UNIVERSITÀ
DI TRENTO

Tecniche di clusterizzazione non supervisionata e clusterizzazione di documenti testuali

Daniele Capelli

Supervisore: Luigi Amedeo Bianchi

Università di Trento

9 Maggio 2024

Indice

Contesto e presentazione del problema

Tecniche supervisionate e non supervisionate

Tecniche di apprendimento supervisionato

- Addestramento su dati etichettati, con una coppia input-output
- Obiettivo: creare previsioni accurate su nuovi dati
- Esempi: regressione logistica, alberi di classificazione

Tecniche di apprendimento non supervisionato

- Addestramento su dati non etichettati
- Obiettivo: trovare relazioni e strutture nascoste
- Esempi: clustering, PCA

Clustering

La clusterizzazione (o clustering) consiste nella divisione di un dataset in sottogruppi, detti **cluster**, senza avere informazioni a priori sulla loro forma.

- Metodo di classificazione non supervisionato
- Utilizzato in vari contesti: ricerca clinica, marketing...

Misure di dissimilarità

Matrice di dissimilarità

Per dividere un dataset abbiamo bisogno di stabilire il livello di differenza tra gruppi. Questo si può calcolare combinando la differenza tra due campioni, che viene riassunta nella matrice di dissimilarità. Questa matrice presenta le seguenti caratteristiche:

- Matrice $\mathbf{D}$ di dimensione $N \times N$, dove N indica il numero di campioni che abbiamo a disposizione nel dataset
- L'elemento d_{jk} di $\mathbf{D}$ rappresenta la differenza tra il j-mo oggetto ed il k-mo.

Differenza tra oggetti I

Supponiamo di avere un dataset che contiene p attributi su N campioni $x_1, \ldots, x_N$, ovvero per $i = 1, \ldots, N$ abbiamo delle misure x_{ij} per $j = 1, 2, \ldots, p$. Il primo passo per calcolare la differenza tra x_i e x_k consiste nel fornire una differenza a livello del j-mo attributo $d_j(x_{ij}, x_{kj})$: la scelta più comune consiste nel quadrato della distanza

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

Abbiamo delle alternative che si basano sulla tipologia di predittore a disposizione.

Differenza tra oggetti II

- Variabili quantitative: funzione strettamente crescente rispetto al valore assoluto della differenza tra i due valori osservati, oppure correlazione

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$$

- Variabili ordinali: si usano i valori

$$\frac{i - \frac{1}{2}}{M}, \quad i = 1, \ldots, M$$

- Variabili categoriche: si specifica una matrice $M \times M$, il cui elemento di posizione rr' rappresenta il livello di differenza tra le categorie r e r'

Differenza tra oggetti III

La differenza finale tra x_i ed x_k può essere fornita quindi in due modi:

- Sommando le differenze a livello di attributo:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{i'j})$$

- Sommando le differenze pesandole tramite opportuni pesi ω_j:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} \omega_j d_j(x_{ij}, x_{i'j}), \qquad \sum_{j=1}^{p} \omega_j = 1$$

Missing values

A volte capita di ottenere dataset danneggiati in cui alcuni valori relativi ad alcune osservazioni sono mancanti. Ci sono due modi per risolvere il problema dei missing values:

- Omettere le osservazioni $x_{ij}, x_{i'j}$ per le quali almeno uno dei due valori risulta mancante.
- Assegnare ai valori mancanti la media o la mediana dei valori dati, o creare la categoria "missing".

K-Means Clustering

K-means clustering I

- Obiettivo: dividere il dataset in un numero prefissato K di cluster distinti e non sovrapposti. Indichiamo con $C_1, \ldots, C_K$ gli insiemi contenenti gli indici in ciascuno dei cluster.
- Problema da risolvere:

$$\min_{C_1,\ldots,C_K} \{W(C)\}$$

dove $W(C)$ indica la variazione intra-cluster. Possiamo usare

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x_{i'}||^2$$

K-means clustering II

Computazionalmente il problema visto risulta troppo oneroso: vediamone una riscrittura equivalente. In primo luogo se N_k è il numero di indici nel k-mo cluster e $\bar{x}_k$ è il vettore le cui componenti sono date dalla media delle componenti dei campioni nel k-mo cluster, vale che

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2$$

Dato un gruppo di osservazioni S

$$\bar{x}_S = \underset{m}{\text{argmin}} \sum_{i \in S} ||x_i - m||^2$$

K-means clustering III

Il problema è quindi equivalente a

$$\min_{C_1,\ldots,C_K,m_1,\ldots,m_K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2$$

dove i punti $m_1, \ldots, m_K$ sono detti **centroidi**.

Algoritmo per risolvere il K-means clustering

Assegniamo in maniera casuale ogni indice a un cluster. Successivamente si alternano i passi seguenti.

1. Dati i cluster $C_1, \ldots, C_K$ si calcolano i nuovi centroidi $m_1, \ldots, m_K$ come la media dei punti all'interno dei cluster.
2. Dati i centroidi $\{m_1, \ldots, m_K\}$, assegniamo l'osservazione i al cluster il cui centroide risulta più vicino a i stesso, ovvero assegniamo il cluster $C(i)$ dato da

$$C(i) = \underset{1 \leq k \leq K}{\operatorname{argmin}} ||x_i - m_k||^2$$

Criterio di stop: assegnazione invariata.

Scelta del parametro K

A seconda del problema che stiamo studiando, K è un parametro che può essere noto oppure no: nel caso non sia già assegnato, si scelgono una serie di valori $K \in \{1, 2, \ldots, K_{\text{max}}\}$, si esegue K-means clustering in corrispondenza di questi valori e se ne calcolano i corrispondenti valori di perdita $\{W_1, \ldots, W_{k_{\text{max}}}\}$. Si plottano i risultati in un grafico e si sceglie K con il "metodo del gomito".

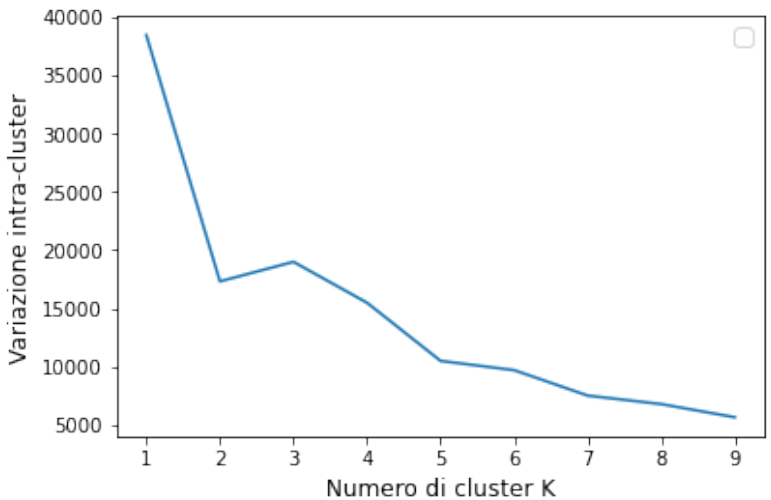


Figura: Esempio del metodo del gomito.

Iris Dataset

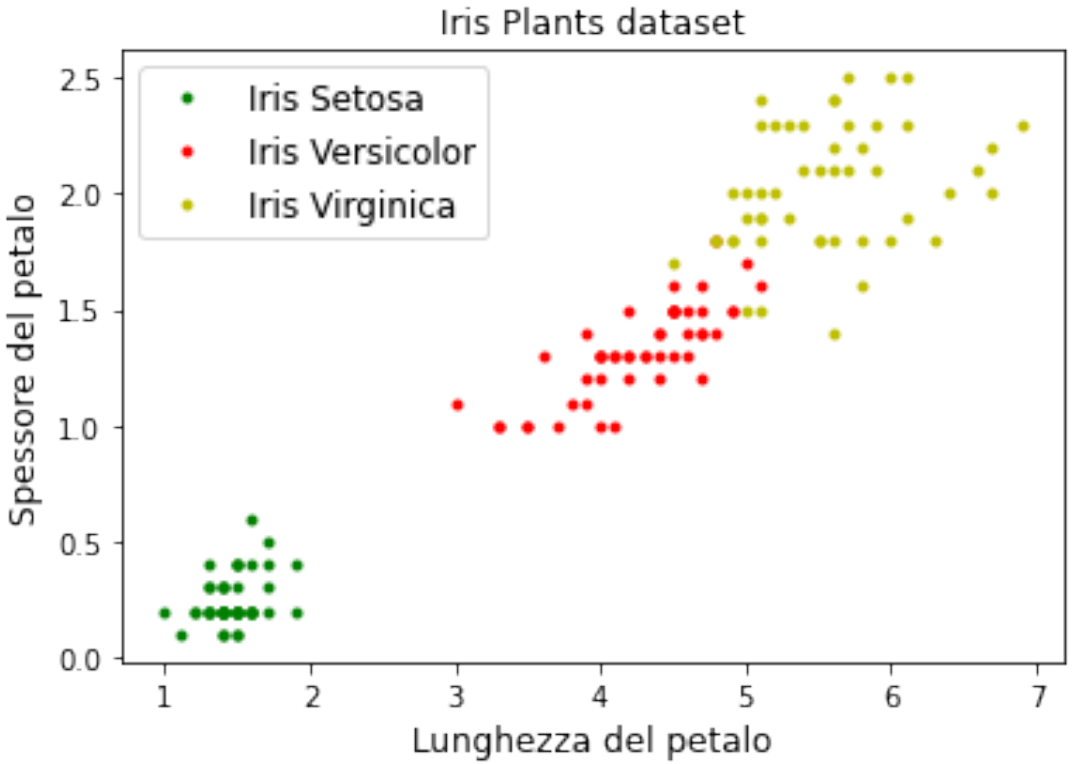
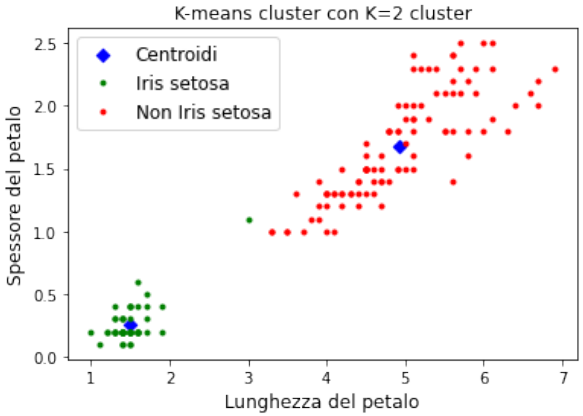
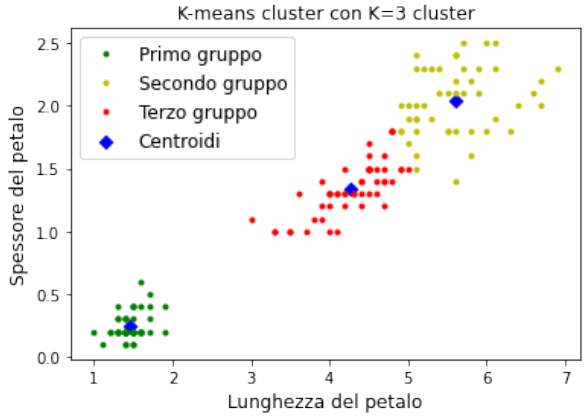


Figura: Iris dataset: tipologia di fiore in corrispondenza di lunghezza e spessore del petalo.

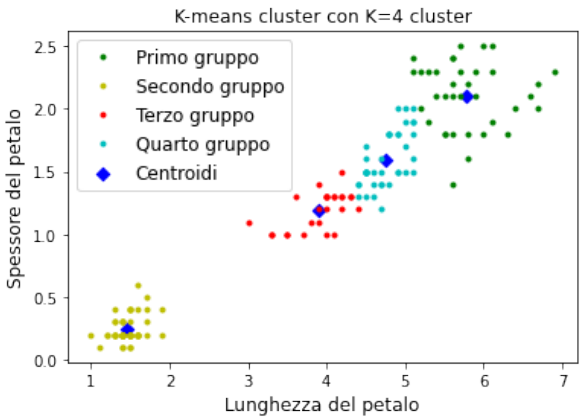
Applicazione di K-Means Clustering a Iris Dataset



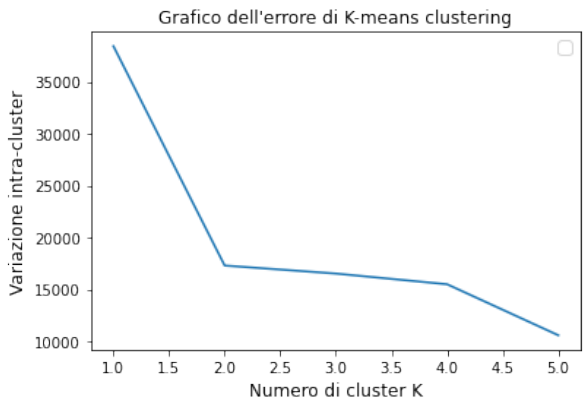
(a) K-means con $K = 2$



(b) K-means con $K = 3$



(c) K-means con $K = 4$



(d) Metodo del gomito

Clustering gerarchico

Clustering gerarchico: setting

Nel clustering gerarchico non si fornisce il numero di cluster da realizzare, ma se ne realizzano diversi tra i quali poter scegliere. Ne vediamo due tipologie:

- Agglomerativo (bottom up): dal basso verso l'alto, in cui ogni osservazione forma inizialmente un singolo cluster
- Divisivo (top-down): dalla cima verso il fondo, in cui ogni osservazione è inizialmente contenuta in un unico grande cluster

Nel seguito utilizzeremo il concetto di diametro di un cluster G, dato da

$$D_G = \max_{i \in G, i' \in G} d_{ii'}$$

Clustering agglomerativo I

Il clustering agglomerativo viene costruito a step:

- Inizialmente, al passo 0, ognuna delle N osservazioni a disposizione viene considerata come parte di un singolo cluster
- Al passo k si uniscono i due cluster del passo $k-1$ che sono più vicini tra loro per andare a ottenere $N-k$ cluster.

Abbiamo bisogno di definire una misura di differenza tra due cluster G ed H: questa si basa sulle possibili differenze $d_{ii'}$ tra un oggetto i in G e uno i' di H. Vediamo tre modi di procedere e i loro svantaggi.

Clustering agglomerativo II

1. Single Linkage nearest-neighbour technique

$$d_{SL}(G, H) = \min_{i \in G, i' \in H} d_{ii'}$$

Risulta molto sensibile alla presenza di punti isolati, inoltre crea la possibilità di collegare elementi che sono vicini tra gruppi diversi ma lontani all'interno del gruppo stesso. Questa misura crea cluster con diametri grandi.

2. Complete Linkage o furthest-neighbor

$$d_{L}(G, H) = \max_{i \in G, i' \in H} d_{ii'}$$

Questa misura si comporta esattamente in maniera opposta al Single Linkage.

Clustering agglomerativo III

3 Group Average

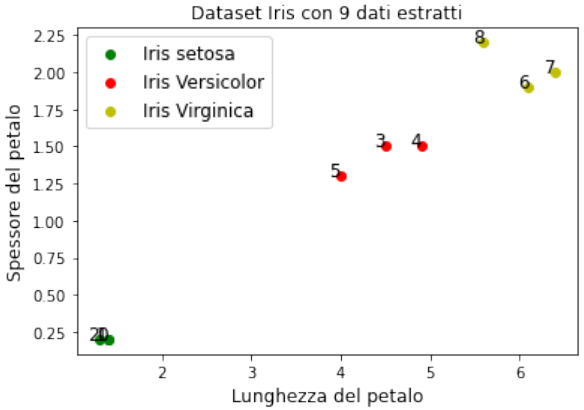
$$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

Questa misura risulta mitigare i problemi di Single e Complete Linkage, ma i risultati ottenuti dipendono fortemente dalla scala numerica utilizzata.

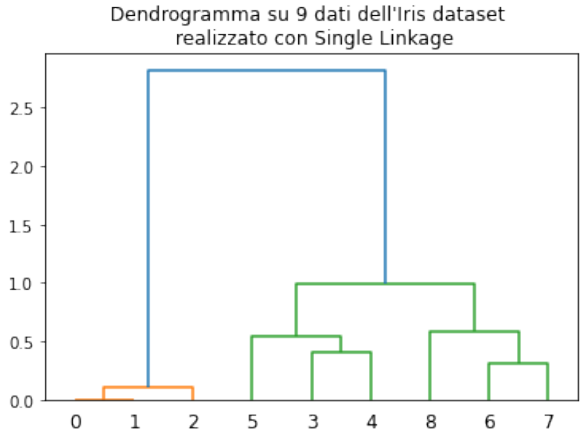
Dendrogrammi

- Per rappresentare il clustering gerarchico si utilizzano i dendrogrammi, che sono alberi binari in cui i nodi rappresentano i cluster e le radici l'intero dataset.
- Il clustering gerarchico gode di una proprietà di monotonia: la differenza tra i cluster che si fondono aumenta all'aumentare del livello di fusione. Questo può essere sfruttato a livello grafico.
- Infine, dato un dendrogramma, la divisione in cluster dell'intero dataset si realizza tagliando il dendrogramma a una determinata altezza.

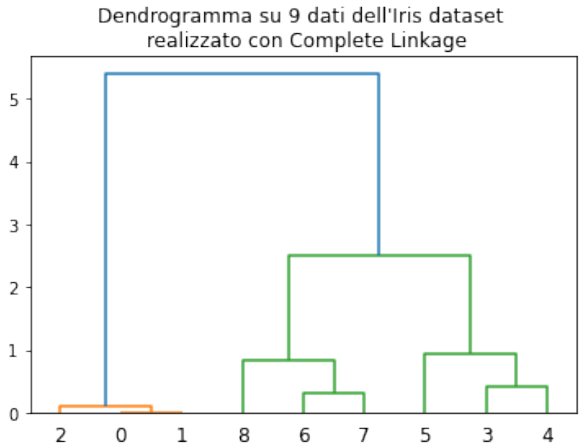
Costruzione di un dendrogramma



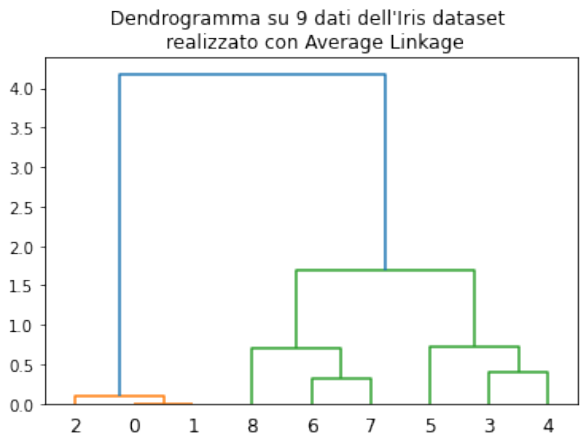
(a) Iris Dataset ristretto



(b) Single Linkage

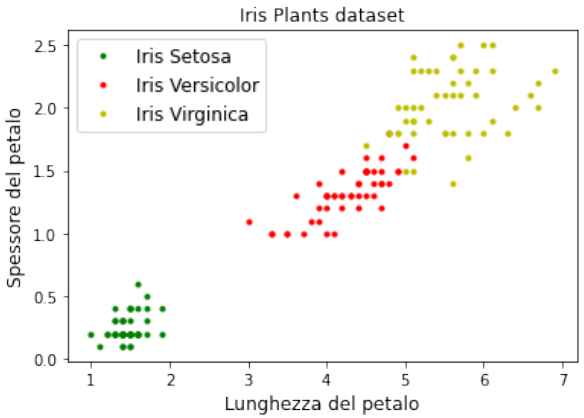


(c) Complete Linkage

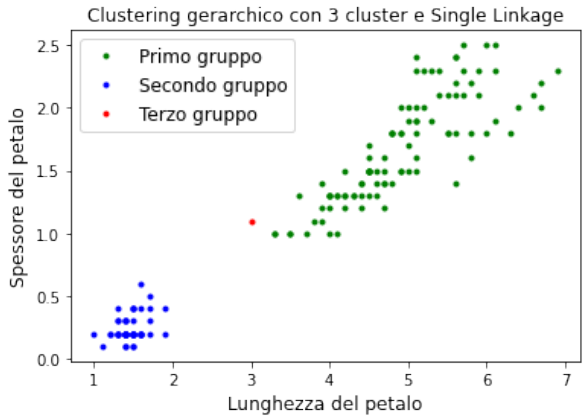


(d) Average Linkage

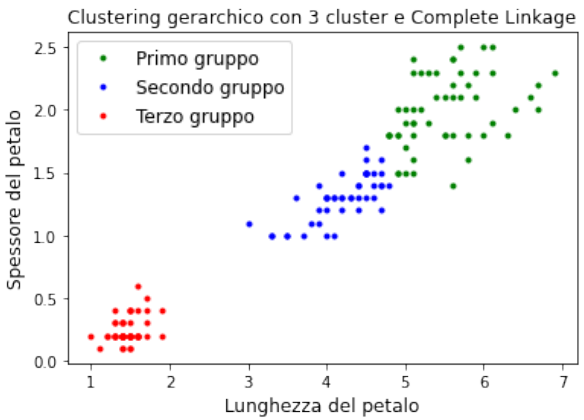
Clustering gerarchico su Iris Dataset



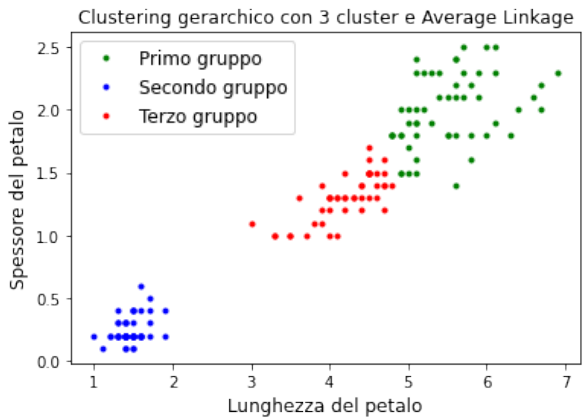
(a) Iris Dataset



(b) Single Linkage

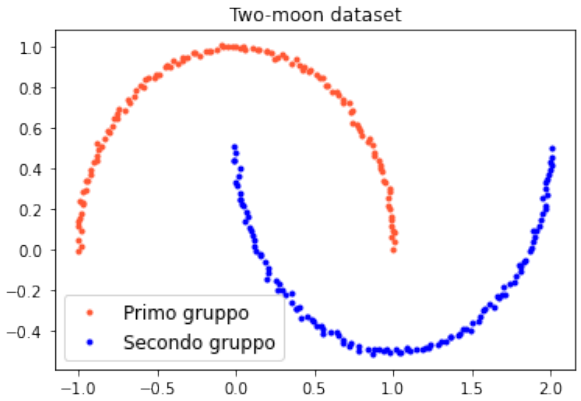


(c) Complete Linkage

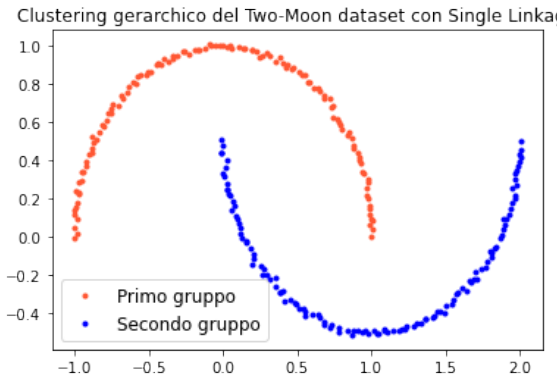


(d) Average Linkage

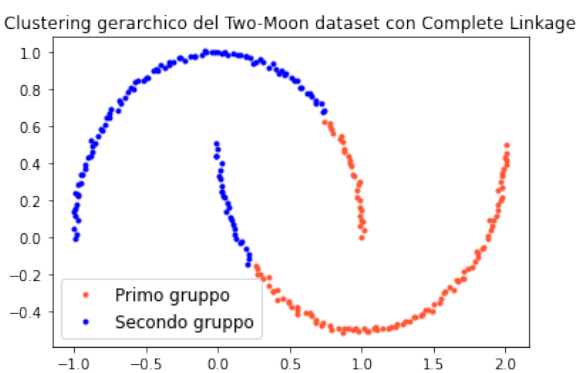
Clustering agglomerativo su Two Moon Dataset



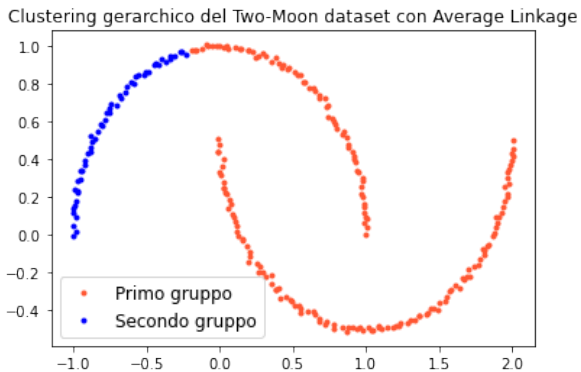
(a) Two Moon Dataset



(b) Single Linkage

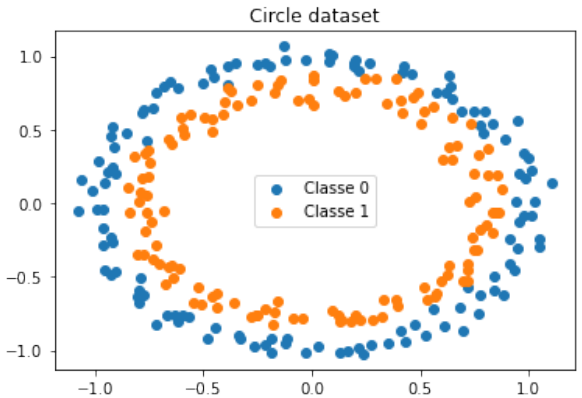


(c) Complete Linkage

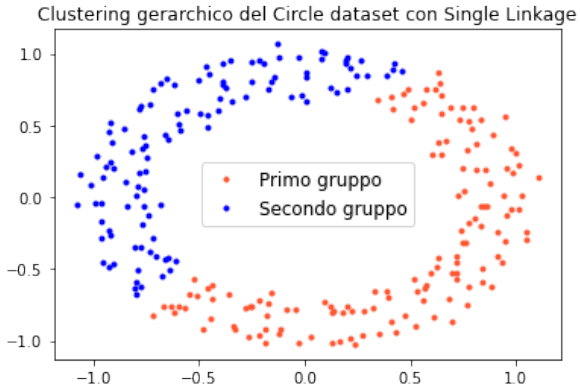


(d) Average Linkage

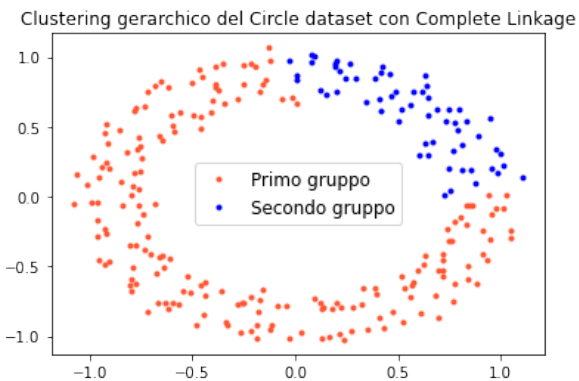
Clustering agglomerativo su Circle Dataset



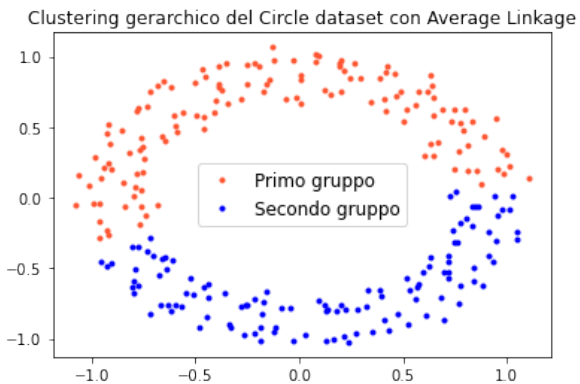
(a) Circle Dataset



(b) Single Linkage



(c) Complete Linkage



(d) Average Linkage

Clustering divisivo

Il clustering divisivo viene realizzato per passi: inizialmente tutti i dati appartengono a un unico cluster. In seguito, al passo k, dividiamo uno dei cluster ottenuti al passo precedente in due sottocluster. Nel dettaglio la divisione avviene tramite la seguente procedura.

- Preso un cluster G si prende l'osservazione con differenza media dalle altre osservazioni massima, che forma il nuovo cluster H
- Ogni osservazione in G la cui distanza media da quelle in H, meno quella dalle altre osservazioni in G è negativa viene trasferita in H

Per scegliere il cluster da dividere si prende quello con diametro maggiore o con dissimilarità media maggiore

$$\bar{d}_G = \frac{1}{N_G} \sum_{i \in G} \sum_{i' \in G} d_{ii'}$$

Clusterizzazione di documenti testuali

Rappresentazione dei dati

Vediamo ora un'applicazione dei metodi studiati tramite la classificazione di documenti testuali. Come primo problema dobbiamo trovare un modo di rappresentare i documenti: si utilizza la rappresentazione **bags of words**.
Dato un dizionario, si ordinano le M parole contenute in esso e, a ogni documento, associamo un vettore di lunghezza M il cui elemento i-mo indica il numero di volte in cui compare la parola i-ma del dizionario.

Accorgimenti sull'uso di bags of words

L'approccio presentato, senza ulteriori specificazioni, presenta come difetto una forte dipendenza del vettore ottenuto dalla lunghezza del documento. Per questo motivo si utilizzano i seguenti accorgimenti.

- Consideriamo al posto della frequenza assoluta la frequenza relativa.
- Eliminiamo quelle parole che compaiono con una frequenza maggiore di una certa soglia (ad esempio articoli e preposizioni).
- Come misura di differenza utilizziamo la metrica del coseno, data da

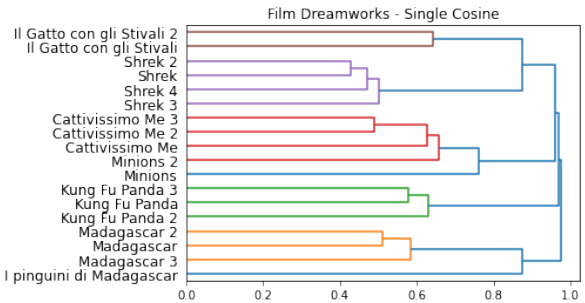
$$S_C(A, B) = \frac{A \cdot B}{||A|| \, ||B||}$$

Film Dreamworks

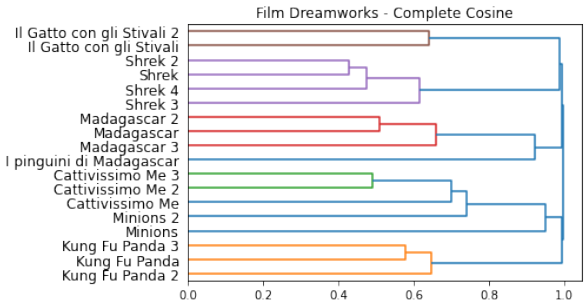
Vediamo come raggruppare una serie di film data la loro trama (in inglese).

- Consideriamo la trama di 18 film Dreamworks: Shrek, Shrek 2, Shrek 3, Shrek 4, Il Gatto con gli stivali, Il gatto con gli stivali 2, Kung fu Panda, Kung fu Panda 2, Kung fu Panda 3, Cattivissimo Me, Cattivissimo Me 2, Cattivissimo Me 3, Minions, Minions 2, Madagascar, Madagascar 2, Madagascar 3 e Madagascar 4.
- Usiamo una rappresentazione bags of words ristretta alle parole che appaiono con una frequenza minore dell'80% nei vari documenti.

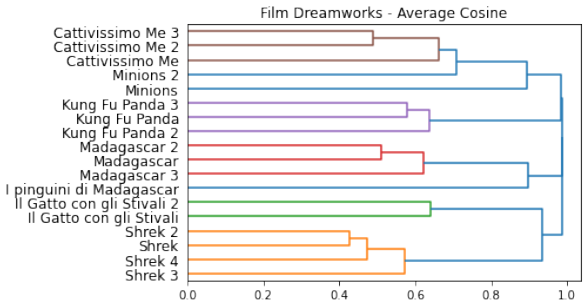
Clustering gerarchico con metrica del coseno



(a) Single Linkage

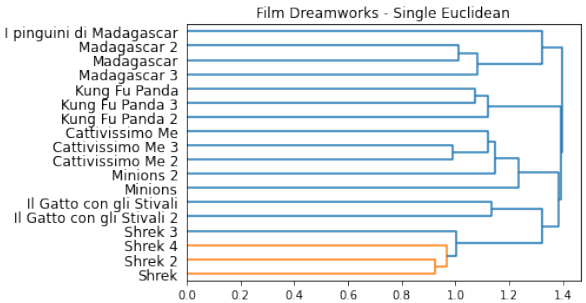


(b) Complete Linkage

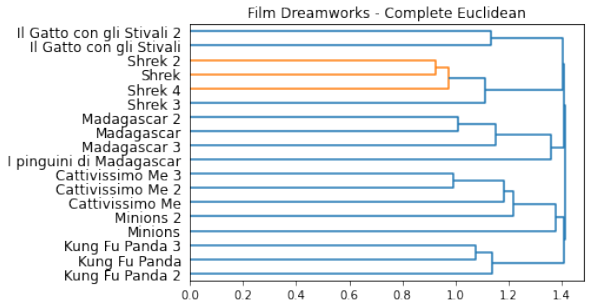


(c) Average Linkage

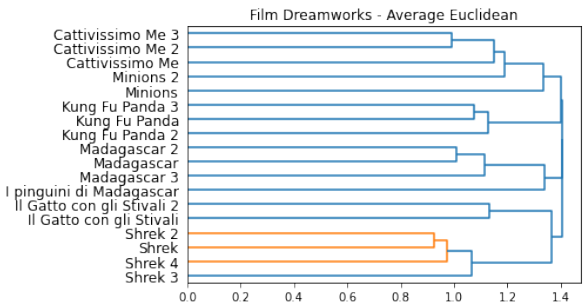
Clustering gerarchico con metrica euclidea



(a) Single Linkage



(b) Complete Linkage



(c) Average Linkage

K-Means clustering

Gruppo	Film
Primo gruppo	Kung fu Panda, Kung fu Panda 2, Kung fu Panda 3
Secondo gruppo	Cattivissimo Me, Cattivissimo Me 2, Cattivissimo Me 3, Minions, Minions 2
Terzo gruppo	Shrek, Shrek 2, Shrek 3, Shrek 4, Il gatto con gli stivali, Il gatto con gli stivali 2
Quarto gruppo	Madagascar, Madagascar 2, Madagascar 3, I pinguini di Madagascar

Tabella: Divisione dei film con K-means clustering e $K = 4$ cluster.

Link al codice

Figura: https://github.com/dcapelli02/Percorso-di-Eccellenza