Valutazione del clustering

Algoritmi per l'intelligenza artificiale

Vincenzo Bonnici
Corso di Laurea Magistrale in Scienze Informatiche
Dipartimento di Scienze Matematiche, Fisiche e Informatiche
Universitá degli Studi di Parma

2025-2026

Valutare/Validare il clustering

The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

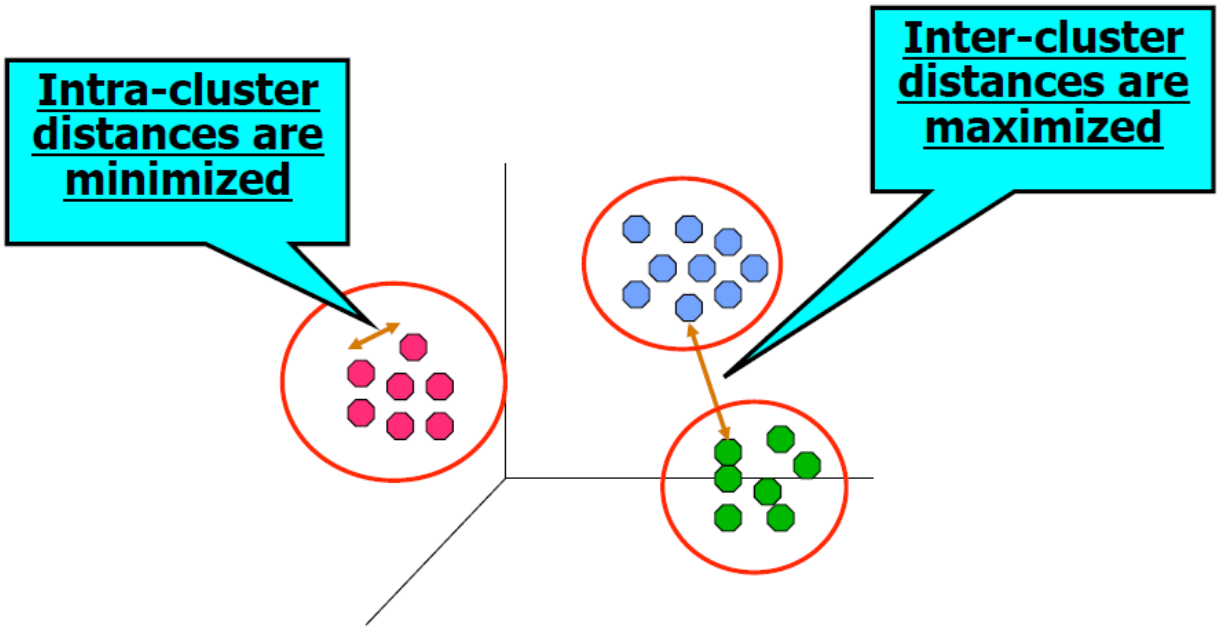
Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.

Algorithms for Clustering Data, Jain and Dubes.

Bontà del clustering

Un buon metodo di clustering produrrà cluster di **alta qualità** con

- **alta** similarità **intra**-classe
- **bassa** similarità **inter**-classe

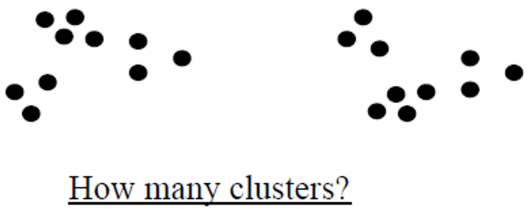


Bontà del clustering

La qualità del risultato del clustering dipende:
dalla **misura di similarità** usata, o dallo **specifico algoritmo** usato.

La qualità del clustering è anche misurato in base alla sua abilità di scoprire alcuni o tutti i **pattern nascosti**.

Purtroppo la nozione di cluster può essere **ambigua**:



How many clusters?

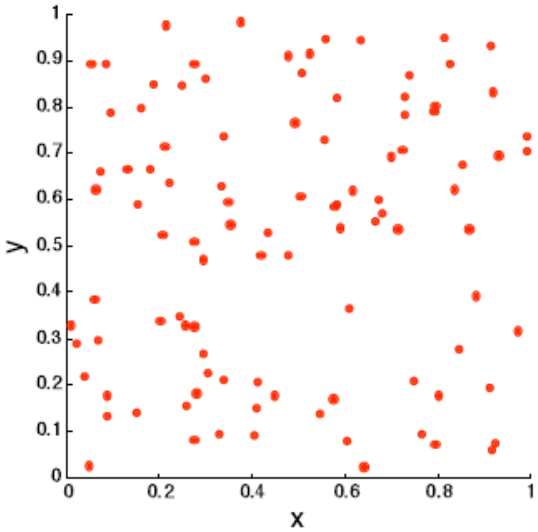
Six Clusters

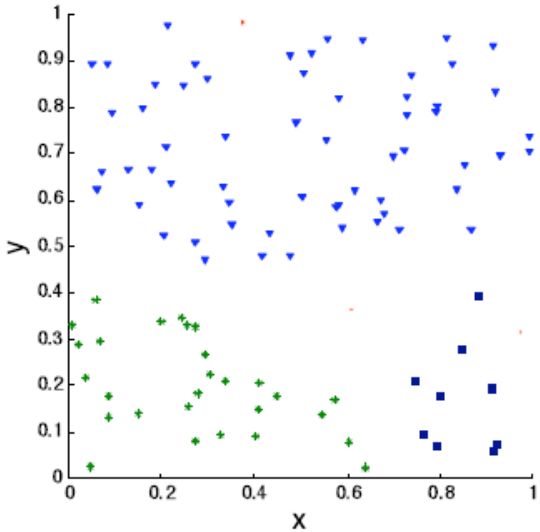
Two Clusters

Four Clusters

Dati randomici

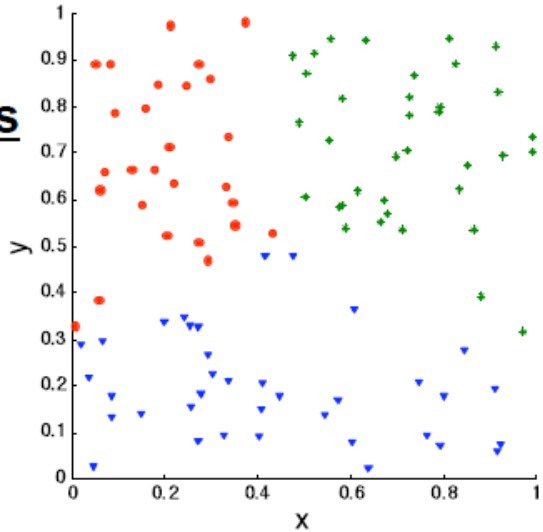
Random Points

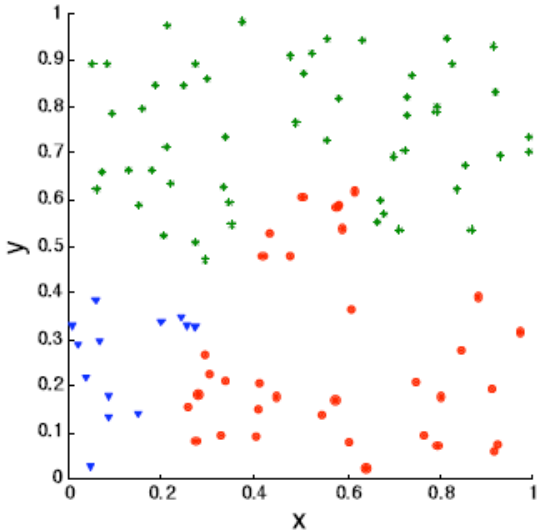




DBSCAN

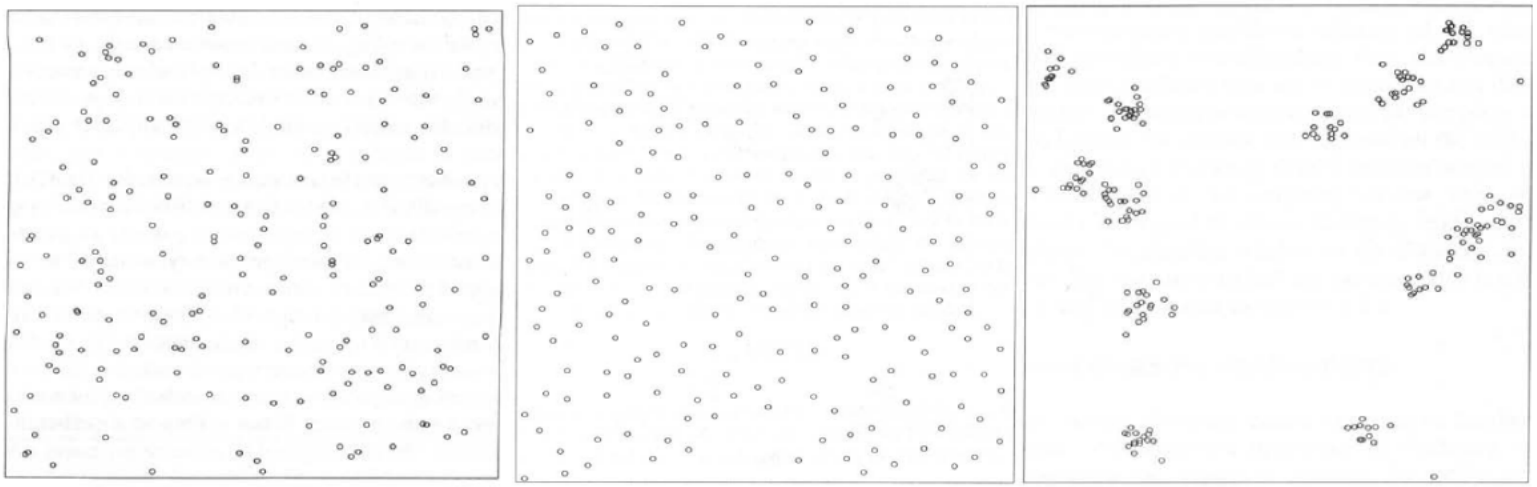
K-means





Hierachical (MAX)

Dati randomici e non



random regular cluster

Validare il clustering

Come facciamo a validare la bontà di un cluster?

Perché valutare?

- Per evitare di trovare pattern quando invece trattasi di rumore
- Per comparare algoritmi diversi
- Per valutare due insiemi di cluster (due risultati globali)
- Per comparare due cluster

Validare il clustering

Determinare la **clustering tendency** di un data set:

- Una struttura non-random esiste realmente nei dati?
- Qual è il numero corretto di cluster?

Confrontare i risultati rispetto a **conoscenze esterne**.

Valutare i risultati con **parametri interni**, senza usare conoscenza pregressa.

Nota che le stesse metodologie possono essere usate per confrontare due cluster, oppure due insiemi di cluster (clustering) ottenuti da un algoritmo.

Misure per validare il clustering

External Index:

- Misura quanto i cluster individuati corrispondono a etichette di classe fornite esternamente (conoscenza pregressa).
  - Entropia

Internal Index:

- Misura la qualità del clustering senza informazione/conoscenza esterna
  - Sum of Squared Error (SSE)

In letteratura sono spesso riferiti come **criteri** invece di **indici**.
Comunque, il criterio è la strategia generale, mentre l'indice è la misura numerica che la implementa.

Misura tramite correlazione (indice interno)

Due matrici

- **Matrice di Prossimità** $(n \times n)$ per gli n oggetti
  - Similarità tra ogni coppia di oggetti. Dovrebbe essere ≈ 1 per coppie che stanno nello stesso cluster
- **Matrice di Incidenza** $(n \times n)$
  - Entry (i,j) $= 1$: i due oggetti appartengono allo stesso cluster
  - Entry (i,j) $= 0$: i due oggetti appartengono a cluster differenti

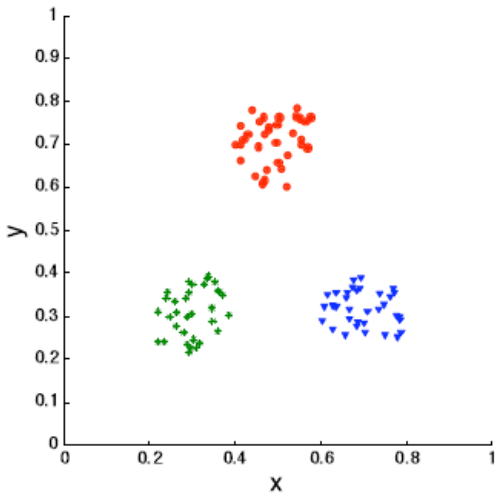
Calcola la **correlazione** tra due matrici

- Le matrici sono simmetriche, per cui solo la correlazione tra $(n-1)/2$ entry delle due matrici deve essere calcolata
- **Alta correlazione**: I punti appartenenti allo stesso cluster sono vicini

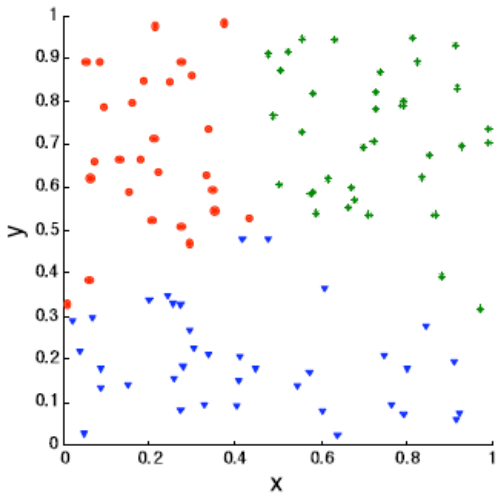
Misura non adatta per cluster costruiti sulla base della densità o contiguità spaziale dei punti.

Misura tramite correlazione (indice interno)

Correlazione delle due matrici (incidenza e prossimità) per K-means su due diversi dataset.



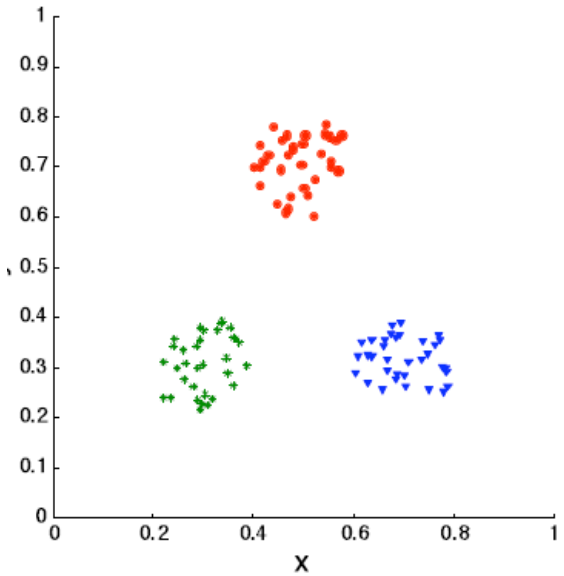
Corr = -0.9235

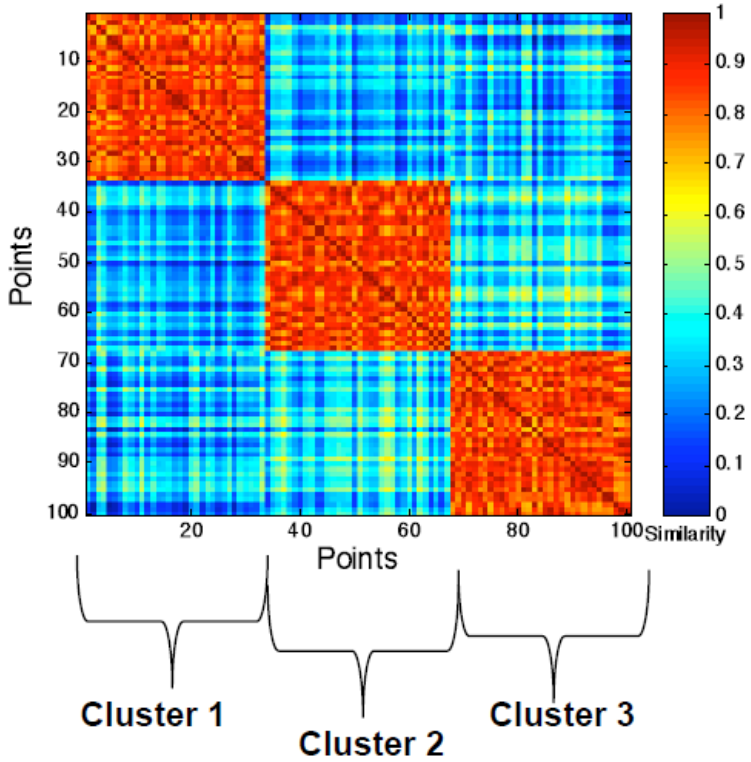


Corr = -0.5810

Matrice di similarità per validazione visuale

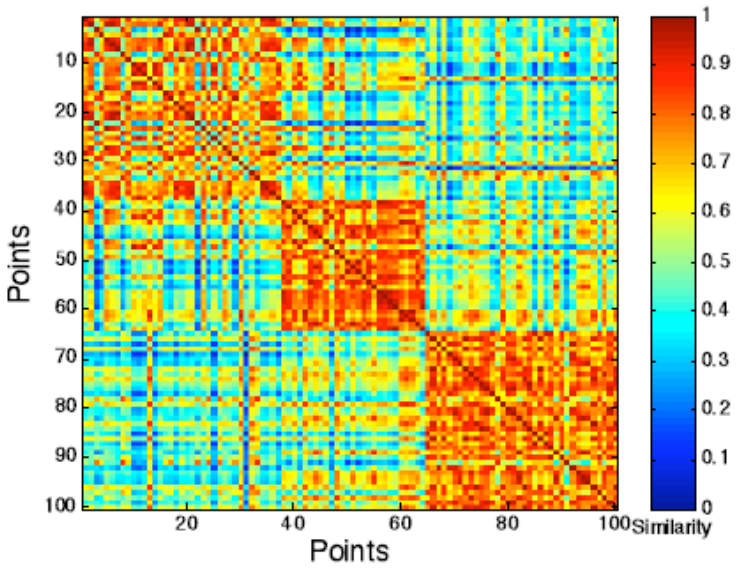
Ordinare righe (e colonne) rispetto alle etichette dei cluster.

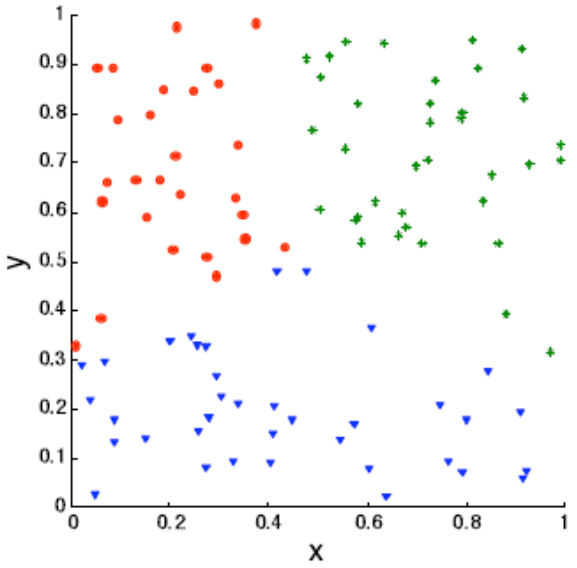




Matrice di similarità per validazione visuale

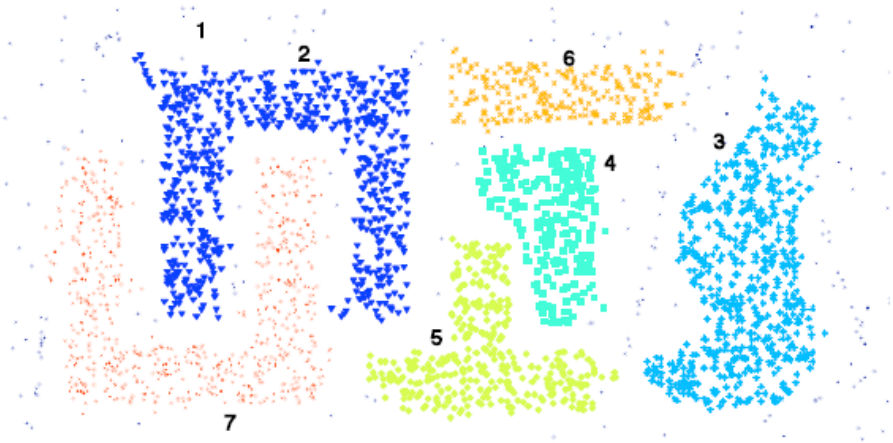
I cluster in dati random non sono molto definiti. (k-means)

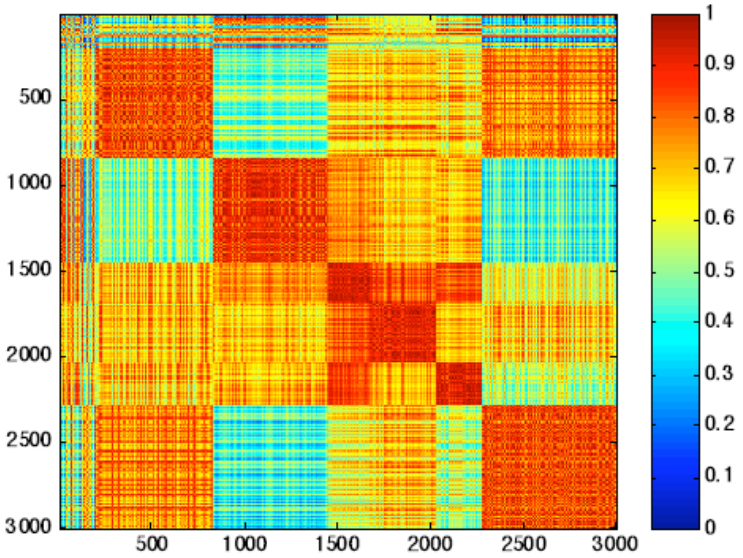




Matrice di similarità per validazione visuale

La misura di similarità non è adatta per valutare DBSCAN.



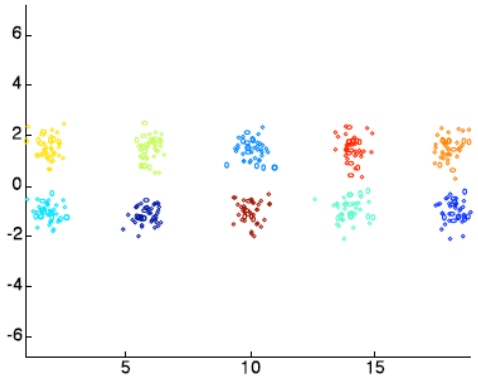


Validazioen tramite SSE (indice interno)

Si può usare solo se è definito un prototipo (centroide/medoide) degli elementi del cluster.

SSE è un buon indice per confrontare sia due clustering e sia due cluster.

Può essere usato anche per stimare il numero di cluster ottimale nel K-means.

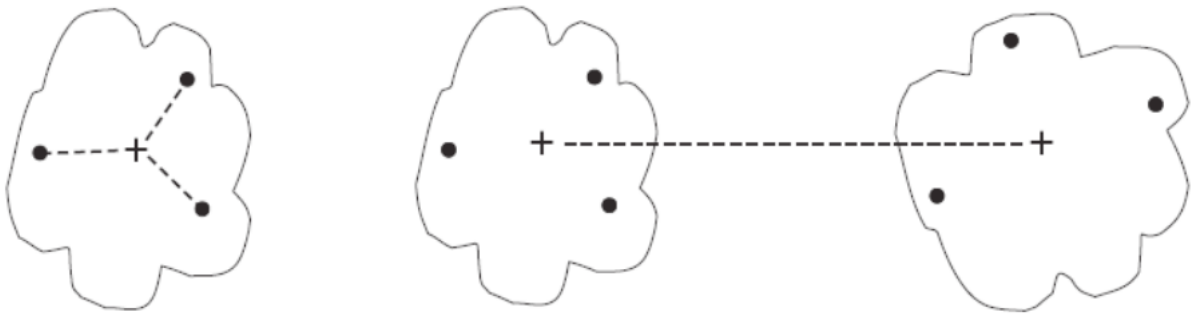


Altre misure interne: coesione e separazione

$$overall\_validity = \sum_{i=1}^{K} validity(C_i)$$

dove $validity(C_i)$ può essere l'indice di coesione o di separazione dell'i-esimo cluster.

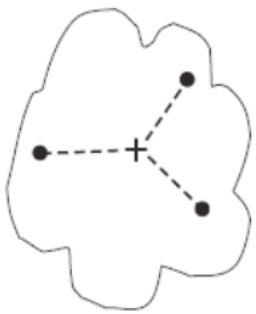
- **coesione**:misura l'affinità tra gli oggetti di un cluster
- **separazione**: misura quanto i cluster sono distinti e ben separati rispetto agli altri cluster



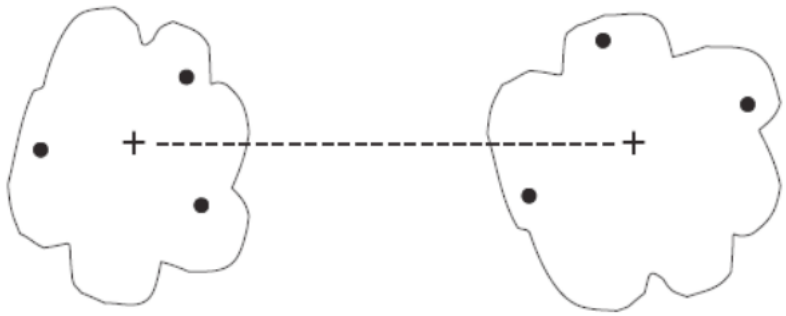
(a) Cohesion. (b) Separation.

Altre misure interne: coesione e separazione

Prototype-based view:

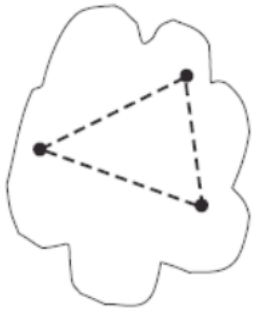


(a) Cohesion.

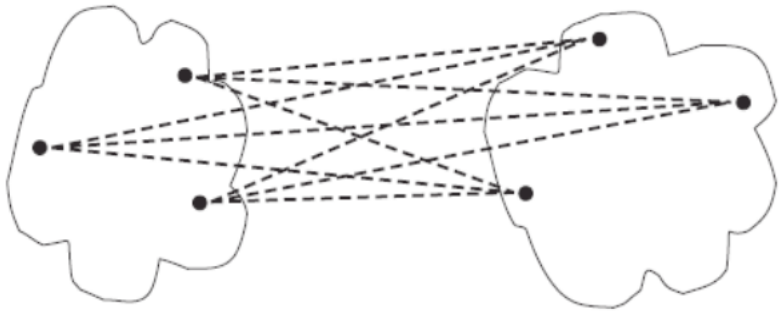


(b) Separation.

Graph-based view:



(a) Cohesion.



(b) Separation.

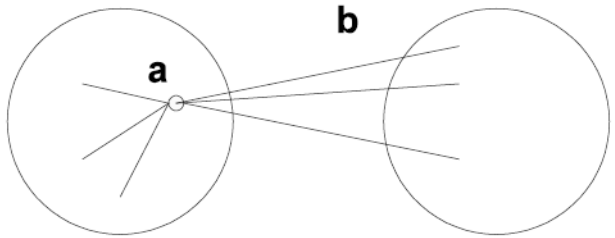
Altre misure interne: coeffciente di silhouette

Il **Silhouette Coefficient** combina le idee della coesione e della separation (per singoli punti, cluster singoli, o risultati del clustering).

Per un punto i

- Sia C_i il cluster di i
- Calcola: a_i = distanza media di i dagli altri punti di C_i
- Calcola: b_i = distanza media minima di i dai punti del cluster C t.c. $C \neq C_i$
- Silhouette Coefficient s_i per il punto i:

$$s_i = (b_i - a_i)/max(a_i, b_i)$$



Altre misure interne: coeffciente di silhouette

- Sempre tra -1 e 1.
- Caso **-1** non desiderabile, perché questo succederebbe se $a_i > b_i$
- Vorremmo avere un valore positivo (ovvero $a_i < b_i$), con a_i molto piccolo (≈ 0), in questo caso s_i tende a **1**

Coefficiente per un **singolo cluster**

- media dei coefficienti di tutti i punti del cluster

Coefficiente per un **clustering completo**

- media dei coefficienti di tutti i punti

Metodo del gomito

Il **metodo del gomito** utilizza un grafico tra la media della somma della somma intra-cluster dei quadrati delle distanze tra i rispettivi centroidi del cluster e i punti del cluster e il numero di cluster (o K).

Per determinare il numero ottimale di cluster, dobbiamo selezionare un tale valore di K nel punto "gomito", o il punto dopo il quale l'errore inizia a diminuire in modo lineare.

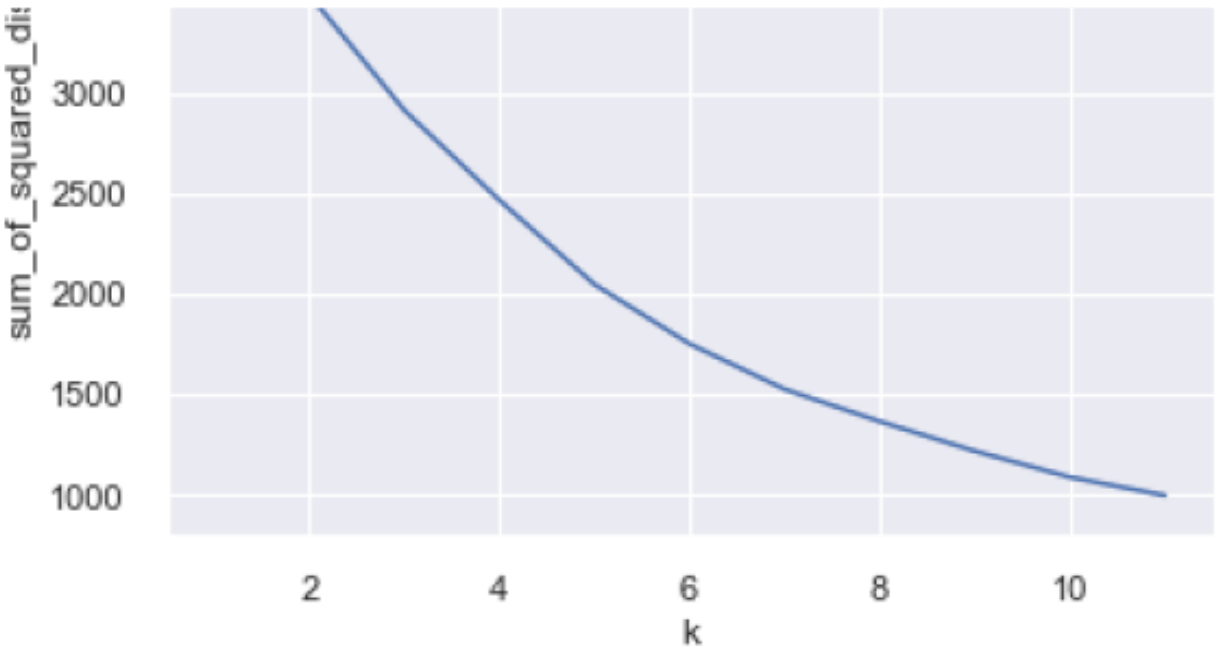
La curva può salire o scendere, ma se c'è un forte punto di flesso, è una buona indicazione che il modello sottostante si adatta meglio in quel punto per quanto riguarda il numero di cluster.

Il metodo del gomito non funziona bene se i dati non sono intrinsecamente molto raggruppati.

Metodo del gomito

Per impostazione predefinita, viene calcolato il punteggio di distorsione.

Prendendo la somma delle distanze al quadrato come metrica, otteniamo il seguente grafico del gomito per i nostri dati:



Qui, non possiamo vedere un punto del gomito molto distinto. Si potrebbe dedurre che il valore ottimale di K sia5, 6 o 7.

Metodo del gomito: Calinski_Harbasz

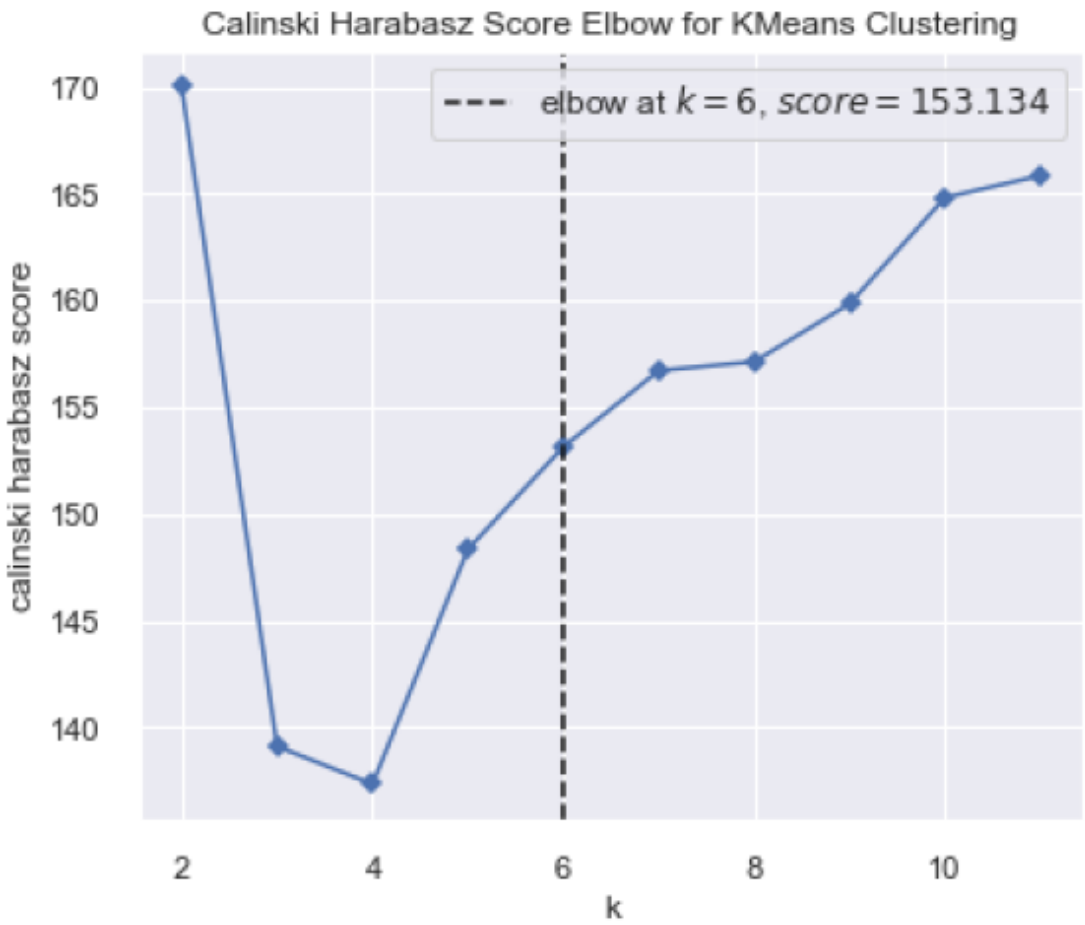
È possibile utilizzare anche altre metriche,tra cui il punteggio **Calinski_Harabasz** che calcola il rapporto di dispersione tra e all'interno dei cluster e quindi incorpora anche le informazioni sulla distanza tra i cluster. Più alto è il punteggio, migliore è la prestazione.

E' anche noto come criterio del rapporto di varianza, ed è definito come

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1}$$

dove $tr(B_k)$ è la traccia della matrice di dispersione tra gruppi e $tr(W_k)$ è la traccia della matrice di dispersione all'interno del cluster definita da:
$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$
$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$
con C_q i punti del cluster q, $n_q = |C_q|$, c_q il centroide di q. E l'insieme globale di punti, $n_E = |E|$, c_E in centroide di E.

Metodo del gomito: Calinski_Harbasz



Da questo grafico, possiamo vedere che il valore ottimale di K è 6. Sebbene il punteggio calinski_harbasz sia veloce da calcolare, è generalmente più alto per i cluster convessi rispetto ad altri tipi di cluster come i cluster di densità.

Indice Davies-Bouldin

L'**indice Davies-Bouldin** è definito come la misura di somiglianza media di ciascun cluster con il suo cluster più simile. La somiglianza è il rapporto tra le distanze all'interno del cluster e le distanze tra i cluster. In questo modo, cluster più distanti e meno dispersi porteranno a un punteggio migliore.

Il punteggio minimo è zero e, a differenza della maggior parte delle metriche delle prestazioni, i valori più bassi sono le migliori prestazioni di clustering.

Analogamente al Silhouette Score, il DB Index non richiede la conoscenza a priori delle etichette di verità fondamentale, ma ha un'implementazione più semplice in termini di formulazione rispetto al Silhouette Score.

L'indice è generalmente più alto per i cluster convessi rispetto ad altri concetti di cluster, come i cluster basati sulla densità come quelli ottenuti da DBSCAN.

Indice Davies-Bouldin

E' definito come la similarità media tra ogni cluster C_i, per $i = 1, \ldots, k$, ed il cluster più simile ad esso C_j:

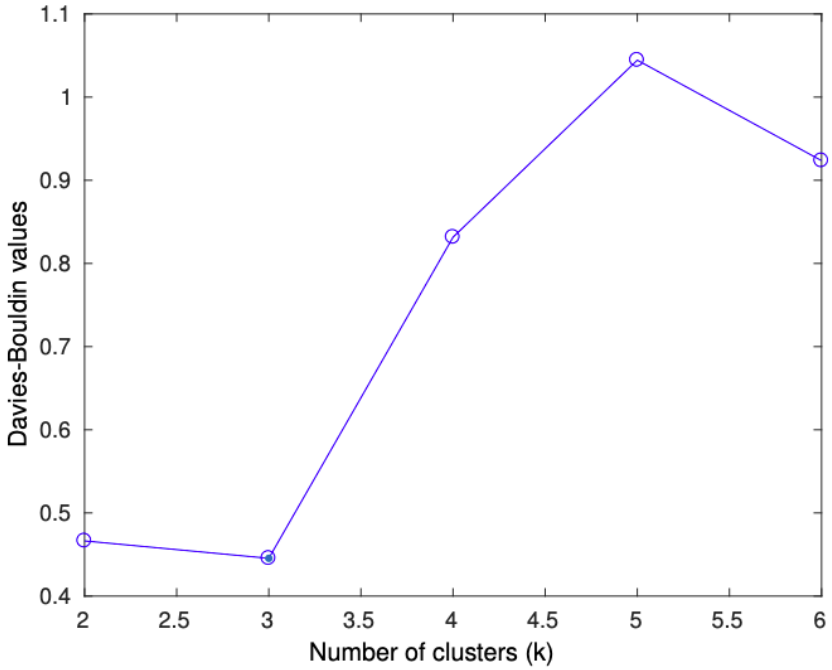
$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

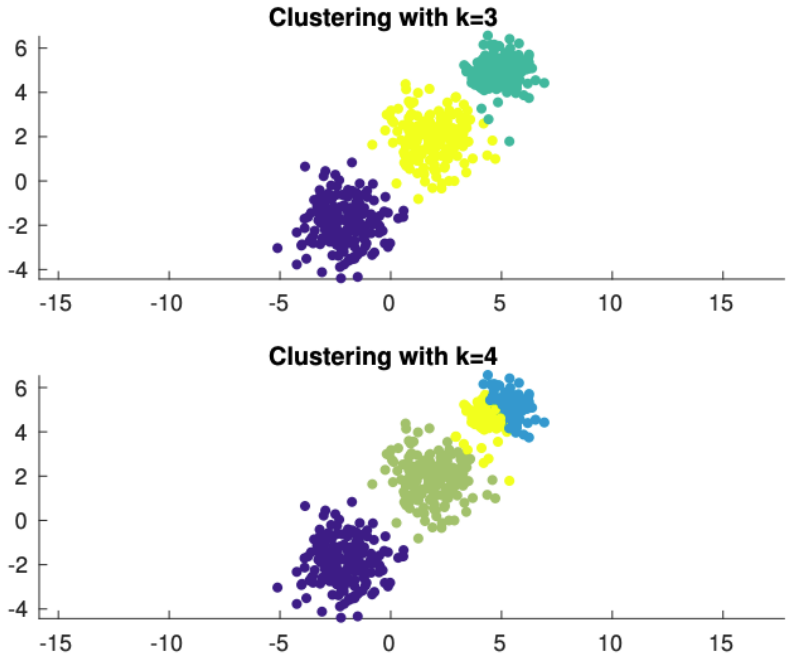
con

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

essendo s_i la distanza media tra ogni punto del cluster i ed il suo centroide (ovvero il diametro del cluter), e d_{ij} la distanza tra i centroidi dei cluster i e j.

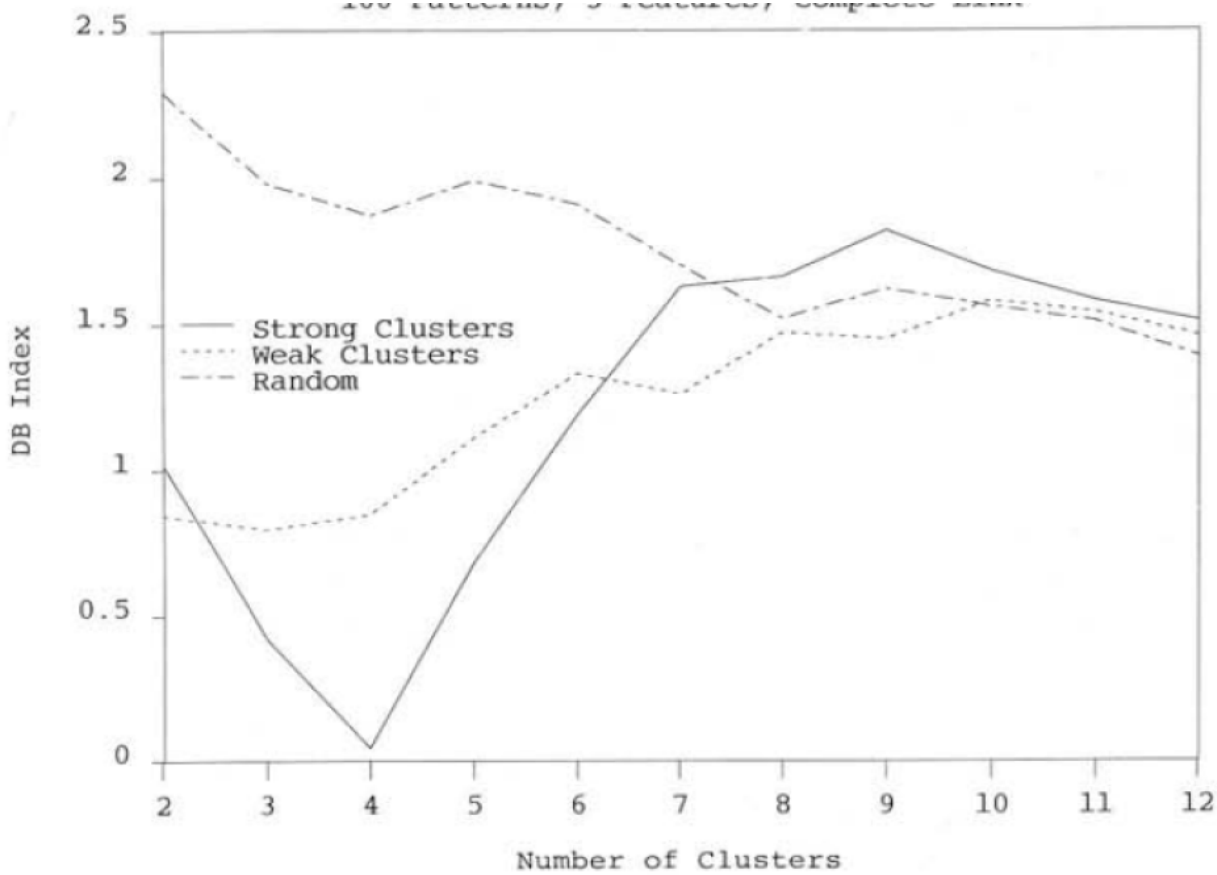
Indice Davies-Bouldin





Indice Davies-Bouldin

Può anche essere utilizzato per determinare la presenza di una struttura di clustering.

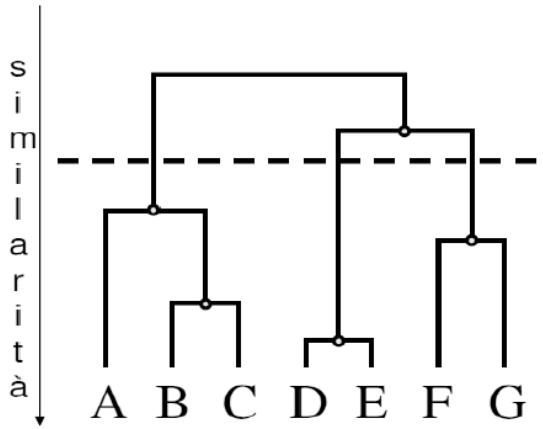


Criteri interni per gerarchie

Rispondono alle seguenti domande:

- Una gerarchia fitta bene i dati su cui è stata calcolata?
- Ci si può fidare di un determinato risultato di clustering gerarchico?

Un esempio: CPCC (**Cophenetic correlation coefficient**):
cophenetic distance: il livello di un dendrogramma dove due oggetti sono stati messi nello stesso cluster per la prima volta



d(D,A) = 6

d(D,E) = 1

7

Criteri interni per gerarchie

La **cophenetic distance** misura quando sono simili due oggetti "dato l'albero" (cioè la misura di distanza espressa dall'albero)

$$CPCC = \frac{\frac{1}{M}\sum_{i,j} d(i,j)d_c(i,j) - m_d m_c}{\sqrt{\frac{1}{M}\sum_{i,j} d^2(i,j) - m_d}\sqrt{\frac{1}{M}\sum_{i,j} d_C^2(i,j) - m_C}} \text{ con } 1 \leq i < j \leq n$$

$m_D = \frac{1}{M}\sum_{i,j} d(i,j)$ misura la correlazione tra la distanza derivante dai dati e la distanza derivante dal dendrogramma che spiega i dati.

$m_C = \frac{1}{M}\sum_{i,j} d_c(i,j)$: CPCC varia tra -1 e 1: più è vicino a 1 migliore è il clustering.

Misure esterne o supervisionate

Classification-based

- Misure simili a quelle usate per valutare i classificatori sulla base della capacità di **riconoscere correttamente** l'appartenenza di un **test item** alla **classe corretta**
- In questo vogliamo misurare la capacità dell'algoritmo di ritrovare le classi presenti nel **test dataset**

Come si fa a valutare?

- si prende un dataset classificato, e quindi partizionato in classi disgiunte
- si ignora l'etichetta classe
- si clusterizza e si valuta il clustering ottenuto

Misure esterne o supervisionate

Usiamo precision, recall, F-measure per validare il risultato del clustering.

L'usiamo per validare un singolo cluster i rispetto alla classe j:

- m è il numero totale di elementi da clusterizzare
- m_i è il numero di elementi del cluster i
- m_j è il numero di elementi della classe j
- m_{ij} è il numero di elementi del cluster i appartenenti alla classe j

- **precision** $p_{ij} = m_{ij}/m_i$
- **recall** $r_{ij} = m_{ij}/m_j$

Come estendiamo ad un **clustering completo**?

- per ogni cluster scegliamo la massima precision, o la massima recall
- sommiamo su tutti i cluster (precision, recall, F-measure)

Misure esterne o supervisionate

Purezza di un clustering:
stesso concetto della precisione ($p_{ij} = m_{ij}/m_i$), ovvero probabilità che un membro del cluster i appartenga alla classe j.

- purezza del **singolo cluster** (uguale a 1 se tutti appartengono ad una sola classe): $p_i = \max_j p_{ij}$
- purezza del **clustering completo**: $p = \sum_{i=1}^{K} \frac{m_i}{m} p_i$

Entropia di un clustering:
basato ancora sulla probabilità che un membro del cluster i appartenga alla classe j ($p_{ij} = m_{ij}/m_i$)

- entropia del **singolo cluster** (uguale a 0 se tutti appartengono ad una sola classe) : $e_i = \sum_{j=1}^{L} p_{ij} log(p_{ij})$
- entropia del **clustering completo**: $e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$

Misure esterne: matrici di incidenza

Matrice di **Incidenza per il clustering** $(n \times n)$:

- Entry (i,j) $= 1$: i due oggetti appartengono allo stesso cluster
- Entry (i,j) $= 0$: i due oggetti appartengono a cluster differenti

Matrice di **Incidenza per le classi note** $(n \times n)$

- Entry (i,j) $= 1$: i due oggetti appartengono alla stessa classe
- Entry (i,j) $= 0$: i due oggetti appartengono a classi differenti

Possiamo calcolarne la **correlazione**, oppure
misurare vicinanza tramite una misura di **similarità per dati binari**

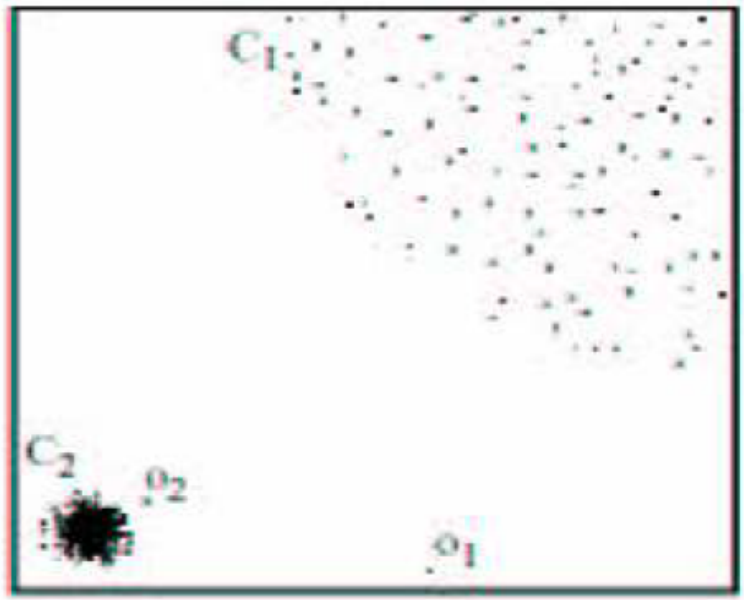
- f_{01} = numero di coppie i,j con classe differente e cluster uguale
- f_{10} = numero di coppie i,j con classe uguale e cluster differente
- f_{11} = numero di coppie i,j con classe uguale e cluster uguale

$$jaccard\_sim = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Identificare gli outlier

Che sono gli outlier?

- Oggetti che sono molto differenti (hanno un comportamento non omogeneo) da tutti gli altri dati.
- Esempi:
  - Utenti di carte di credito: frode bancarie
  - Analisi mediche: risposte non comuni a trattamenti medici



Un metodo basato sulla distanza non individuerebbe o_2 come outlier.

Identificare gli outlier

Local outlier factor (LOF):

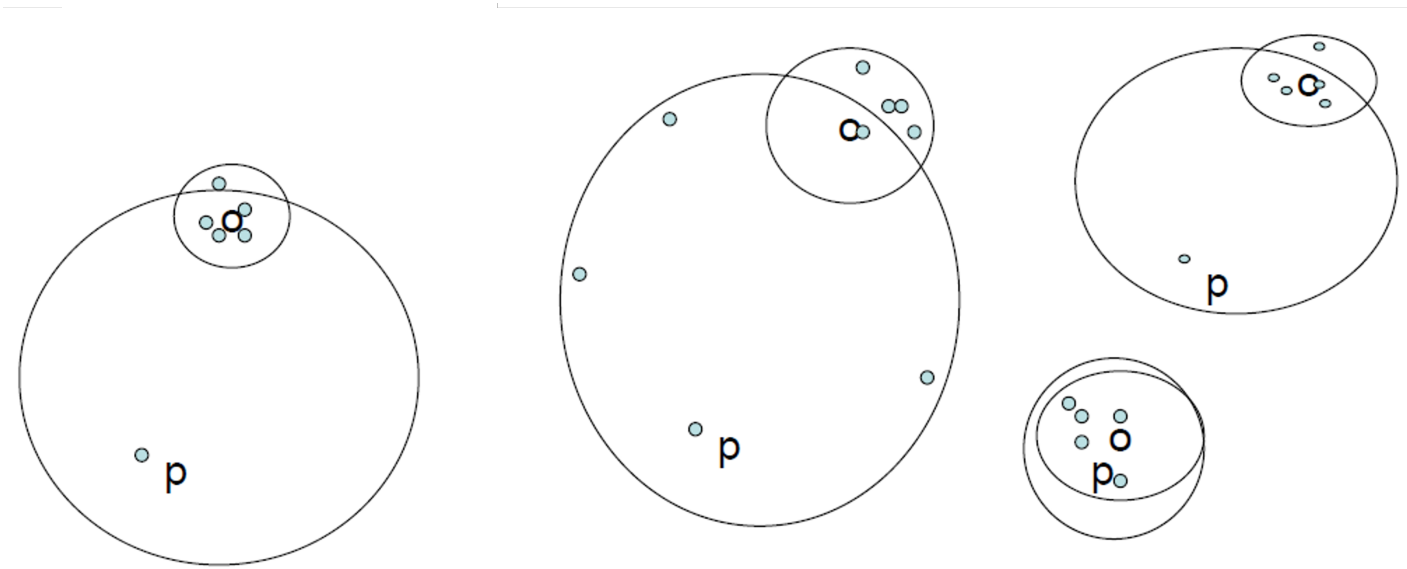
- Essere un outlier in base al comportamento (densità) del suo vicinato (intorno).
- Definire un grado di outlierness cioe' LOCAL OUTLIER FACTOR (quanto isolato e' un oggetto confronto al suo intorno)

K-distance(p) di un oggetto p è la massima distanza di p dai suoi k più vicini punti (K per i metodi quali DBSCAN e' il *MinPts*, ovvero minimo numero di punti da usare per definire un cluster).

K-distance intorno $N_{\text{k-distance}(p)}(p)$ sono I k punti la cui distanza è al massimo k-distance da p.

Identificare gli outlier

Reachability distance reach-dist(p, o) (con o un punto dell'intorno di p con raggio MinPts-distance) $= max\{d(p, o), \text{MinPts-distance}(o)\}$



Identificare gli outlier

Local reachability distance:
$lrd_{MinPts}(p) = |N_{MinPts}(p)| / \sum_{o \in N_{MinPts}(p)} \text{reach-dist}(p, o)$.

Local outlier factor:
$LOF_{MinPts}(p) = (\sum_{o \in N_{MinPts}}(p) / lrd_{MinPts}(o) / lrd_{MinPts}(p)) / |N_{MinPts}(p)|$.

$LOF = 1$ allora p non è un outlier.
Per valori grandi di LOF p è un outlier perchè distanza tra p e o è grande e sarà al numeratore nella sommatoria mentre al denominatore avremo le MinPts-distance di o le quali saranno piccole.