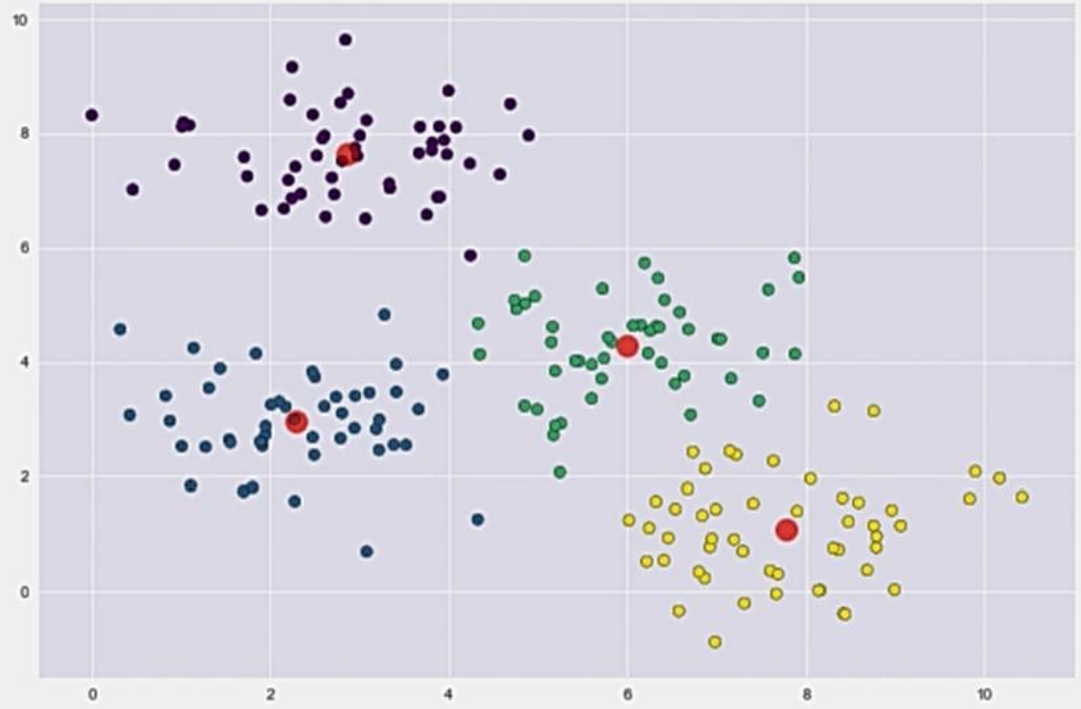
K-means clustering



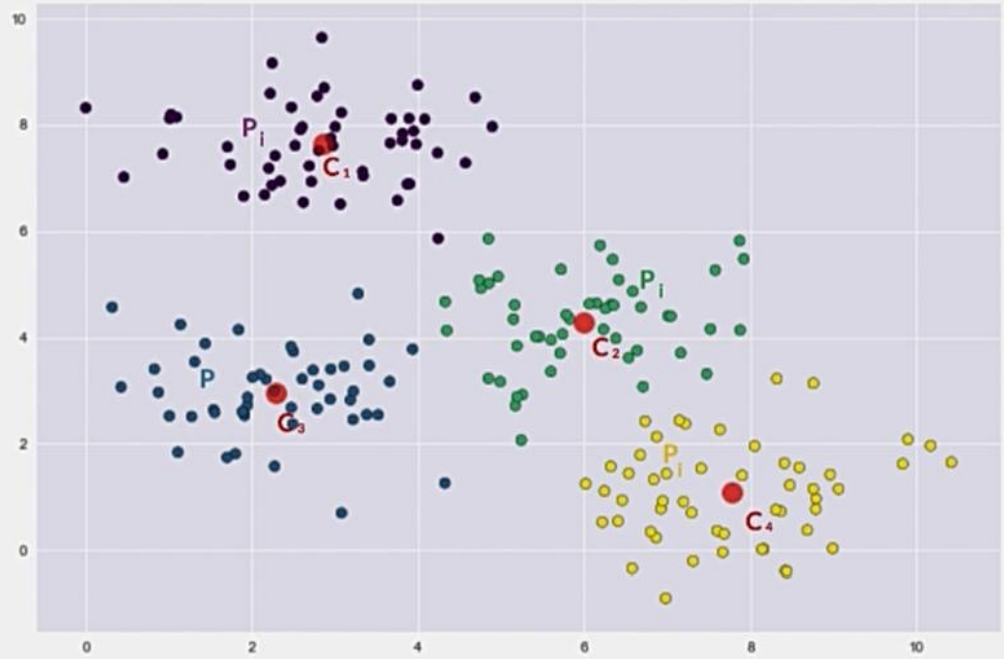
1. Scelgo il numero di clusters K da creare
2. Seleziono casualmente K **centroidi**
3. Calcolo la **distanza** tra ogni centroide e tutte le osservazioni
4. Assegno le osservazioni al cluster rappresentato dal centroide **più vicino**
5. Ricalcolo i centroidi come **la media** degli esempi per ogni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

Come scegliere il numero di Clusters ?

(valore di K)

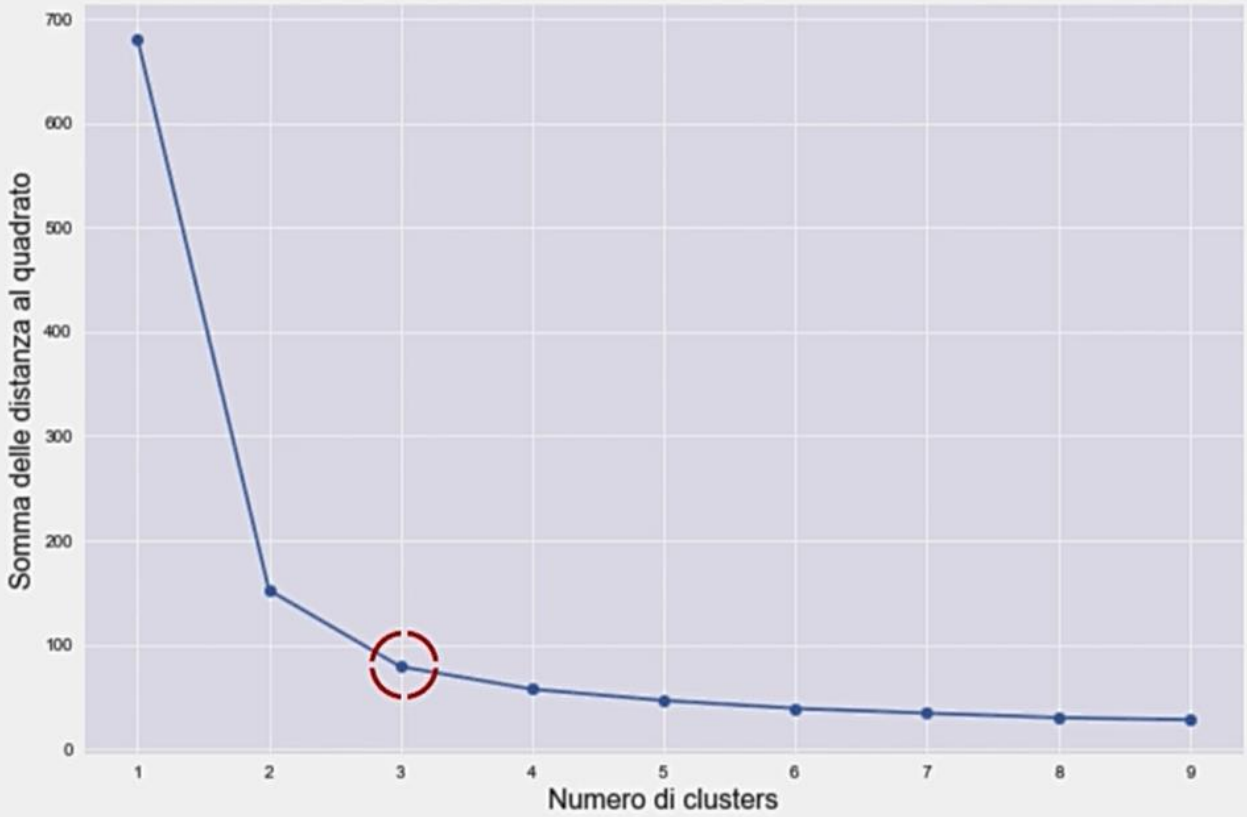
Testando diversi valori di K e confrontando i risultati

SOMMA DELLE DISTANZE AL QUADRATO



$$SSD = \sum_{i \text{ in } P} dist(c_1, P_i) + \sum_{i \text{ in } P} dist(c_2, P_i) + \sum_{i \text{ in } P} dist(c_3, P_i) + \sum_{i \text{ in } P} dist(c_4, P_i)$$

Come scegliere il numero di Clusters ?

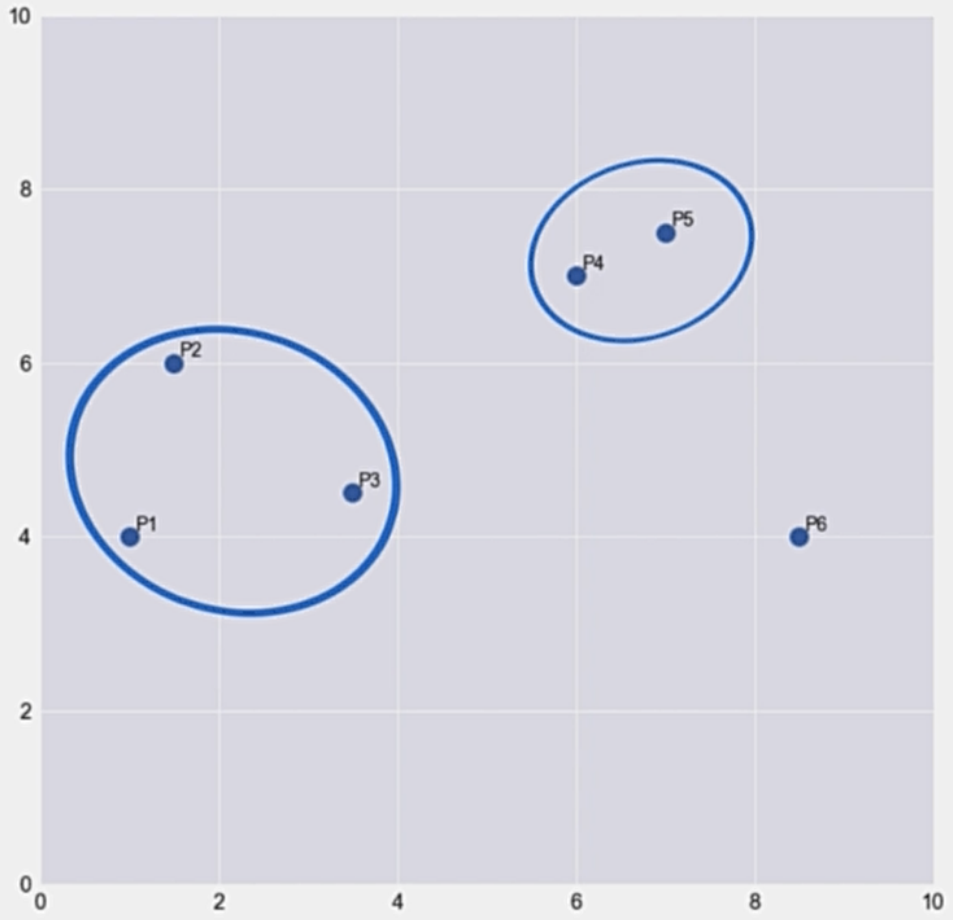


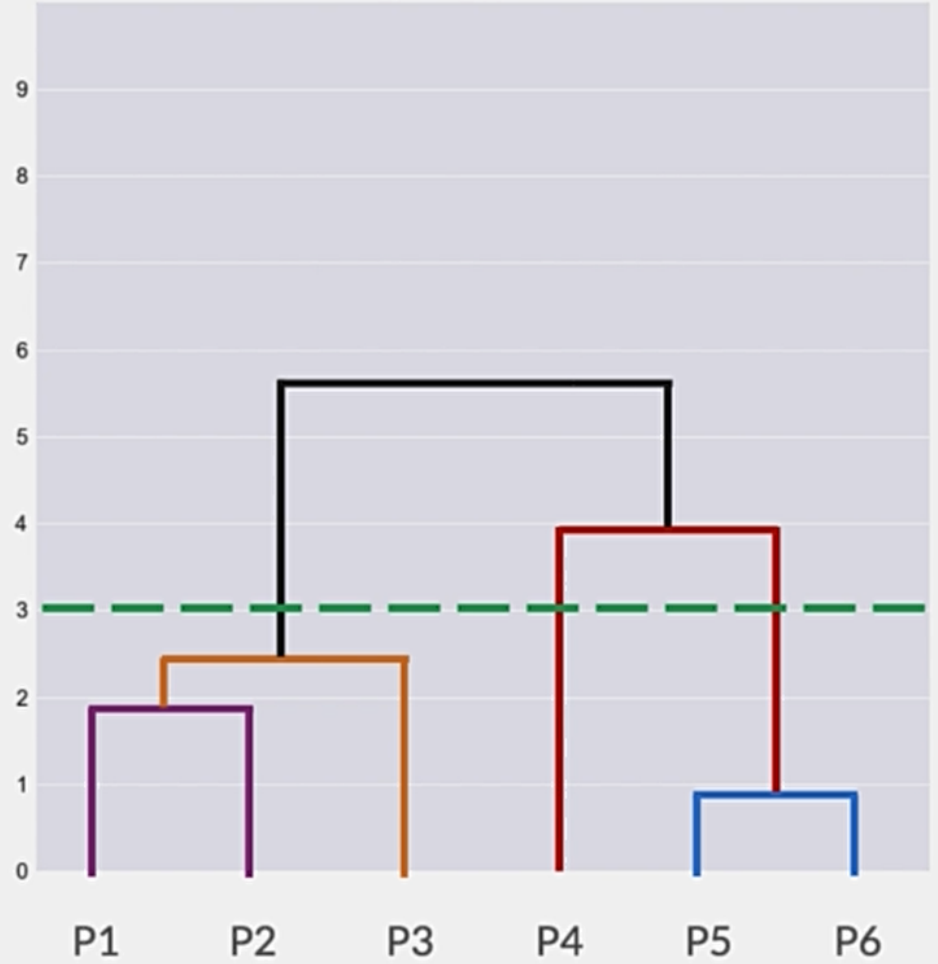
Utilizzare **l'elbow method** per determinare il valore di k ottimale

Clusters

Dendrogramma

ITERAZIONE 1





Clustering Gerarchico

PRO

Non serve definire
il numero di cluster a priori

CONTRO

E' dispendioso in termini
di risorse di calcolo

K-Means	Il numero di cluster va definito a priori
Clustering gerarchico	Il numero di cluster va definito a posteriori
DBSCAN	Non serve definire il numero di cluster

Parametri del DBSCAN

ε **(eps)**	Distanza massima tra due osservazioni nello stesso vicinato
minPts	Numero minimo di osservazioni richieste per formare un cluster ($\geq$ numDims+1, min=3)

DBSCAN

- Scelgo i valori di **eps** e **minPts**
- Per ogni osservazione:

  Ci sono più di **minPts** osservazioni in un raggio di distanza **eps** dall'osservazione ?

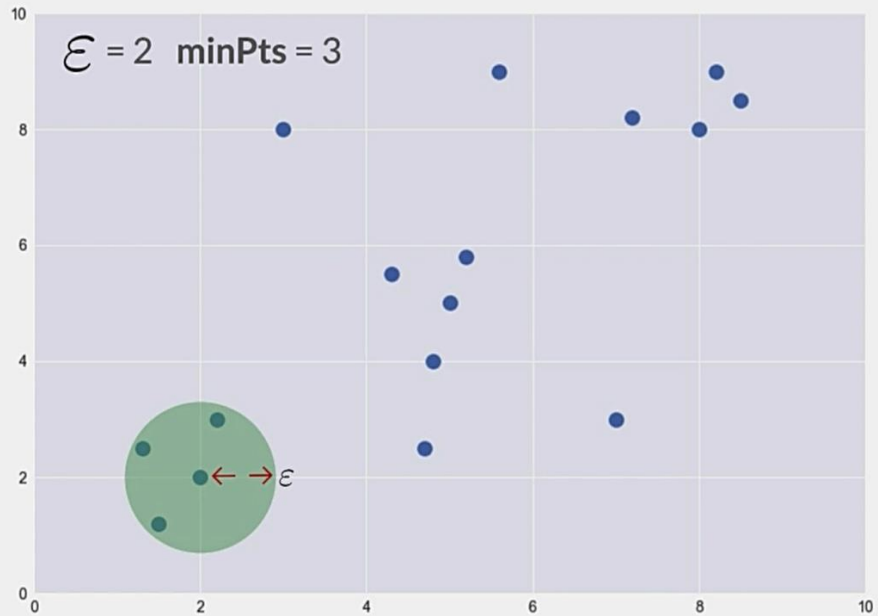
  SI → l'osservazione è un **core point** e forma un cluster

  NO → c'è un core point nel raggio di distanza eps dall'osservazione ?

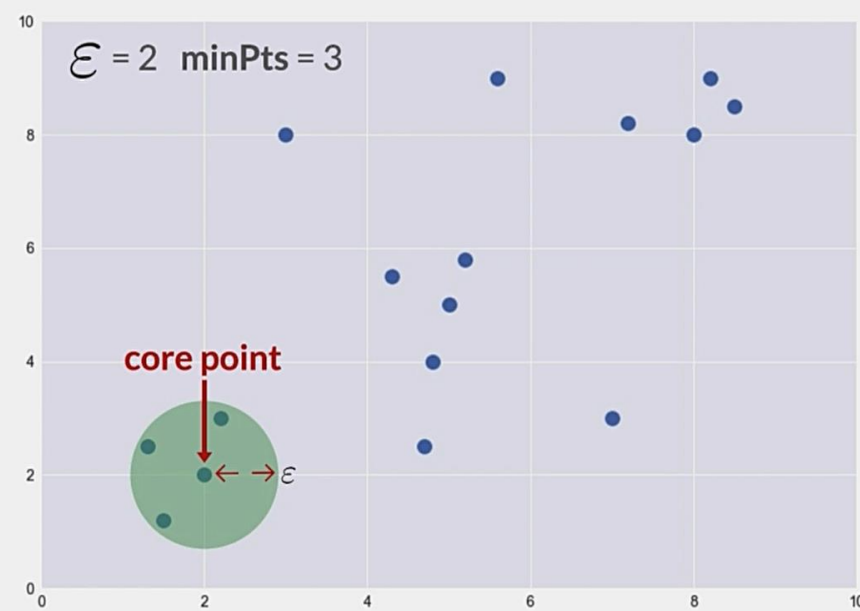
  SI → l'osservazione è un **border point**
  e viene assegnata al cluster rappresentato dal core point

  NO → l'osservazione è un **noise point**
  e non viene assegnata a nessun cluster

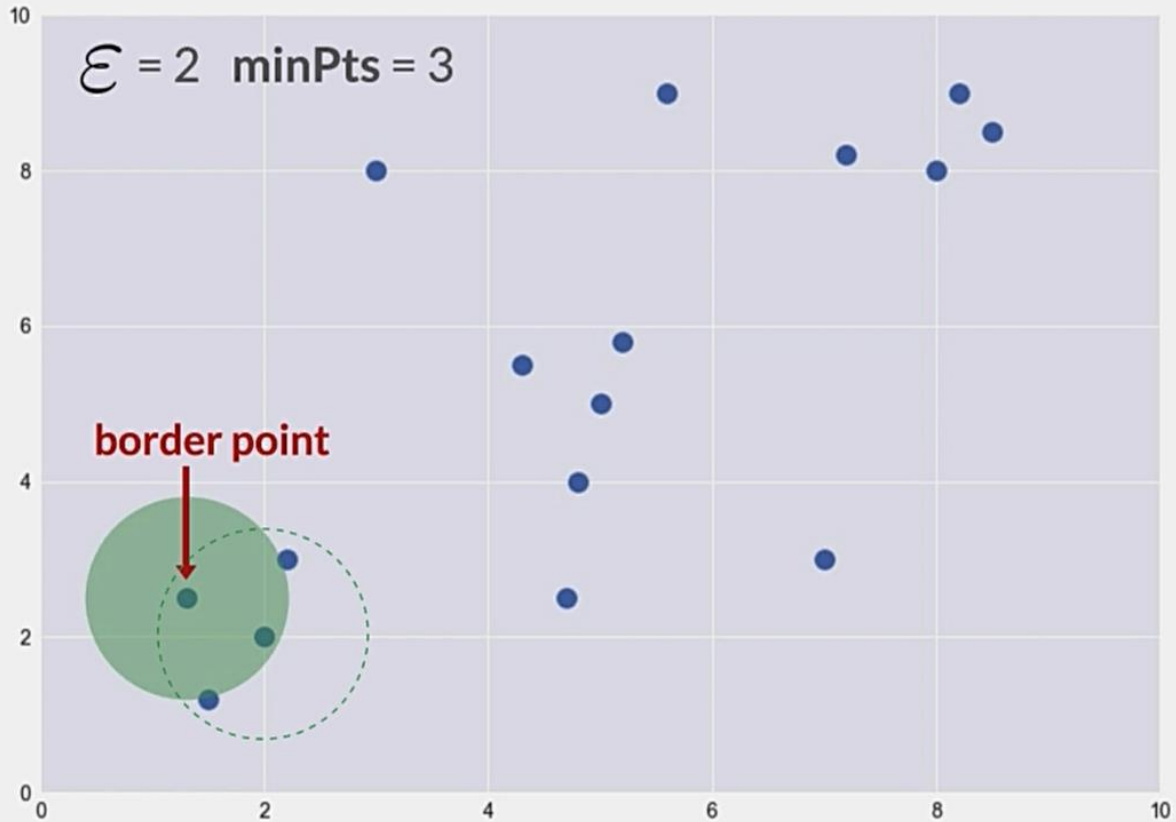
DBSCAN: UN ESEMPIO

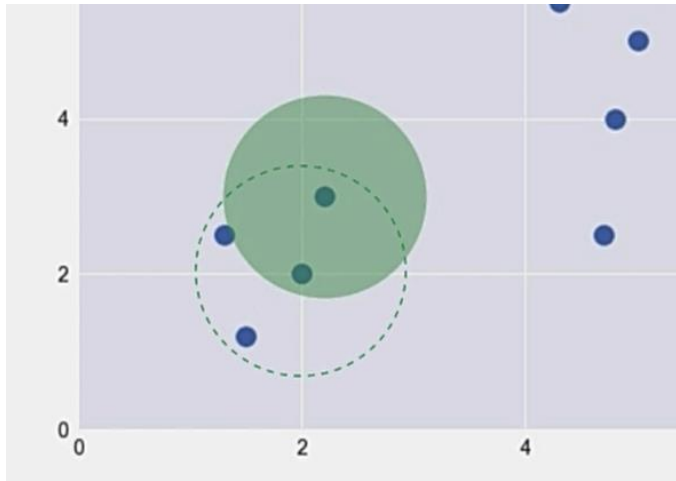


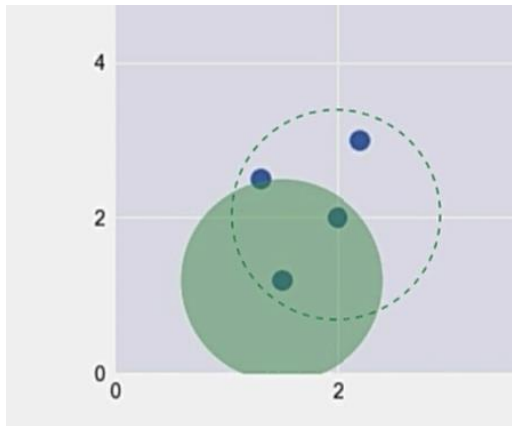
DBSCAN: UN ESEMPIO



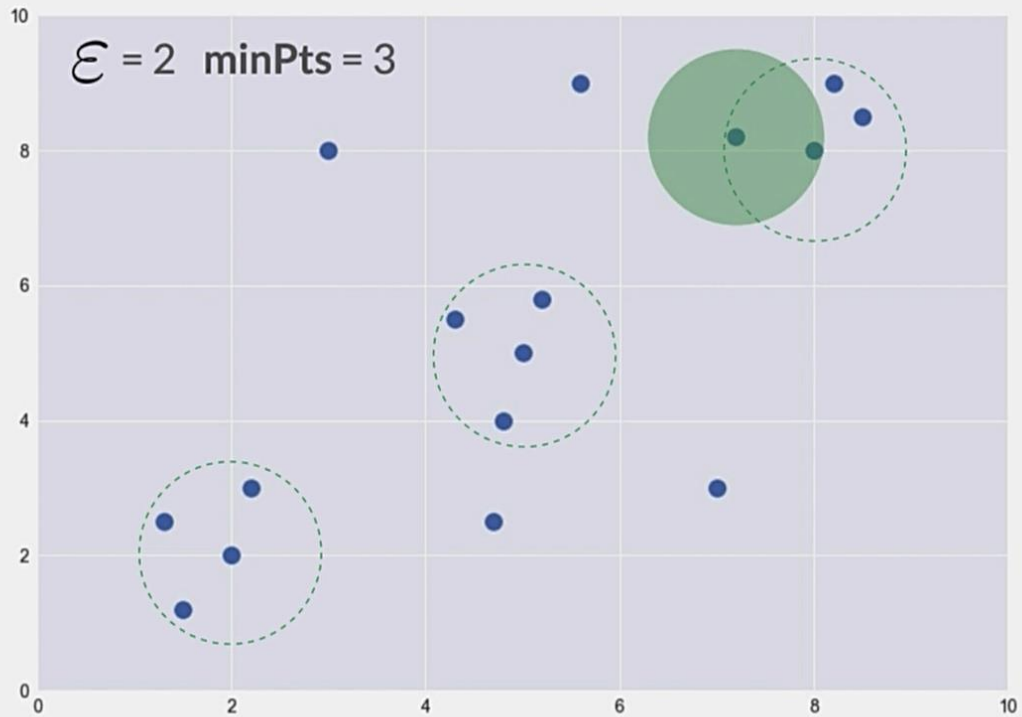
DBSCAN: UN ESEMPIO



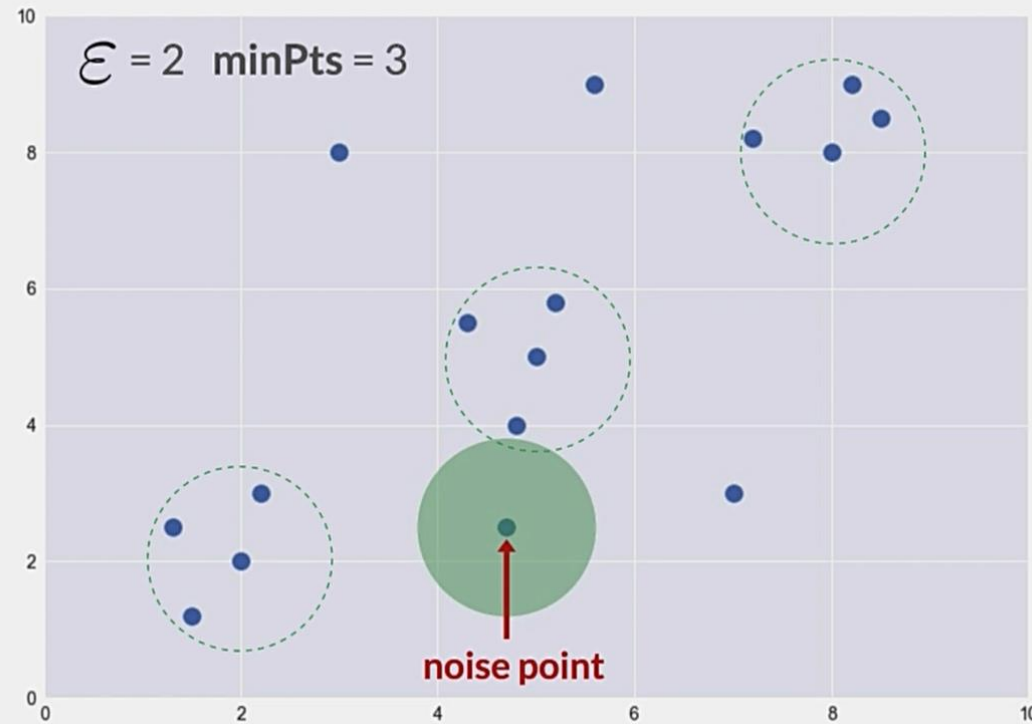




DBSCAN: UN ESEMPIO

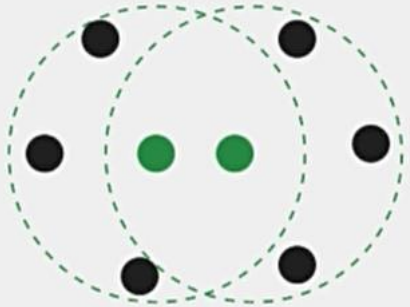


DBSCAN: UN ESEMPIO

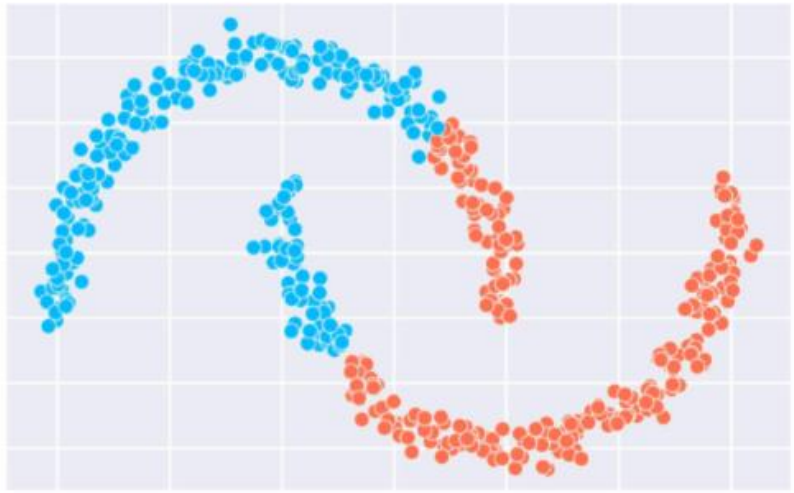


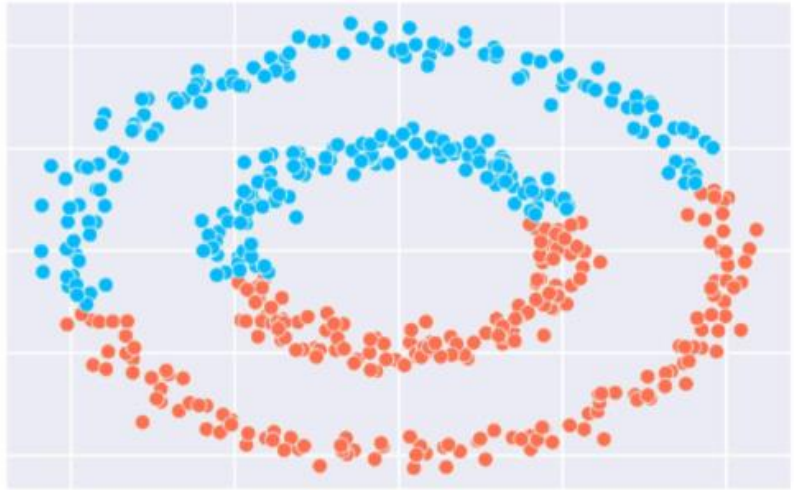
DBSCAN: VANTAGGI

- Non serve definire il numero di cluster
- E' resistente agli outlier
- Non limita i cluster ad una forma sferica

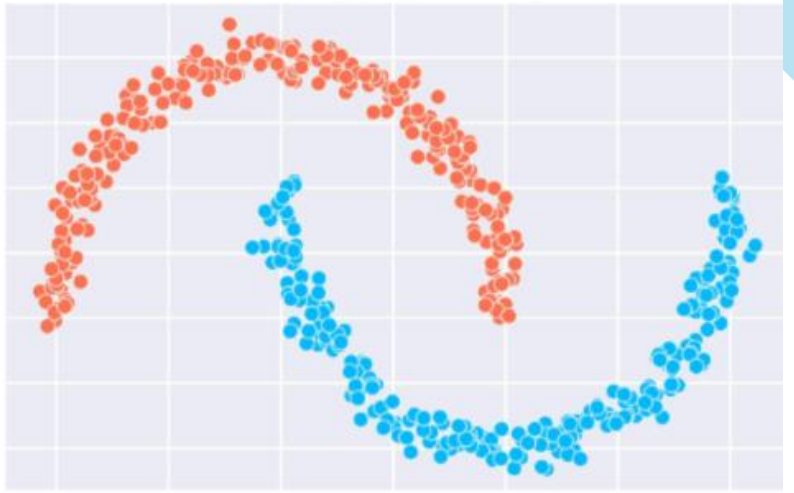


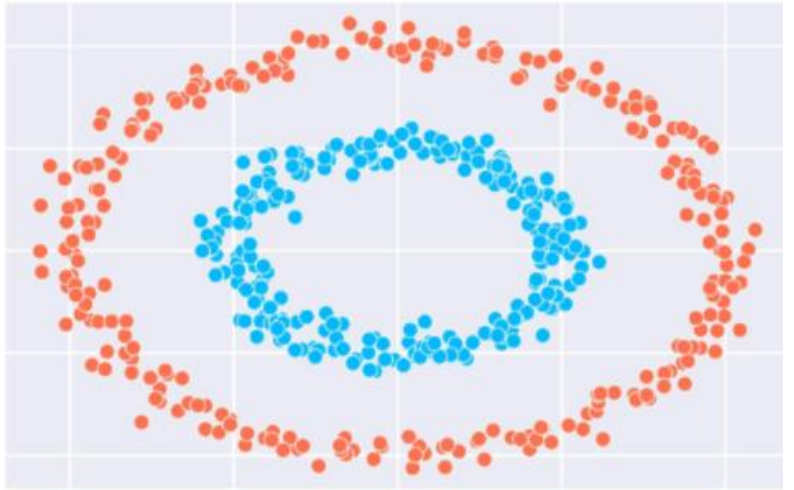
KMeans



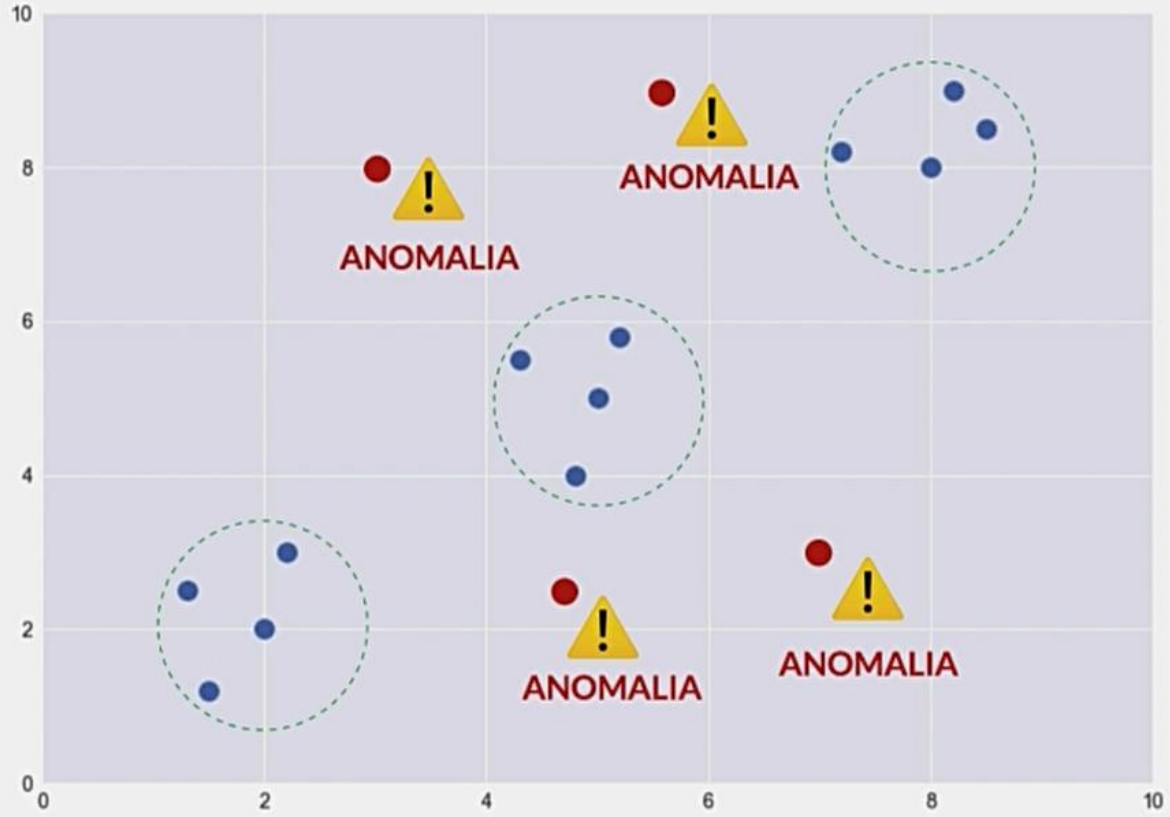


DBSCAN





DBSCAN: ANOMALY DETECTION



Identificare pattern inaspettati nei dati

ANOMALIA

=

OUTLIER

=

NOISE POINT