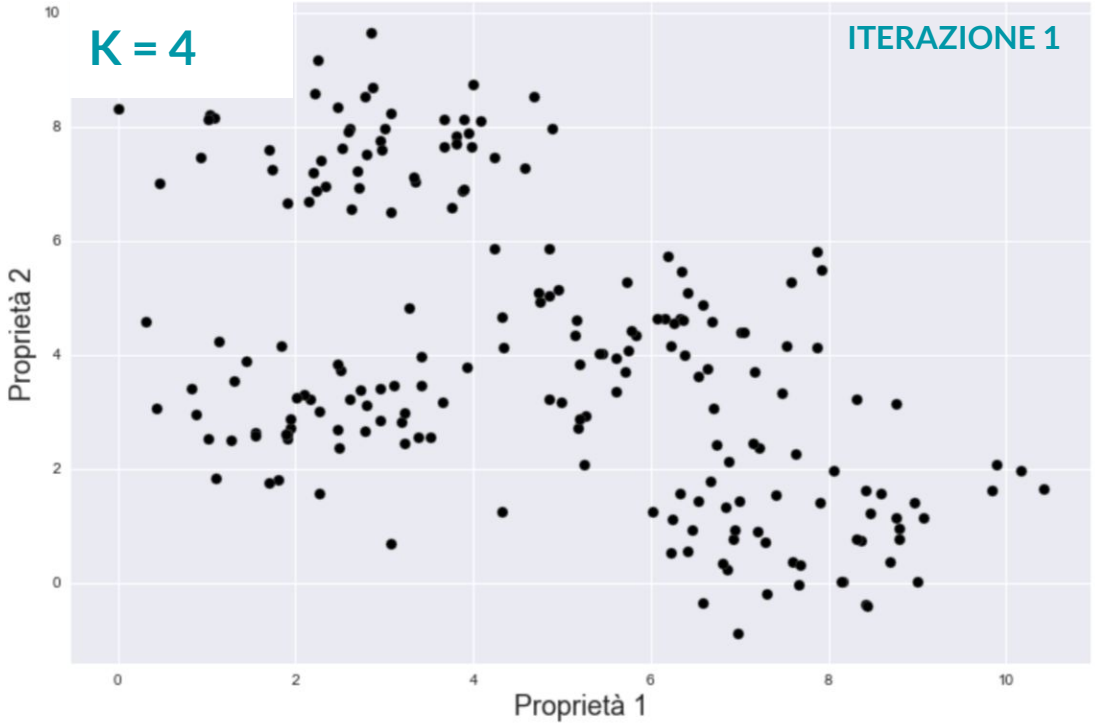
1. Scelgo il numero di clusters K da creare.
2. Seleziono casualmente K centroidi.
3. Calcolo la distanza tra ogni centroide e tutte le osservazioni.
4. Assegno le osservazioni al cluster rappresentato dal centroide più vicino.
5. Ricalcolo i centroidi come la media delle osservazioni per ogni cluster.
6. Ripeto dal punto 2 fino a quando nessuna osservazione cambia più cluster.

K-means: un esempio



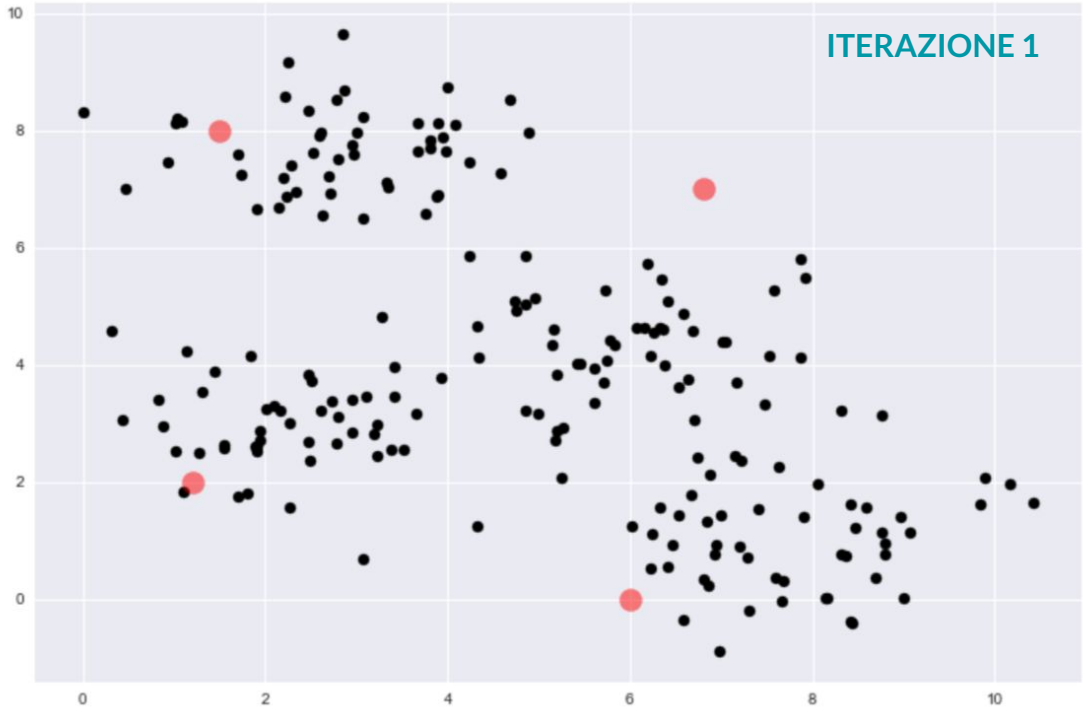
K-means: un esempio

1. Scelgo il numero di clusters K da creare.



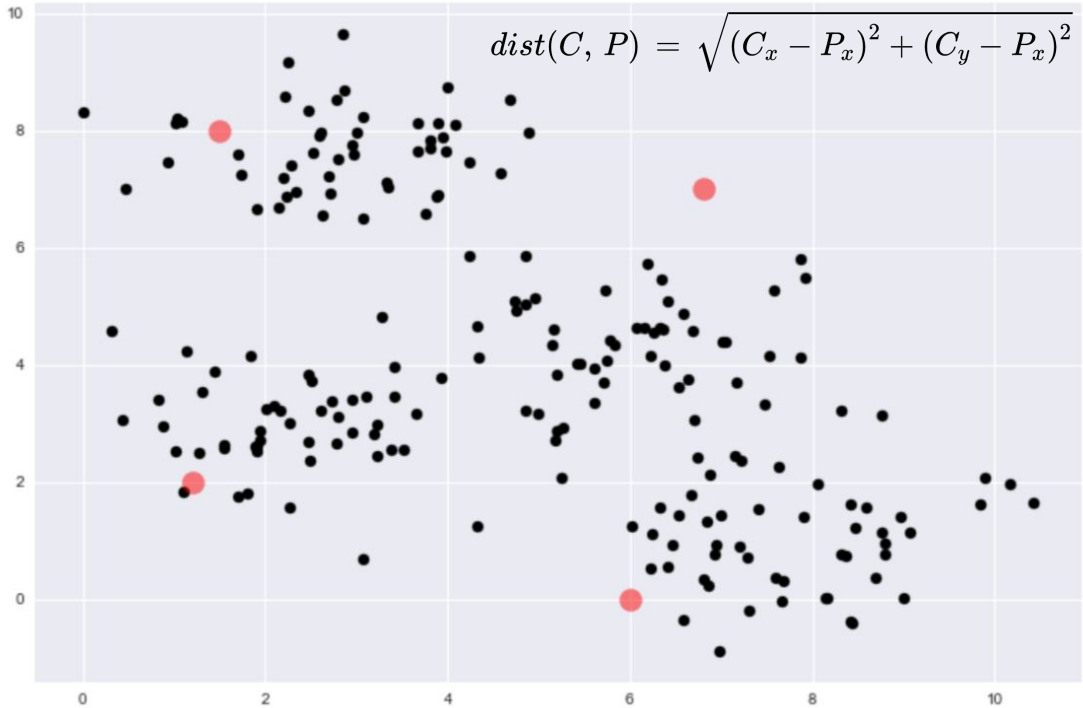
K-means: un esempio

2. Seleziono casualmente K centroidi.



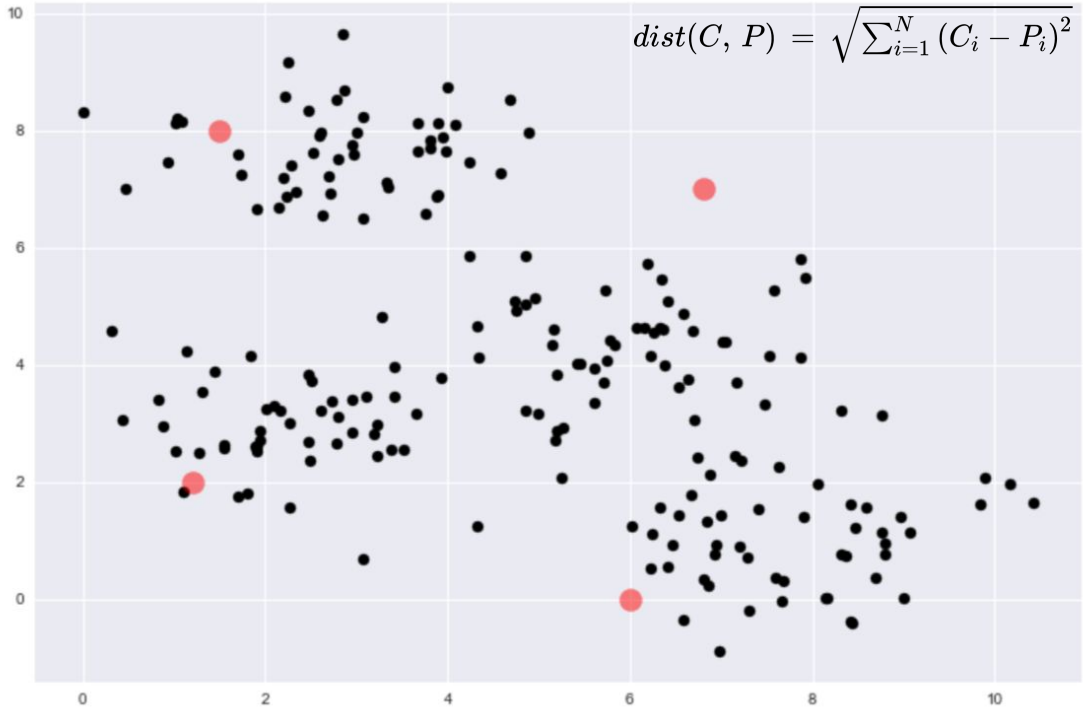
K-means: un esempio

3. Calcolo la distanza tra ogni centroide e tutte le osservazioni.



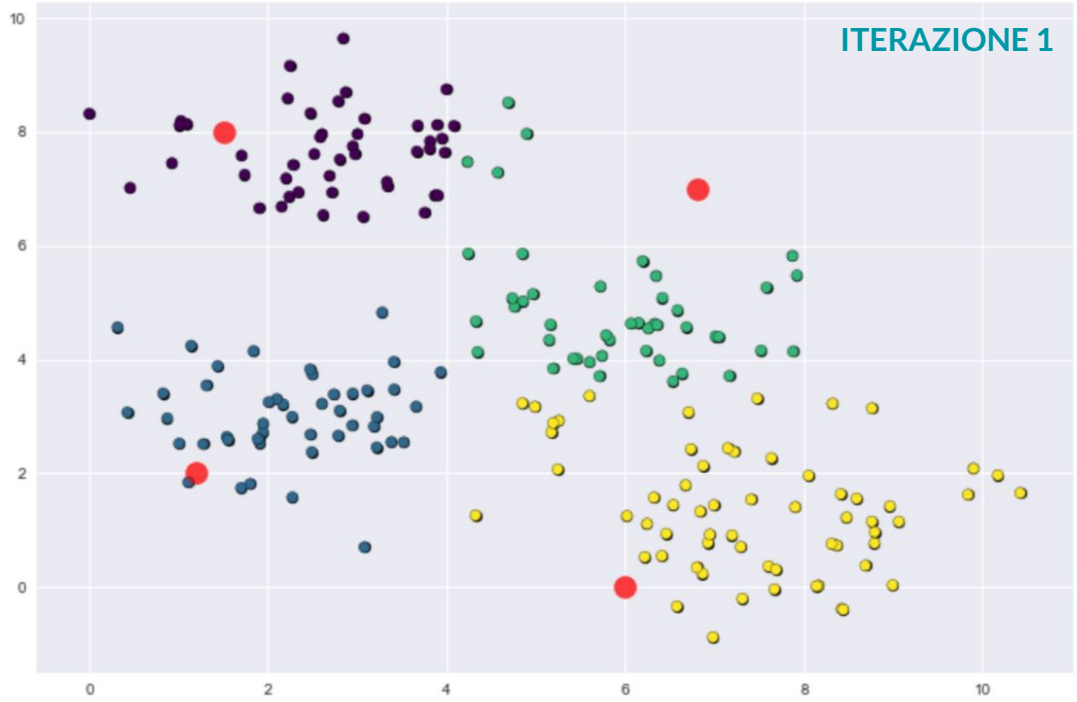
K-means: un esempio

3. Calcolo la distanza tra ogni centroide e tutte le osservazioni.



K-means: un esempio

4. Assegno le osservazioni al cluster rappresentato dal centroide più vicino.

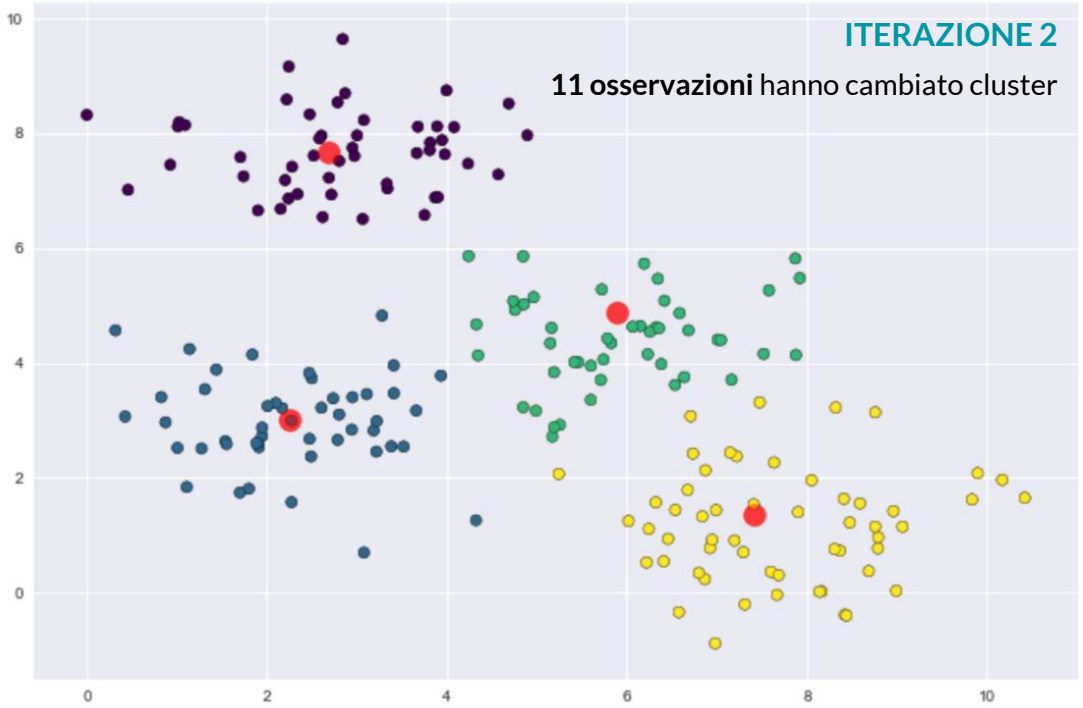


K-means: un esempio

5. Ricalcolo i centroidi come la media degli esempi per ogni cluster.

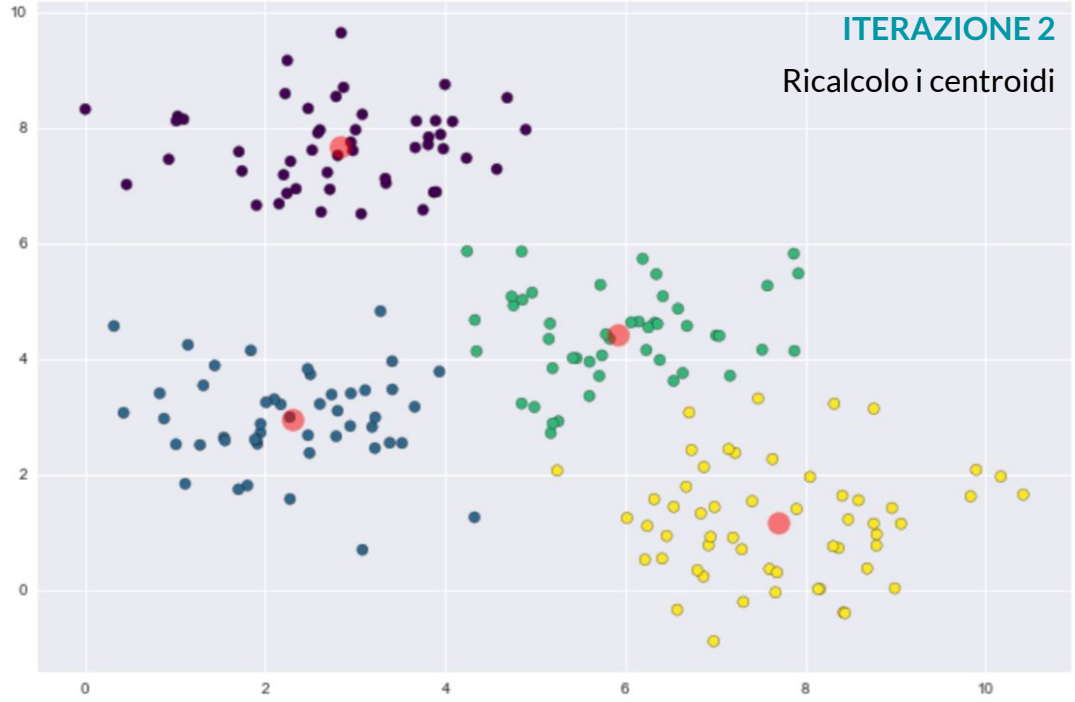
K-means: un esempio

6. Ripeto dal punto 2 fino a quando nessun esempio cambia più cluster.



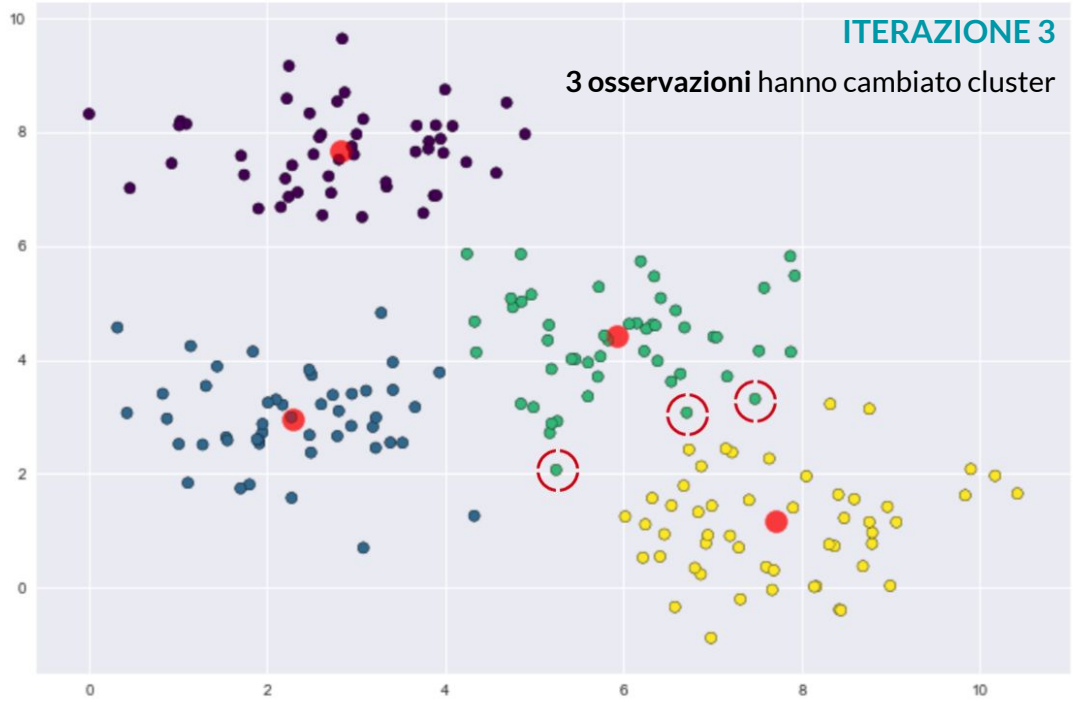
K-means: un esempio

6. Ripeto dal punto 2 fino a quando nessun esempio cambia più cluster.



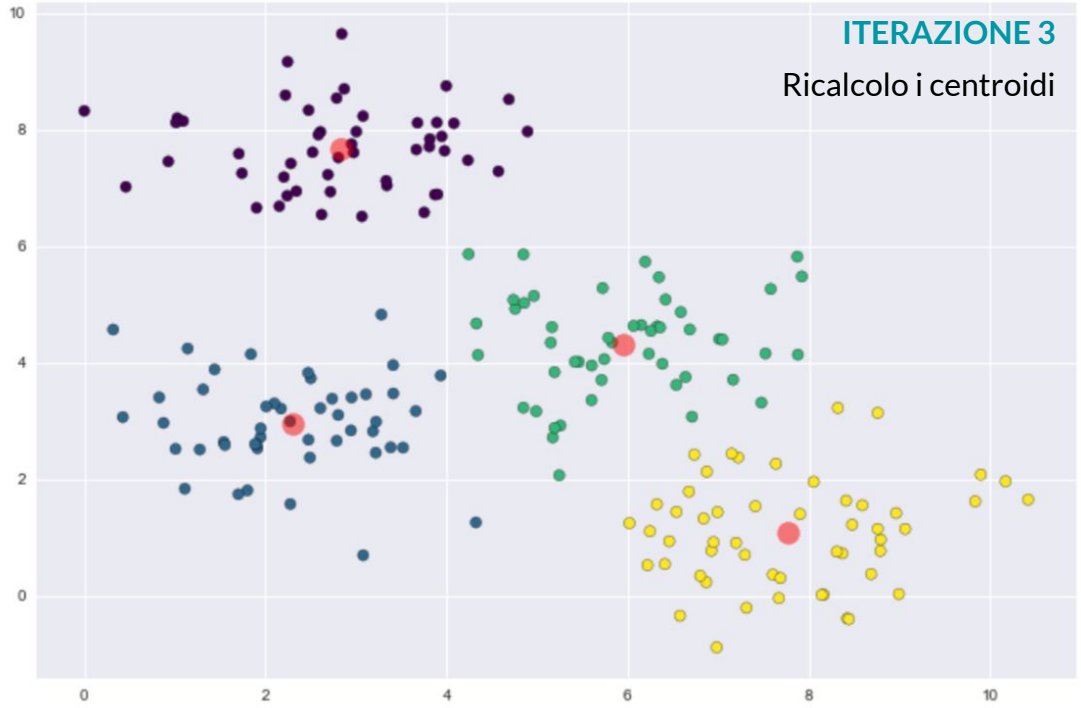
K-means: un esempio

6. Ripeto dal punto 2 fino a quando nessun esempio cambia più cluster.



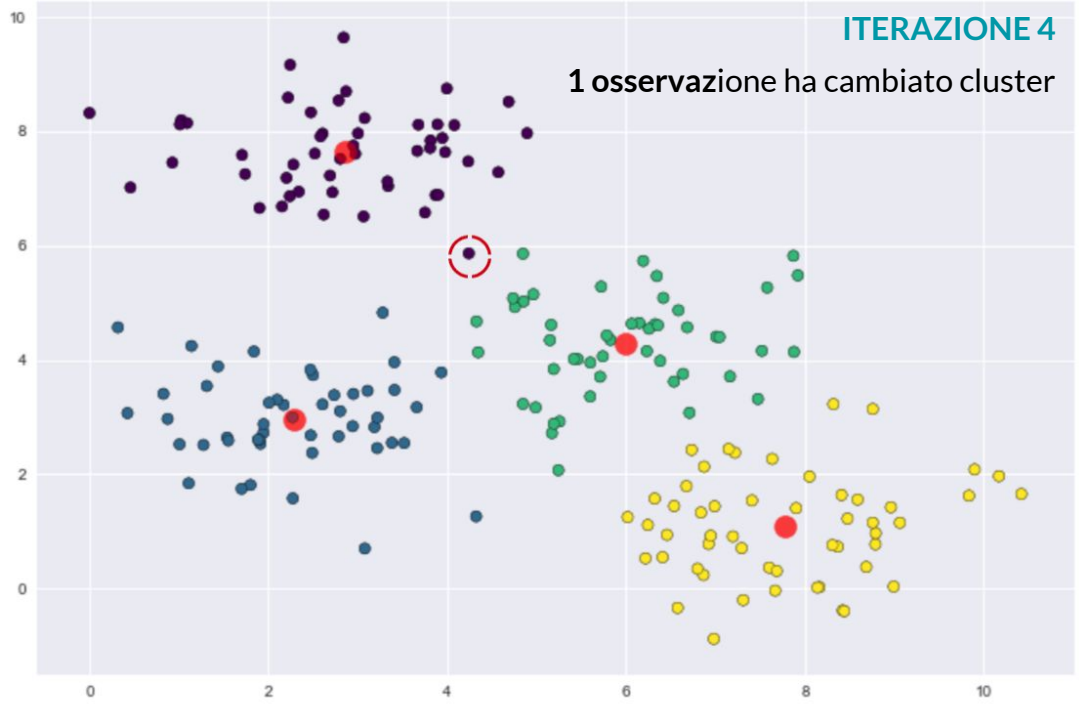
K-means: un esempio

6. Ripeto dal punto 2 fino a quando nessun esempio cambia più cluster.



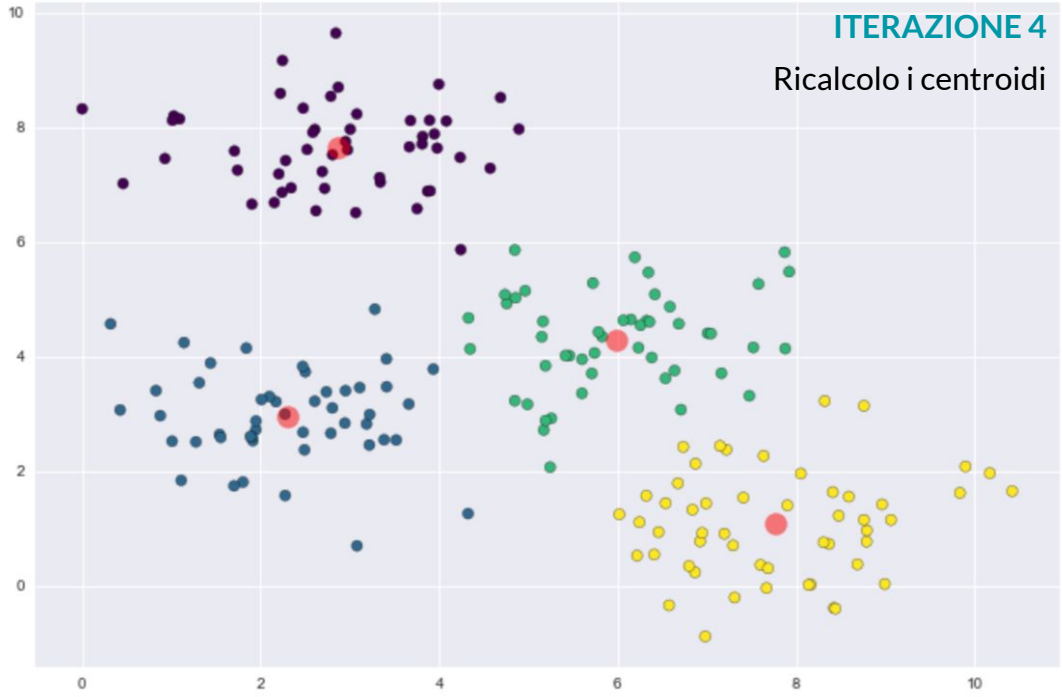
K-means: un esempio

6. Ripeto dal punto 2 fino a quando nessun esempio cambia più cluster.



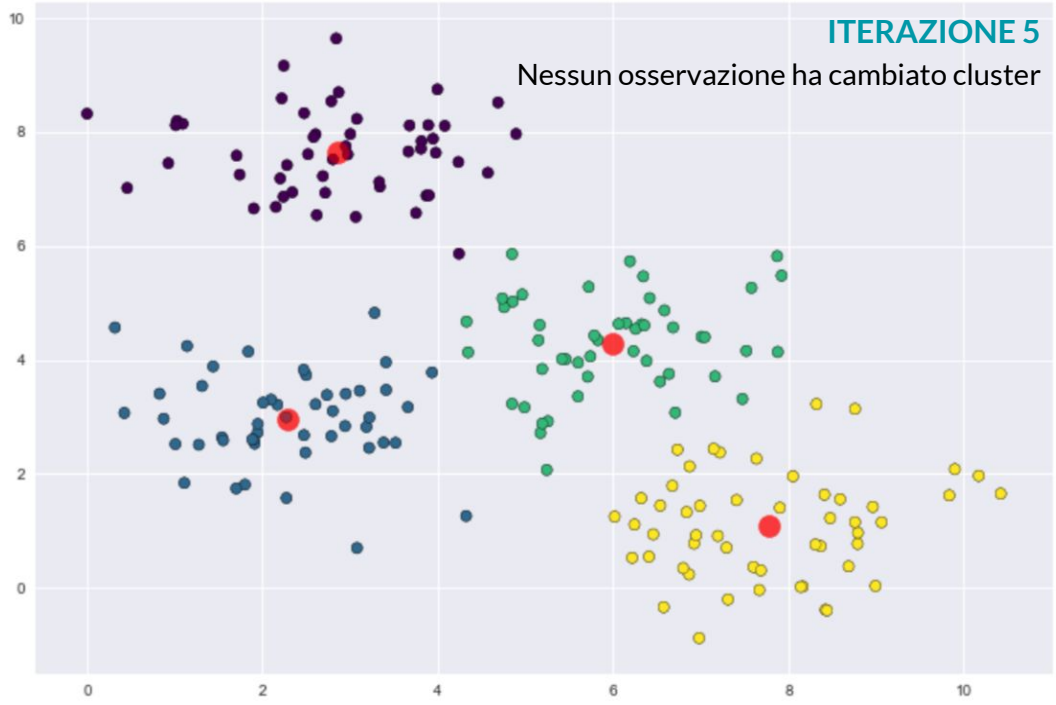
K-means: un esempio

6. Ripeto dal punto 2 fino a quando nessun esempio cambia più cluster.



K-means: un esempio

6. Ripeto dal punto 2 fino a quando nessun esempio cambia più cluster.



Fondamenti di Machine Learning

Il Clustering

Valutare un modello di clustering

presentato da
Giuseppe Gullo

PROFESSION AI

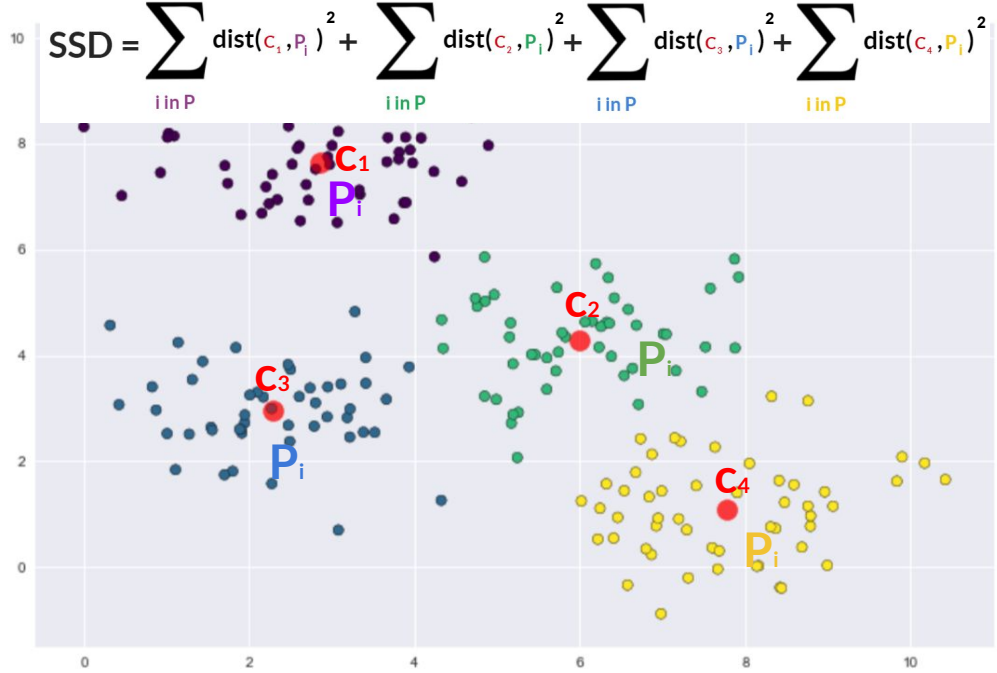
Valutare un modello di clustering

La valutazione di un modello di clustering è complessa perché non abbiamo a disposizione i valori reali.

Somma delle distanze al quadrato

a.k.a Within Cluster Sum of Squares (WCSS) o inertia

E’ la somma delle somme delle distanze al quadrato dei punti dal cluster



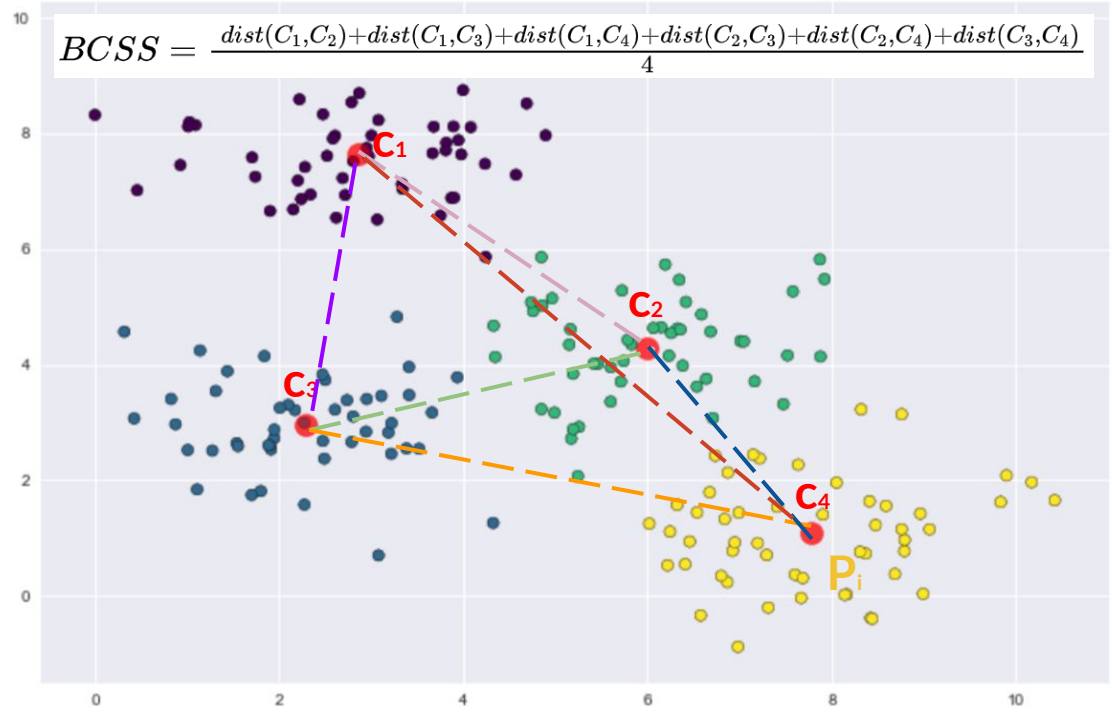
Media delle distanze al quadrato

a.k.a distortion

E’ la media delle medie della distanza al quadrato dei punti dal cluster.

Between Clusters Sum of Squares (BCSS)

E’ la media della distanza tra tutti i centroidi



Fondamenti di Machine Learning

Il Clustering

Determinare il numero di cluster

presentato da
Giuseppe Gullo

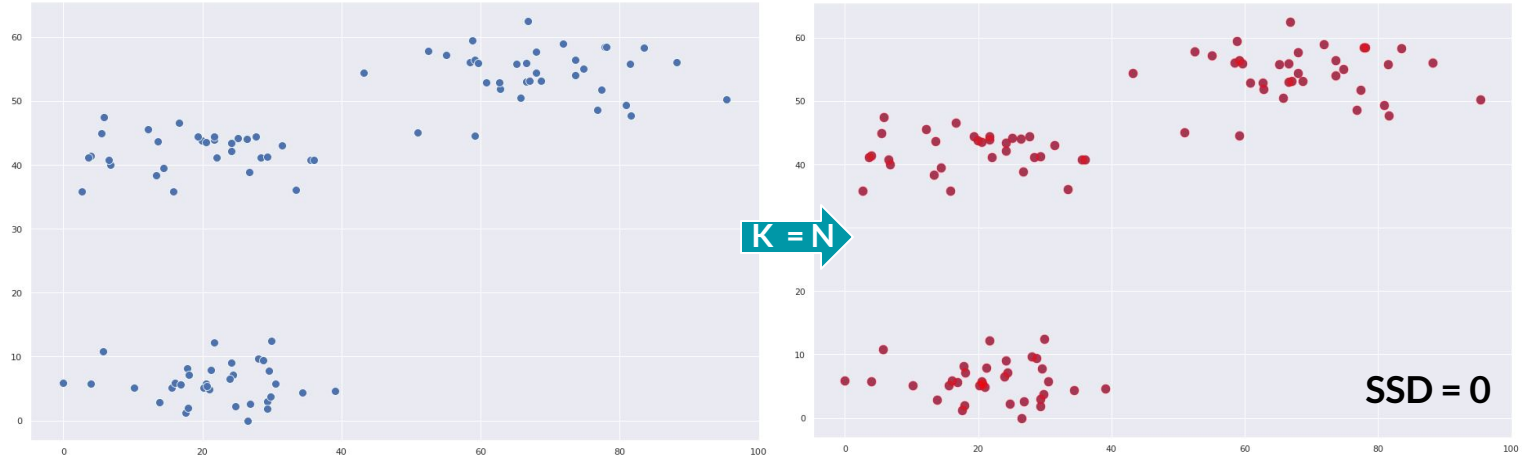
PROFESSION AI

Determinare il numero di cluster

Possiamo testare diversi modelli con diversi valori di k e confrontare i risultati.

Come confrontare i risultati?

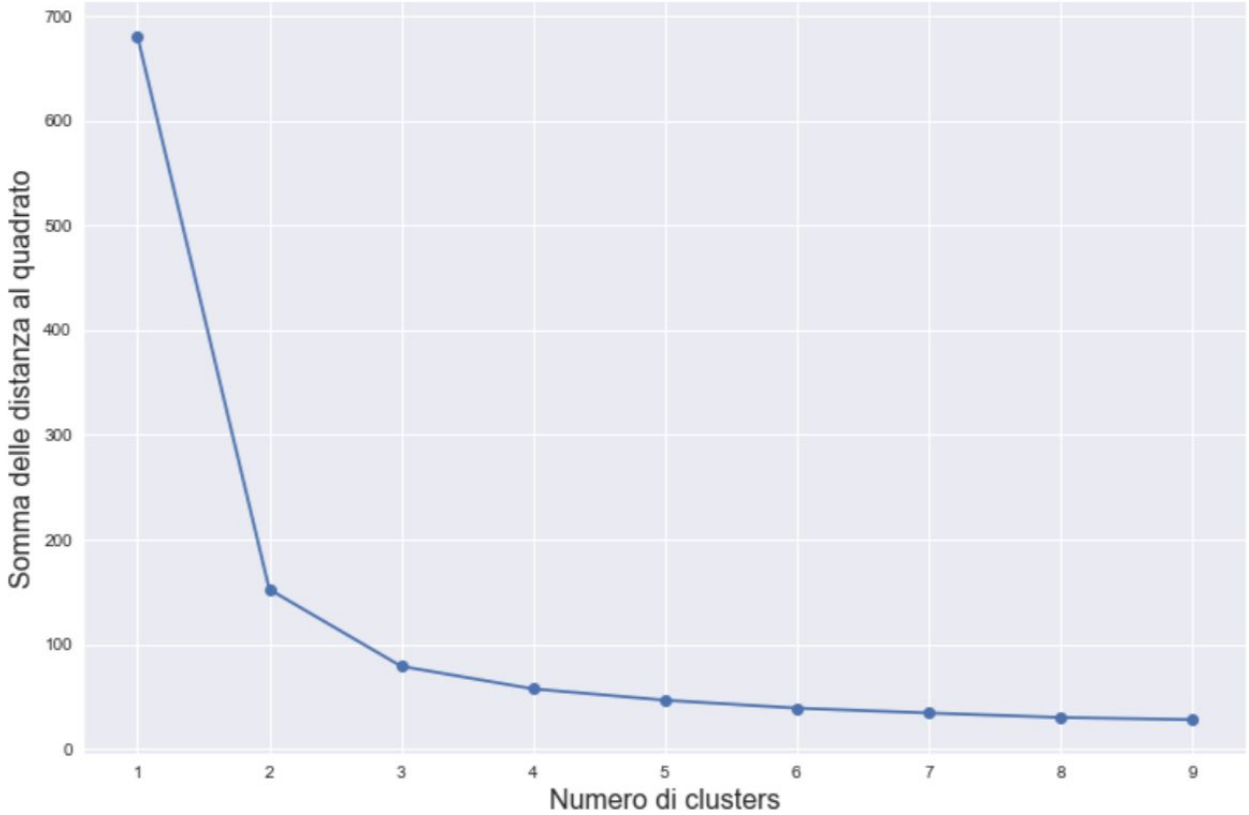
Confrontare la somma delle distanze al quadrato **è sbagliato**!



L’elbow method

E’ un euristica visiva, che consiste nel selezionare il gomito (elbow) della curva della somma delle distanze al quadrato per modelli addestrati con valori di k incrementali.

L’elbow method



L’elbow method