

Filippo Geraci



DATA MINING ED INFORMATION RETRIEVAL



Definizioni

- **Data mining:** attività di scoperta di informazione latente all'interno di un certo insieme di dati (tipicamente molto grande)
- **Information retrieval (IR):** insieme delle tecnologie utilizzate per reperire una specifica informazione all'interno di un certo insieme di dati (tipicamente molto grande)



Definizioni

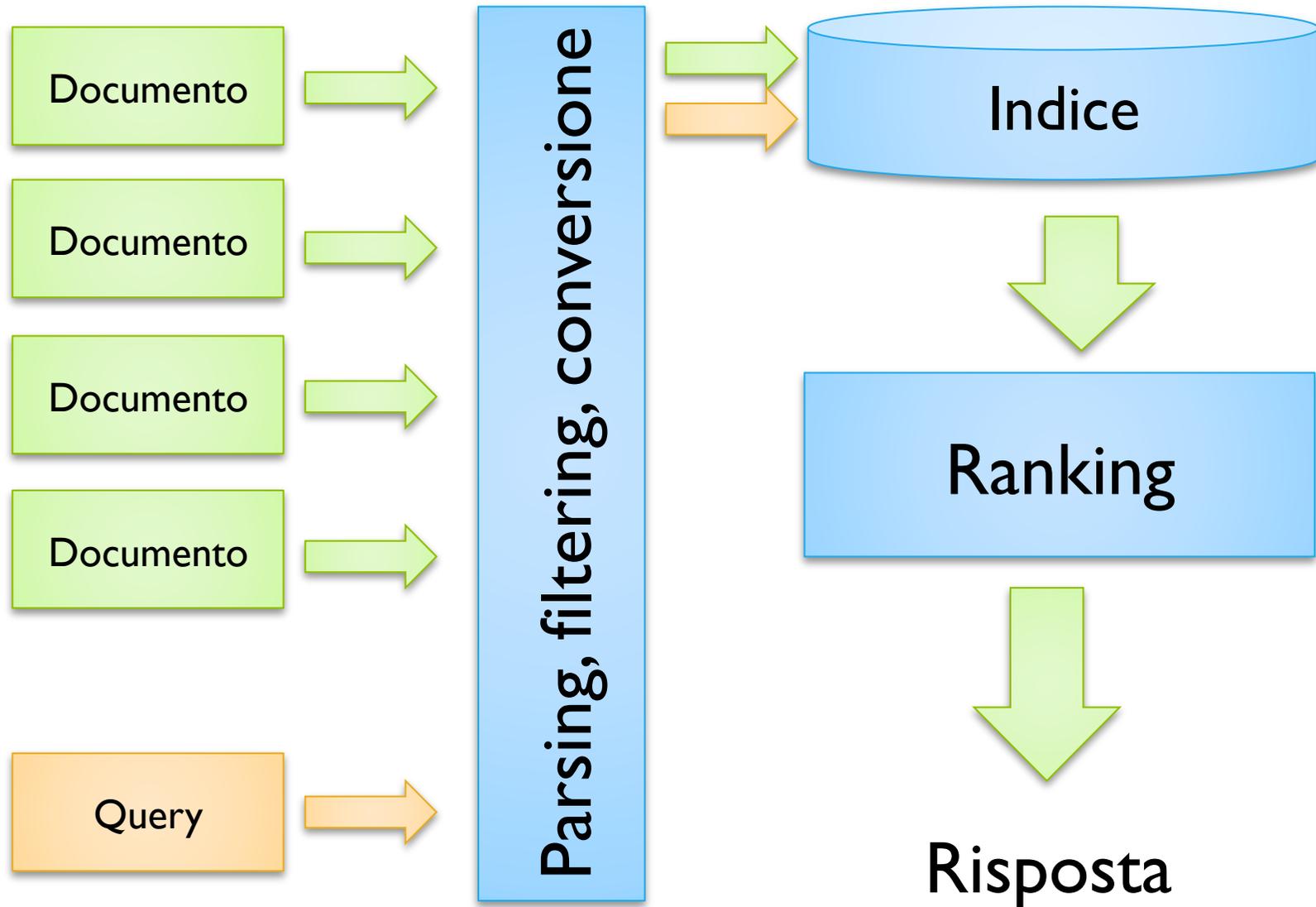
- **Data mining:** attività di **scoperta di informazione latente** all'interno di un certo insieme di dati (tipicamente molto grande)
- **Information retrieval (IR):** insieme delle tecnologie utilizzate **per reperire una specifica informazione** all'interno di un certo insieme di dati (tipicamente molto grande)



Information retrieval (IR)

Parte I

Piattaforma di IR



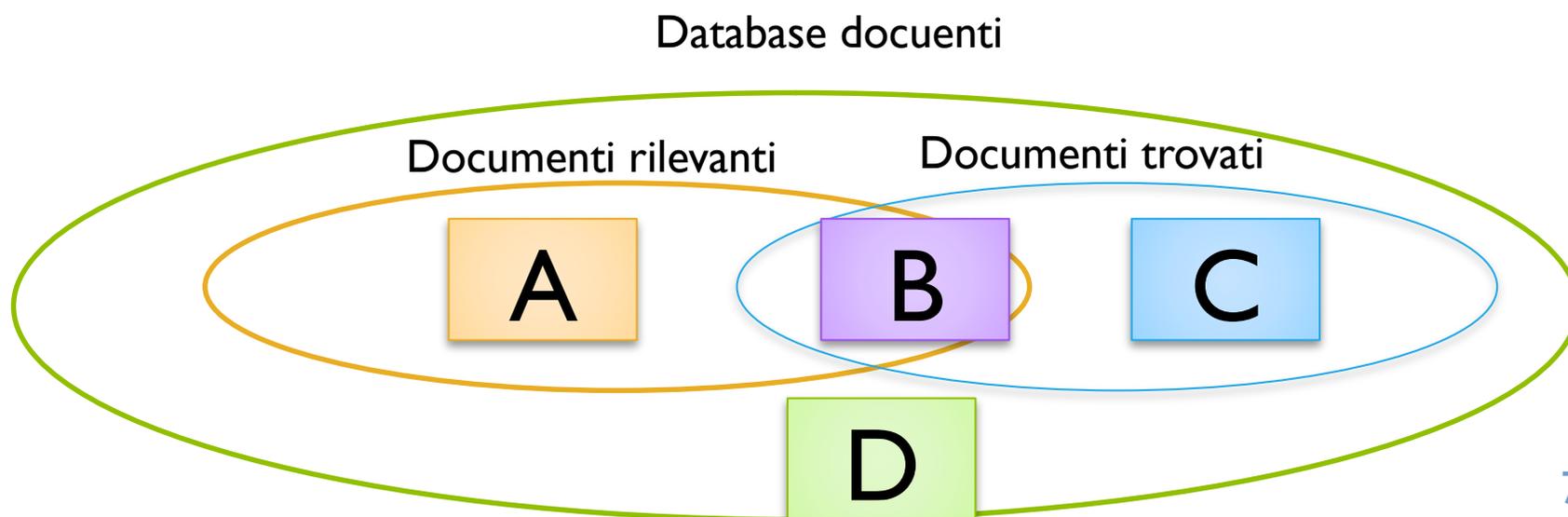


Problemi di progettazione delle piattaforme di IR

- Indicizzazione dei documenti e delle query
 - Qual è il miglior modo di rappresentarle?
- Valutazione della query (processo di retrieval)
 - Qual è il livello di affinità tra query e documento?
- Valutazione del sistema
 - Quant'è efficiente?
 - Quanti dei documenti trovati sono rilevanti? (Precision)
 - Riesce a trovare tutti i documenti rilevanti? (recall)

Precision e recall

- **Precision:** percentuale di documenti rilevanti tra quelli restituiti (B/C)
- **Recall:** percentuale rilevanti restituiti rispetto al totale contenuto nella raccolta (B/A)
- Precision e Recall sono tra loro collegate, per esempio aumentando il numero di documenti ritrovati aumenta la recall e diminuisce la precision.





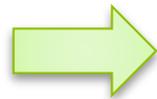
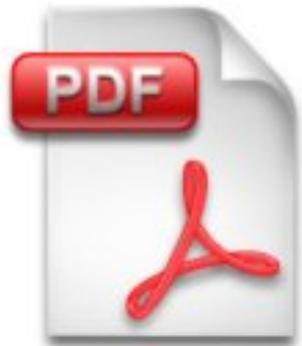
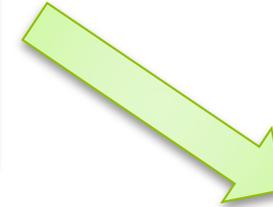
Tipi di documento

- I documenti tipicamente trattati sono di tre tipi:
 - **Testi**, video/immagini, suoni
- Ogni tipo di documento utilizza tecniche diverse
 - In comune rimane lo schema della piattaforma di IR
- Spesso servono molti parser per supportare formati diversi dello stesso tipo di documento

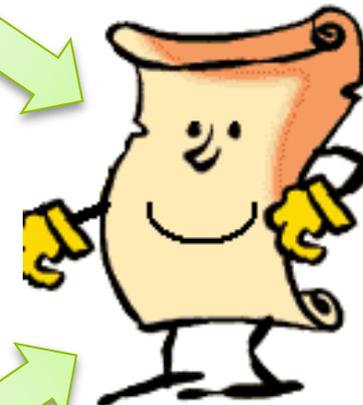
Tipi di documenti



Parser
word



Parser
pdf





Misure di distanza

- In information retrieval si calcola spesso la distanza fra due elementi
 - Maggiore similarità vuol dire minore distanza e viceversa
- Tipicamente gli elementi sono rappresentati da vettori n-dimensionali
- Il passaggio da un documento ad un vettore dipende dal tipo di documento

Correlazione del coseno

- Cosine similarity

$$s(o_a, o_b) = \frac{o_a \cdot o_b}{\|o_a\| \cdot \|o_b\|}$$

- Si trasforma in una distanza

$$d(o_a, o_b) = \sqrt{1 - s^2(o_a, o_b)}$$

Dove o_a, o_b sono due vettori

Coefficiente di Jaccard

- Siano O_a, O_b sono due insiemi di features, il coefficiente di jaccard si definisce come:

$$J(O_a, O_b) = \frac{\#(O_a \cap O_b)}{\#(O_a \cup O_b)}$$

- Per due vettori il coefficiente può essere generalizzato come:

$$GJC(O_a, O_b) = \frac{\min_{i=1}^m (O_{a,i}, O_{b,i})}{\max_{i=1}^m (O_{a,i}, O_{b,i})}$$

Distanza di Minkowski

- Siano O_a, O_b due vettori, la distanza di minkowski è definita come:

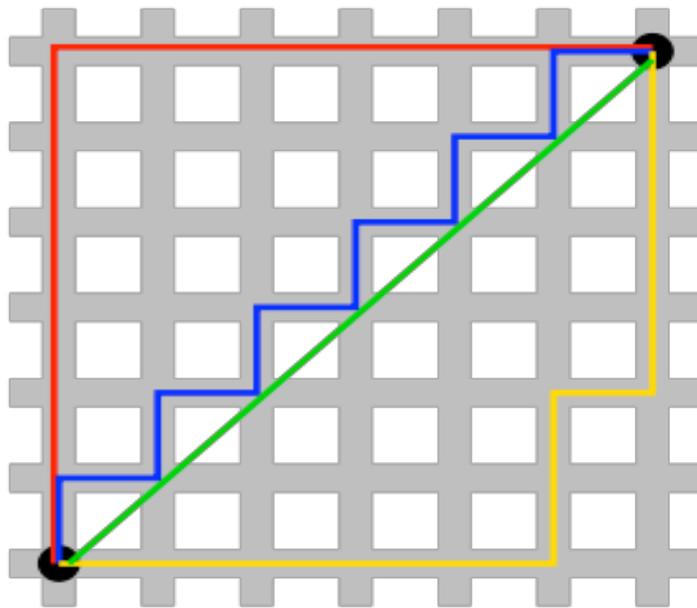
$$L_p(O_a, O_b) = \left(\sum_{i=1}^m |O_{a,i} - O_{b,i}|^p \right)^{1/p}$$

- $p=1$ si riduce alla Manhattan distance
- $p=2$ si riduce alla norma euclidea
- Se p è infinito si ha la norma infinito

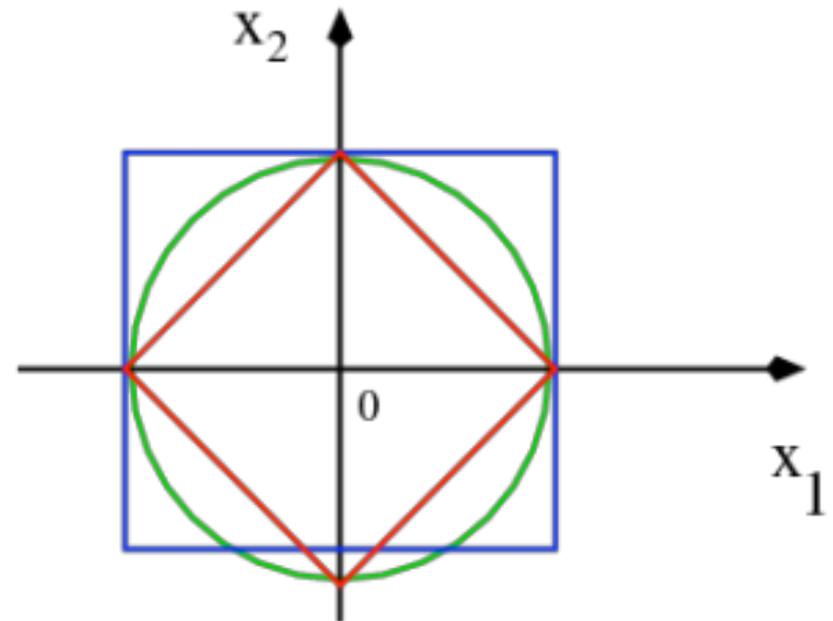
$$L_\infty(O_a, O_b) = \max_{i=1}^m (O_{a,i}, O_{b,i})$$

Distanza di Minkowski

- Le curve in rosso, giallo e blu hanno la stessa Manhattan distance



- Palle unitarie in norma 1 (rosso), norma 2 (verde) e norma infinito (blu)





Text information retrieval



Esempio

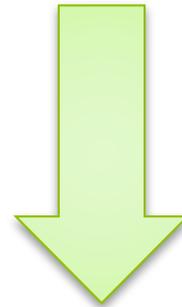
- L'esame di S.I.A è costituito da una prova scritta ed una prova orale. Il voto può variare da 18 a 30.

Trattamento del testo

- Normalizzazione del testo
 - Rimozione della punteggiatura
 - Rimozione numeri e sigle
 - Conversione dei caratteri in minuscolo
 - Ordinamento in ordine lessicografico
 - Rimuovo duplicati, ma ne ricordo la cardinalità
 - Devo ricordare le posizioni nella fase di indicizzazione
- Rimozione delle stop words
 - Vengono rimosse parole poco significative
 - **Esempio:** connettivi, articoli, preposizioni
 - Liste solitamente note a priori, oppure basate su analisi delle frequenze delle parole nel testo

Esempio

- L'esame di S.I.A è costituito da una prova scritta ed una prova orale. Il voto può variare da 18 a 30.



- costituito esame orale prova (2) scritta
variare voto

Stemming

- La stessa parola può apparire in diverse forme pur mantenendo lo stesso significato
- Usato per incrementare l'efficacia in una ricerca e ridurre la dimensione dell'indice
- Stem: porzione della parola rimasta dopo aver rimosso i suoi suffissi o prefissi
- Nasce il problema della correttezza
- Serve un algoritmo per ogni lingua

computer
compute
computes
computing
computed
computation



comput

Stemming - controindicazioni

- Stemming aggressivo

- organization → organ
- army → arm

- Stemming timido

- european → europe
- creation → create

Pesatura dei termini: il modello TF-IDF

- Term frequency: numero di volte in cui un termine compare nel documento

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- Document frequency: numero di documenti in cui compare il termine

INVERSO  $idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$

- Il TF-IDF è un peso associato ad ogni termine in un documento

$$(tf-idf)_{i,j} = tf_{i,j} \times idf_i$$



Pesatura dei termini: il modello TF-IDF

- Uno score alto nel modello TF-IDF si ottiene da un valore alto di term frequency in un dato documento ed un valore basso del document frequency nell'intero dataset.
- Documenti memorizzati come matrice
- Ogni documento è un vettore in uno spazio n-dimensionale



Due aspetti della rappresentazione dei documenti

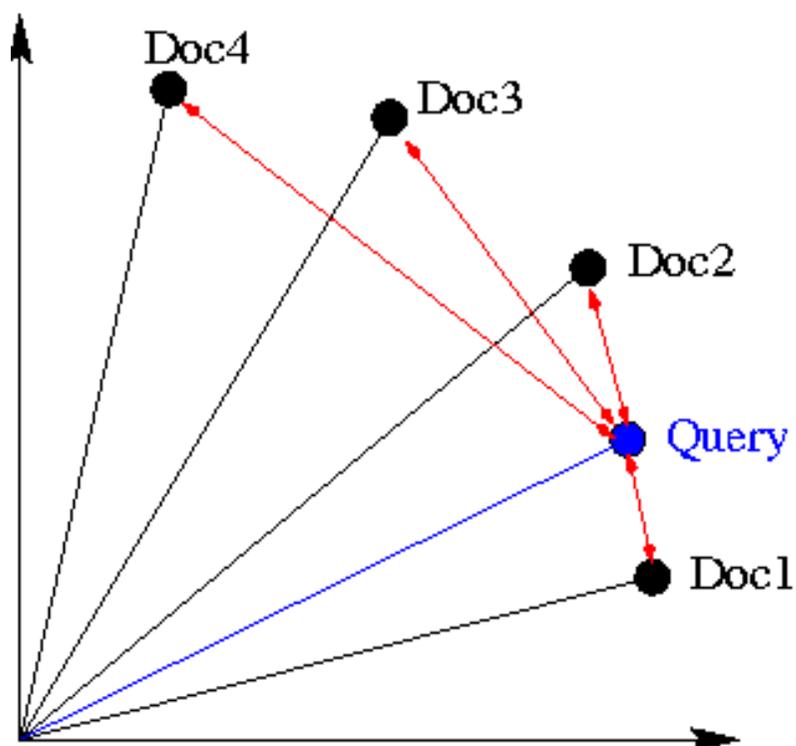
- Descrittività del modello
 - Qual'è il contesto del documento?
 - Importante per recall
- Discriminazione
 - Come distinguo questo documento dagli altri?
 - Importante per precision
- La rimozione di stopwords migliora la discriminazione riducendo il numero di parole in comune tra documenti dissimili
- Lo stemming migliora la descrittività rendendo uguali parole simili
- La migliore rappresentazione dei documenti deve bilanciare questi aspetti

Modelli di retrieval

- Data una query q ed un documento d viene assegnato un punteggio alla rilevanza della coppia (q, d)
- Questo valore è una stima della rilevanza di d per q
- Modelli di retrieval:
 - **Booleano:** documenti e query rappresentati come vettori di booleani
 - **Spazio vettoriale:** documenti e query sono vettori a cui si applica una distanza tra vettori

Spazio vettoriale

- Proiettare i documenti e la query in punti di uno spazio euclideo ad alta dimensione e valutare la distanza



Problema

Vota Antonio, vota Antonio, vota Antonio, vota Antonio, vota Antonio, vota Antonio

- La correlazione tra query e documenti dipende solo dal testo
- Non si considerano i sinonimi



Indici inversi

credevo che
l'azzurro dei
tuoi occhi per
me fosse
sempre cielo

il pomeriggio
è troppo
azzurro e
lungo per me

per sognare
un cielo
azzurro
all'orizzonte
senza
nuvole

1	Azzurro	3	5	5
2	Cielo	11		4
3	Credevo	1		
4	Fosse	9		
5	Lungo		7	
6	Me	8	9	
7	Nuvole			9
8	Occhi	6		
9	Orizzonte			7
10	Per	7	8	1
11	Pomeriggio		2	
12	Sempre	10		
13	Senza			8
14	Sognare			2
15	Troppo		4	
16	Tuoi	5		

cielo azzurro

Cerca

Azzurro	3	5	5
Cielo	11		4



Algoritmi di data mining: Clustering e classificazione

Parte 2

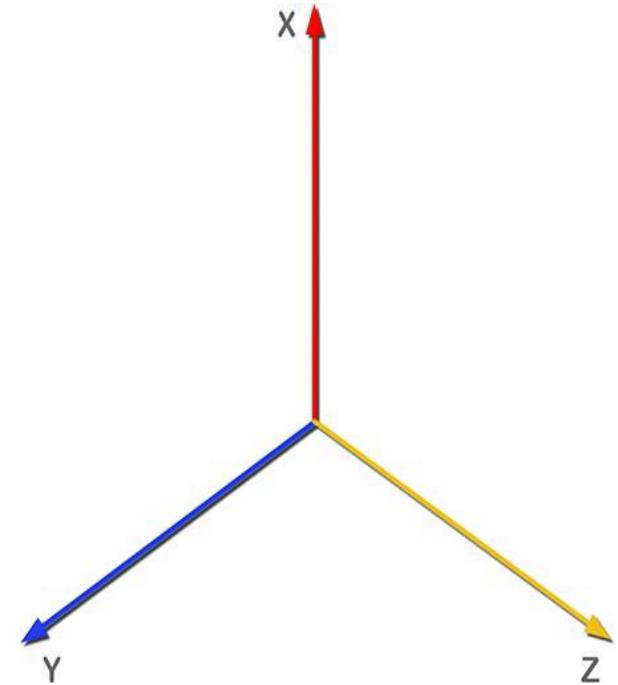
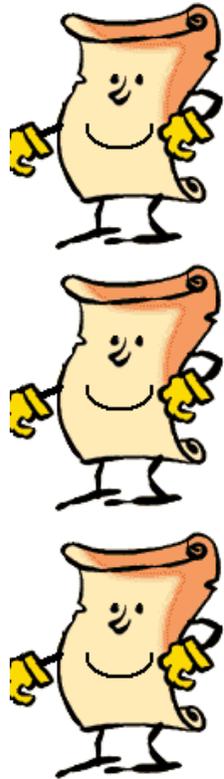
Clustering

- Suddivide un insieme di oggetti in gruppi omogenei
- Procedure indipendenti dal tipo di dati
 - Utilizzabile anche quando non sono disponibili informazioni a-priori sul dominio dei dati;
 - Non possono essere introdotte ottimizzazioni dipendenti dal particolare problema

Cluster hypothesis

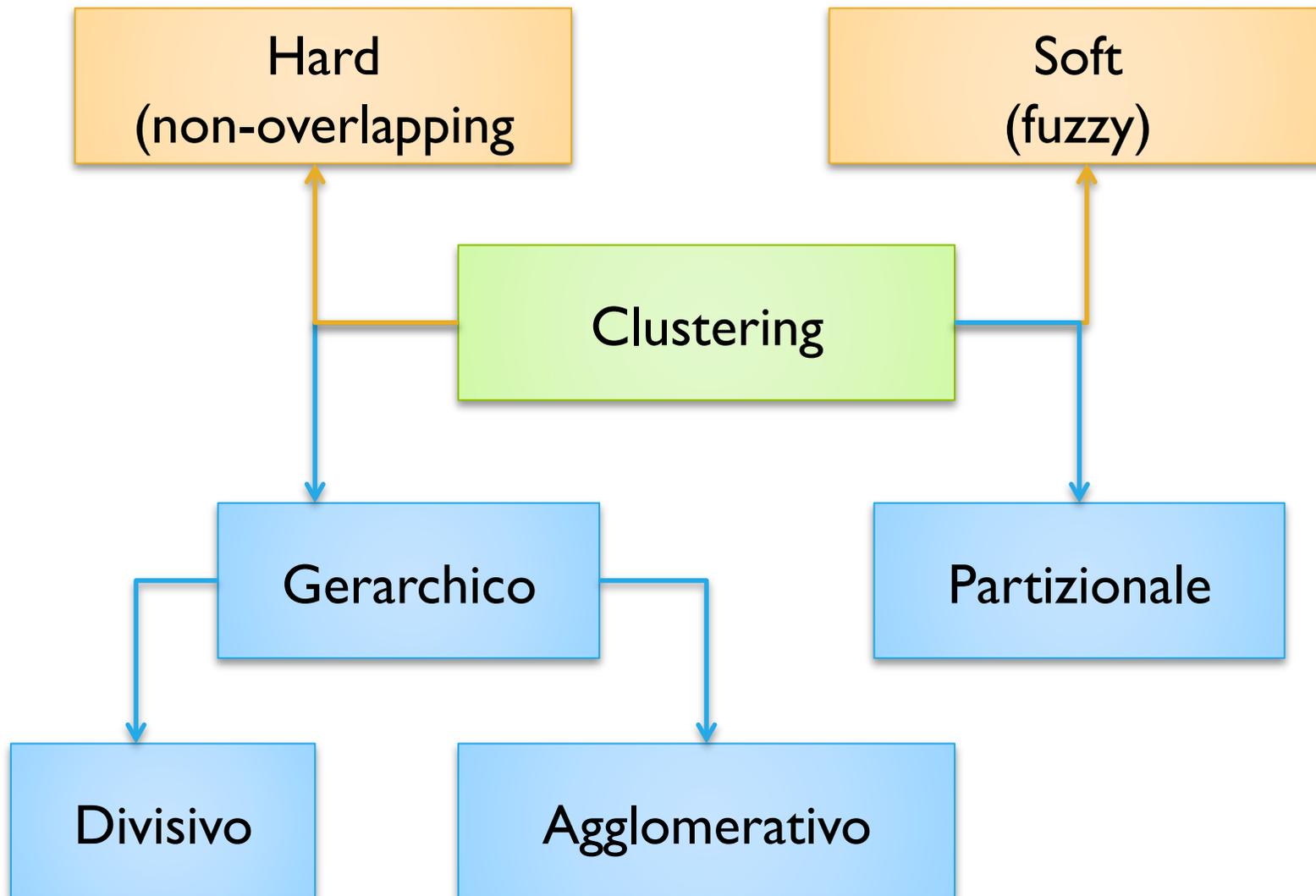
- Se due oggetti sono molto simili ed il primo è anche simile ad un terzo oggetto, molto probabilmente anche tra il secondo ed il terzo oggetto esiste una similarità

Come funziona



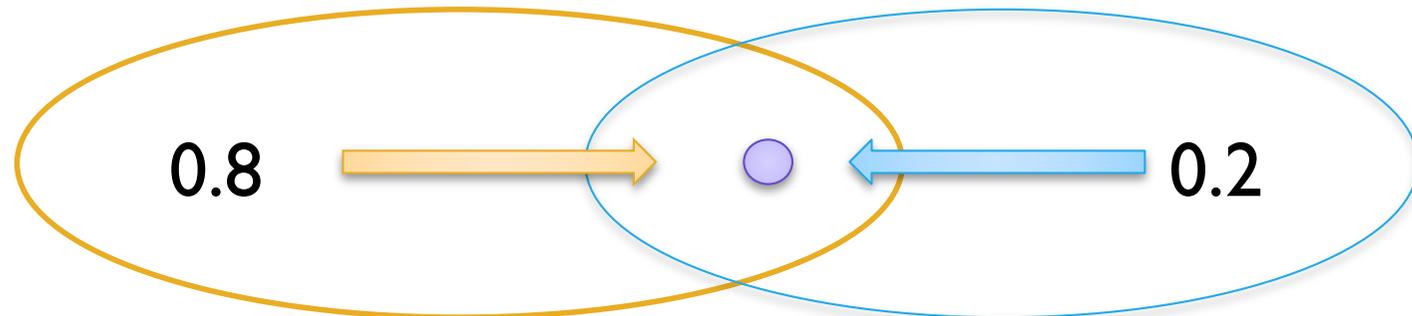
- Tramite tecnologie di information retrieval si trasformano i dati in punti appartenenti ad uno spazio vettoriale (in alcuni casi euclideo) e si lavora su quello
- Indipendente dal tipo di dato

Tassonomia degli algoritmi di clustering



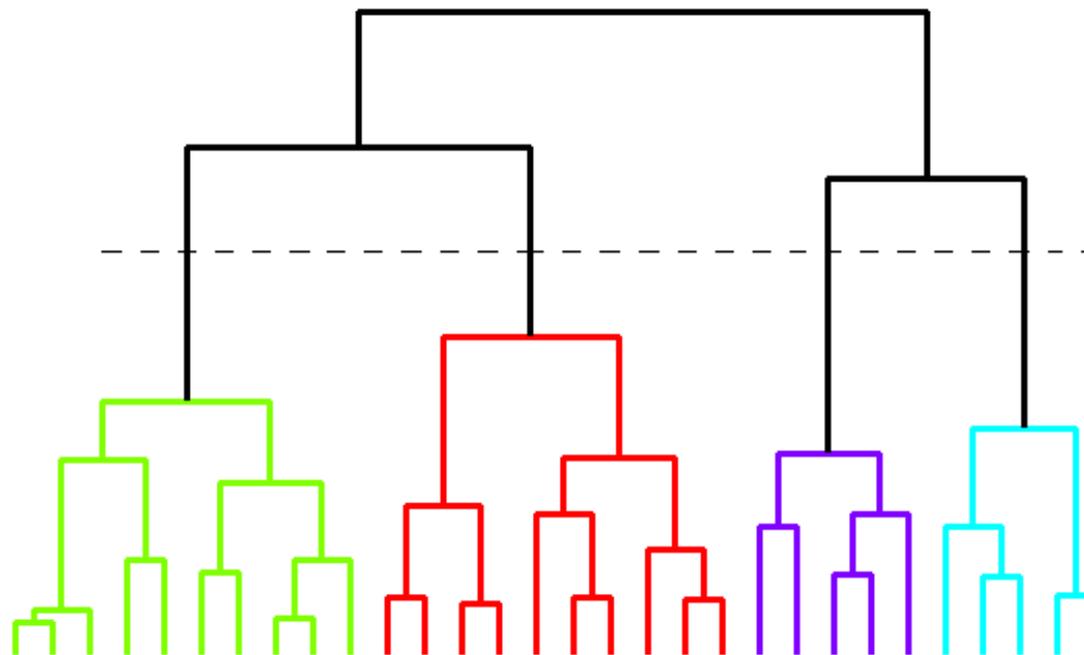
Hard e soft clustering

- Hard clustering: un elemento appartiene solo e soltanto ad un cluster
- Soft clustering: un elemento può appartenere a più cluster
 - Solitamente viene associato un valore di peso (probabilità) di appartenenza dell'elemento ad ogni cluster



Clustering gerarchico

- Richiede l'intera matrice delle distanze
- Produce dendrogramm
- Tagliando l'albero a vari livelli si ottiene un numero diverso di cluster



Clustering gerarchico divisivo

1. Per ogni cluster seleziona una coppia di punti diametrali
2. Se un cluster soddisfa il criterio di partizionamento allora:
 - a) Dividi il cluster usando come centri la coppia selezionata nel punto 1
 - b) Torna al punto 1
3. Se non ci sono più cluster da dividere fermati

Criteri di partizionamento

- Ogni cluster fino a che tutti i cluster non hanno un solo elemento (albero completo)
- Il più grande (albero bilanciato)

Clustering gerarchico agglomerativo

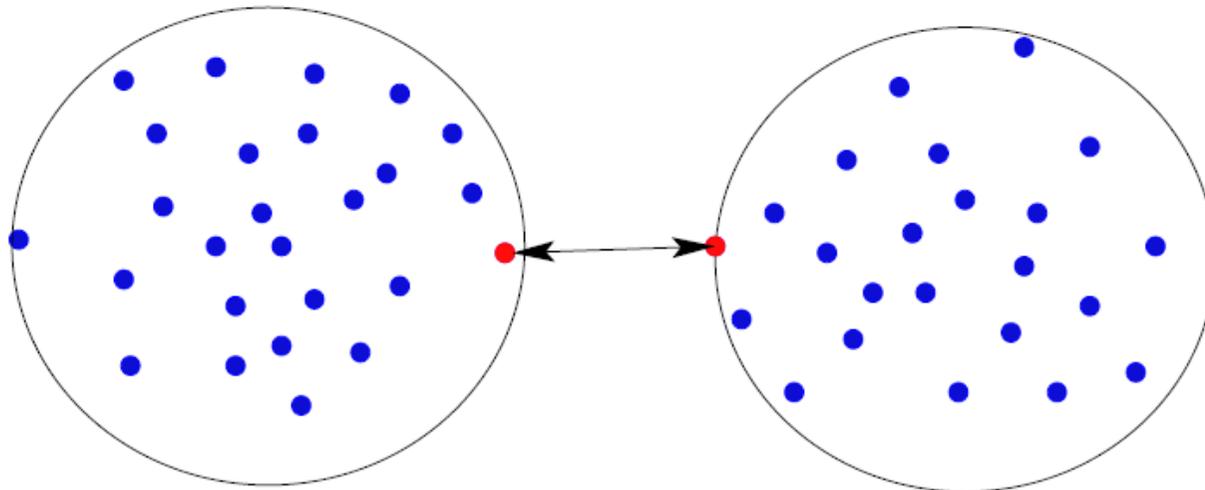
1. All'inizio ad ogni punto si associa un cluster
 2. Ad ogni iterazione
 - a) Costruisci la matrice delle distanze tra cluster
 - b) Unisci i cluster in accordo al criterio di **linkage**
 3. Ricomincia dal punto 2 finché tutti i punti sono in un unico cluster
-

- HAC è l'algoritmo gerarchico più usato
- Computazionalmente costoso
- Produce intero albero

Criteri di linkage: single linkage

$$d(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p, q)$$

- Distanza tra i due elementi più vicini appartenenti a cluster diversi

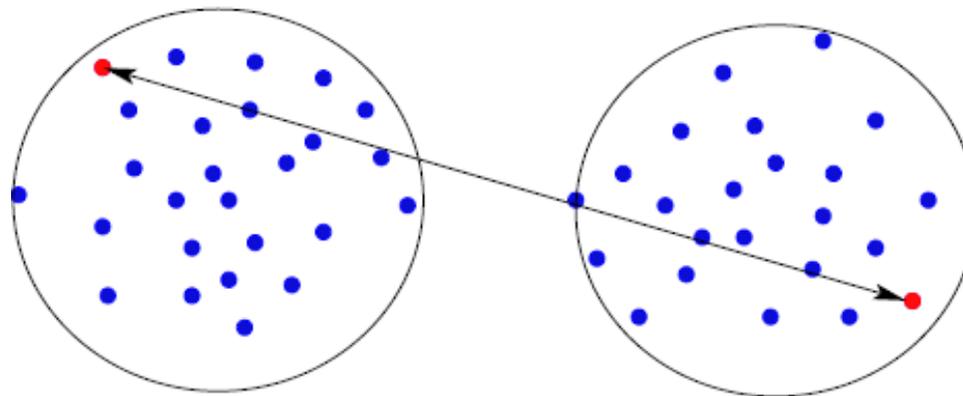


- Forza due cluster ad unirsi solo in base a due elementi non considerando gli altri
 - Crea effetto concatenazione

Criteri di linkage: complete linkage

$$d(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$$

- Distanza tra i due elementi più lontani appartenenti a cluster diversi

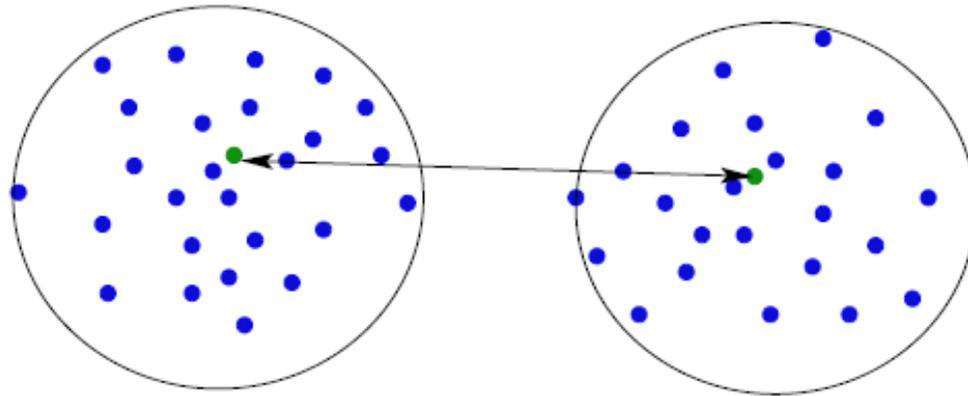


- Crea cluster molto compatti
- Non tollerante a dati contenenti “rumore”

Criteri di linkage: complete linkage

$$d(C_i, C_j) = \frac{1}{|C_i| |C_j|} \sum_{p \in C_i} \sum_{q \in C_j} d(p, q)$$

- Media tra le distanze di tutti gli elementi appartenenti a cluster diversi



- Robusto rispetto a concatenazione e “rumore”
- Computazionalmente costoso

Clustering partitivo

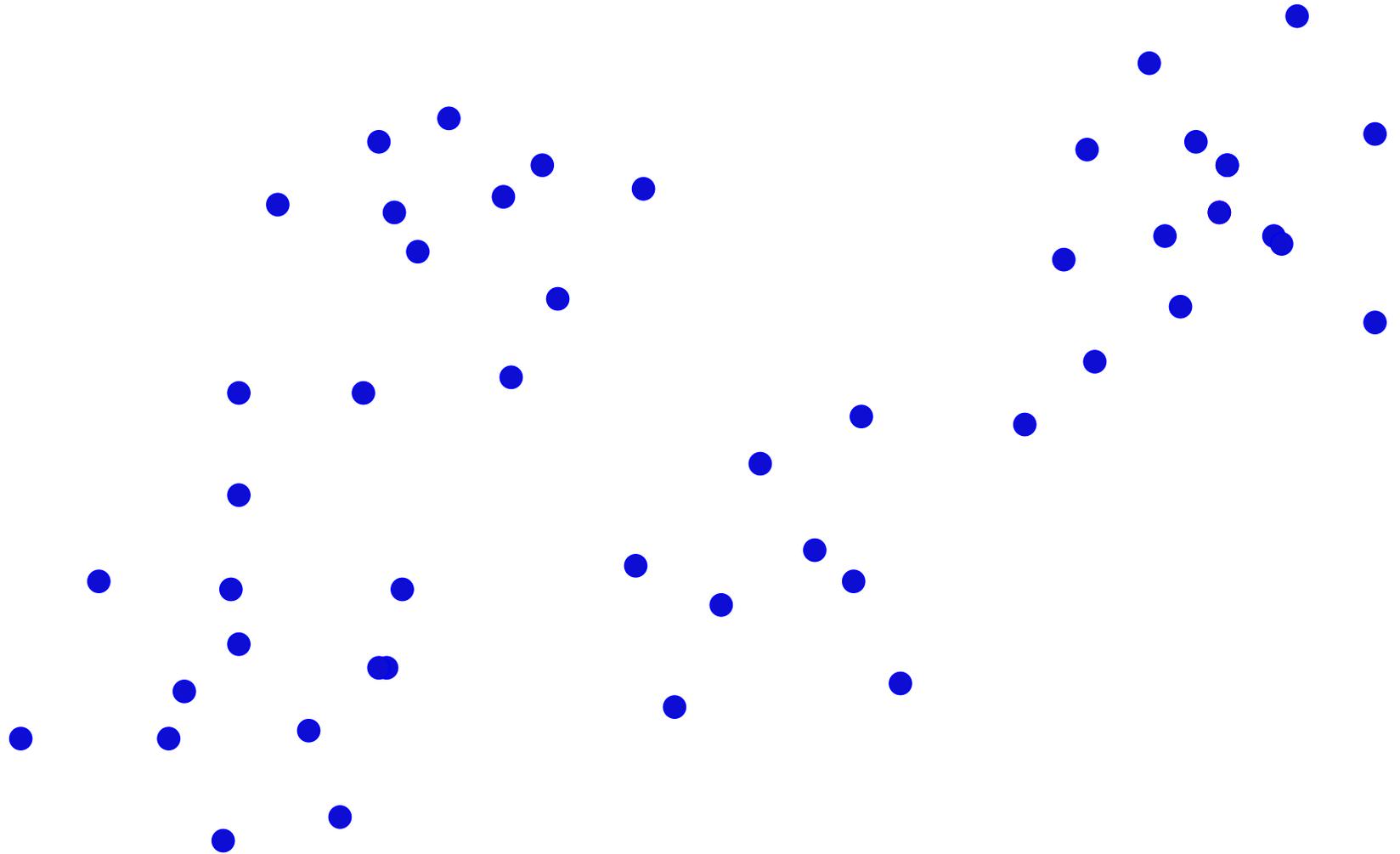
- Algoritmo FPF (furthest point first)
 - Il modo più intuitivo e veloce
 - Minimizza il massimo raggio dei cluster

$$\min_j \max_{x \in C_j} M_v(x, \mu_j)$$

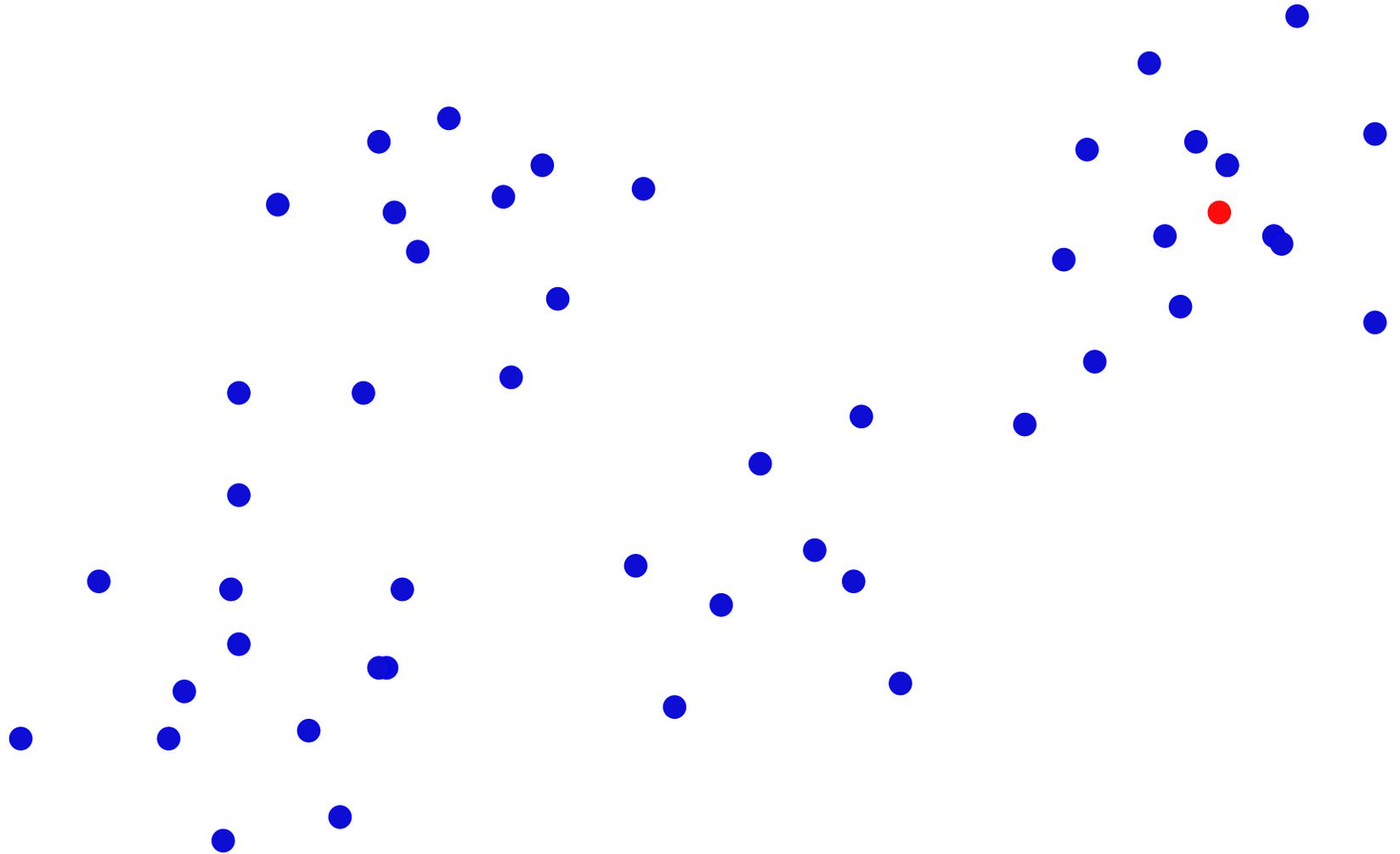
- Algoritmo k-means
 - Il più usato
 - Minimizza la somma dei quadrati delle distanze dei punti dal centroide di riferimento

$$\min_j \sum_{x \in C_j} (M_v(x, \mu_j))^2$$

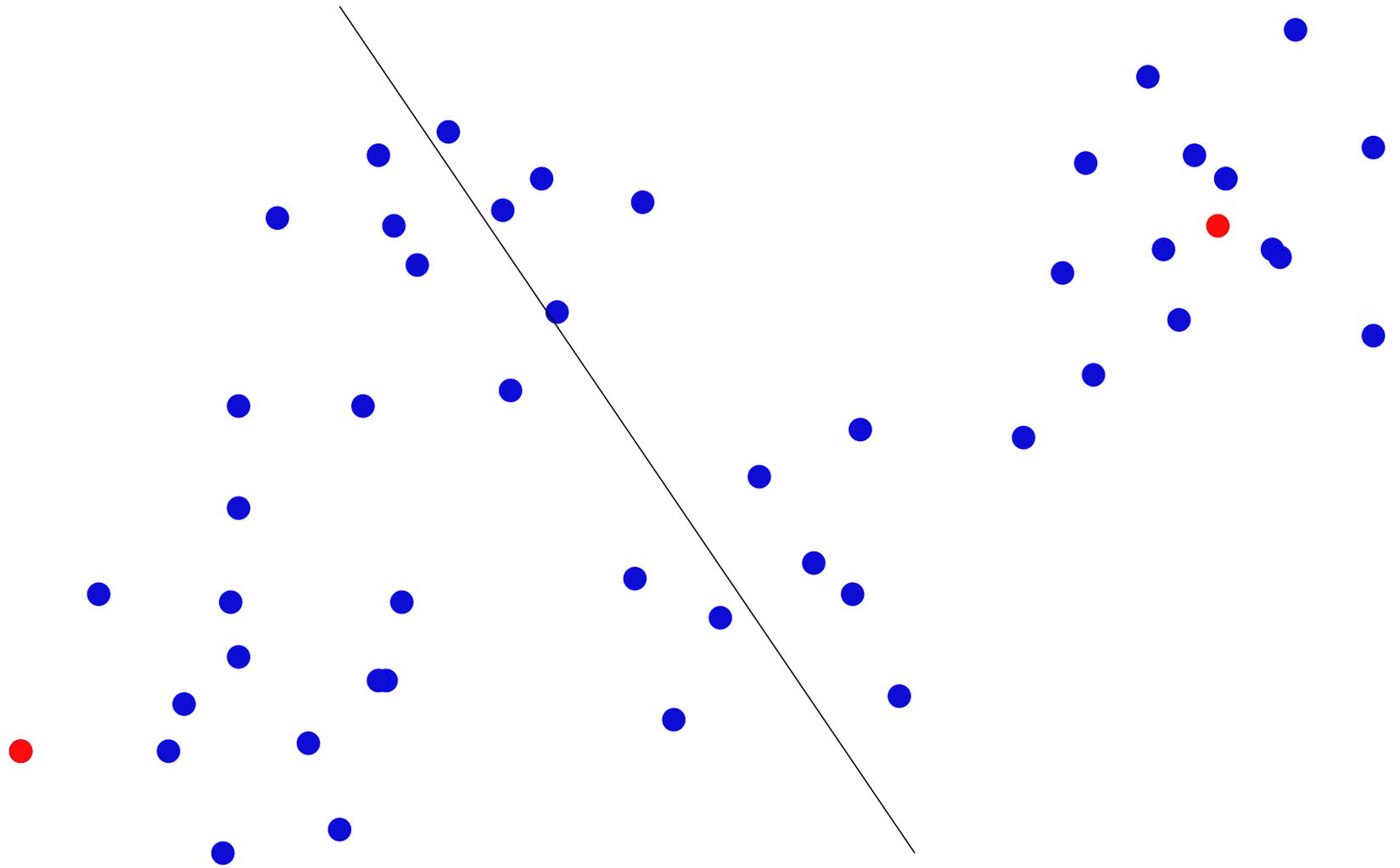
FPF con un esempio



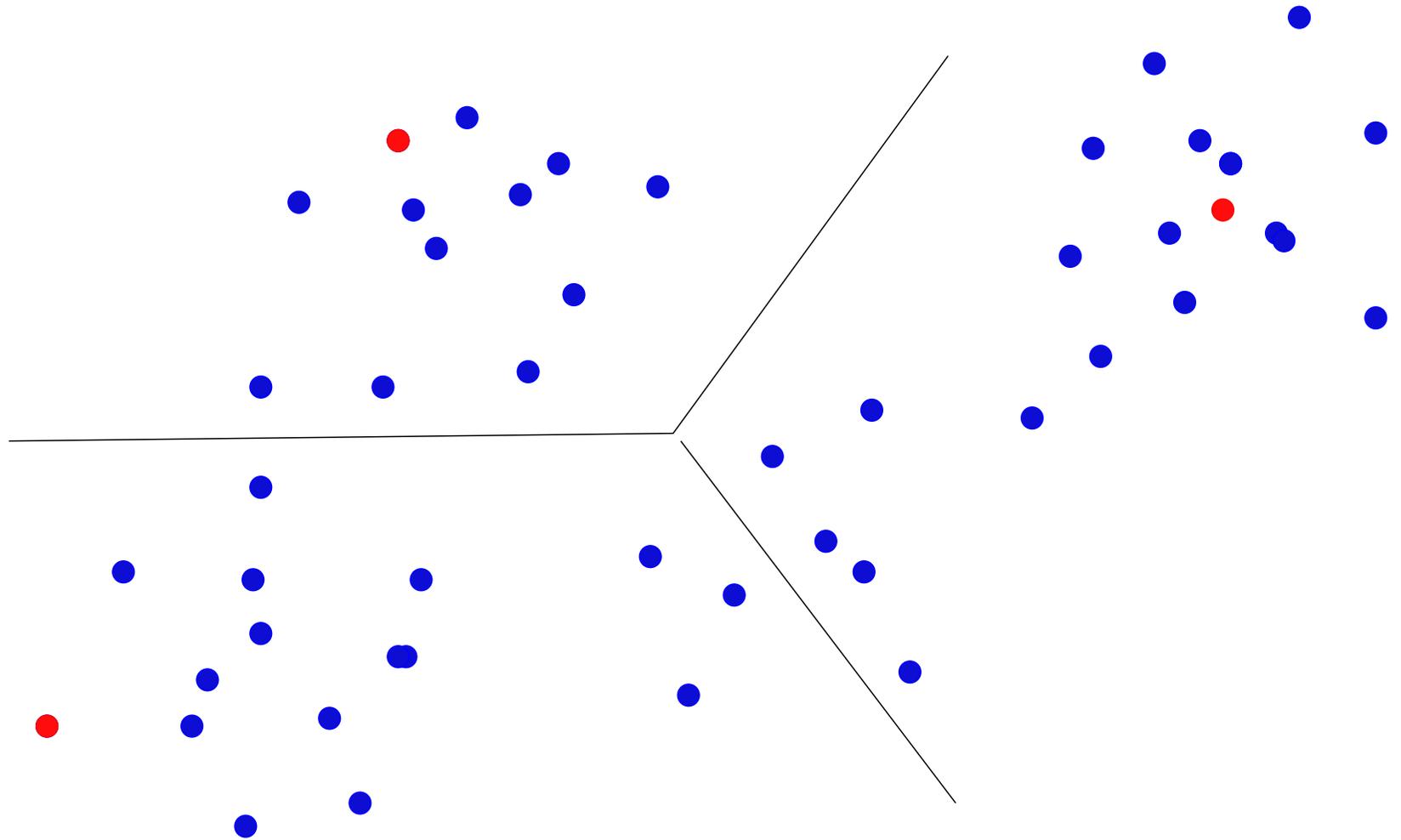
FPF con un esempio



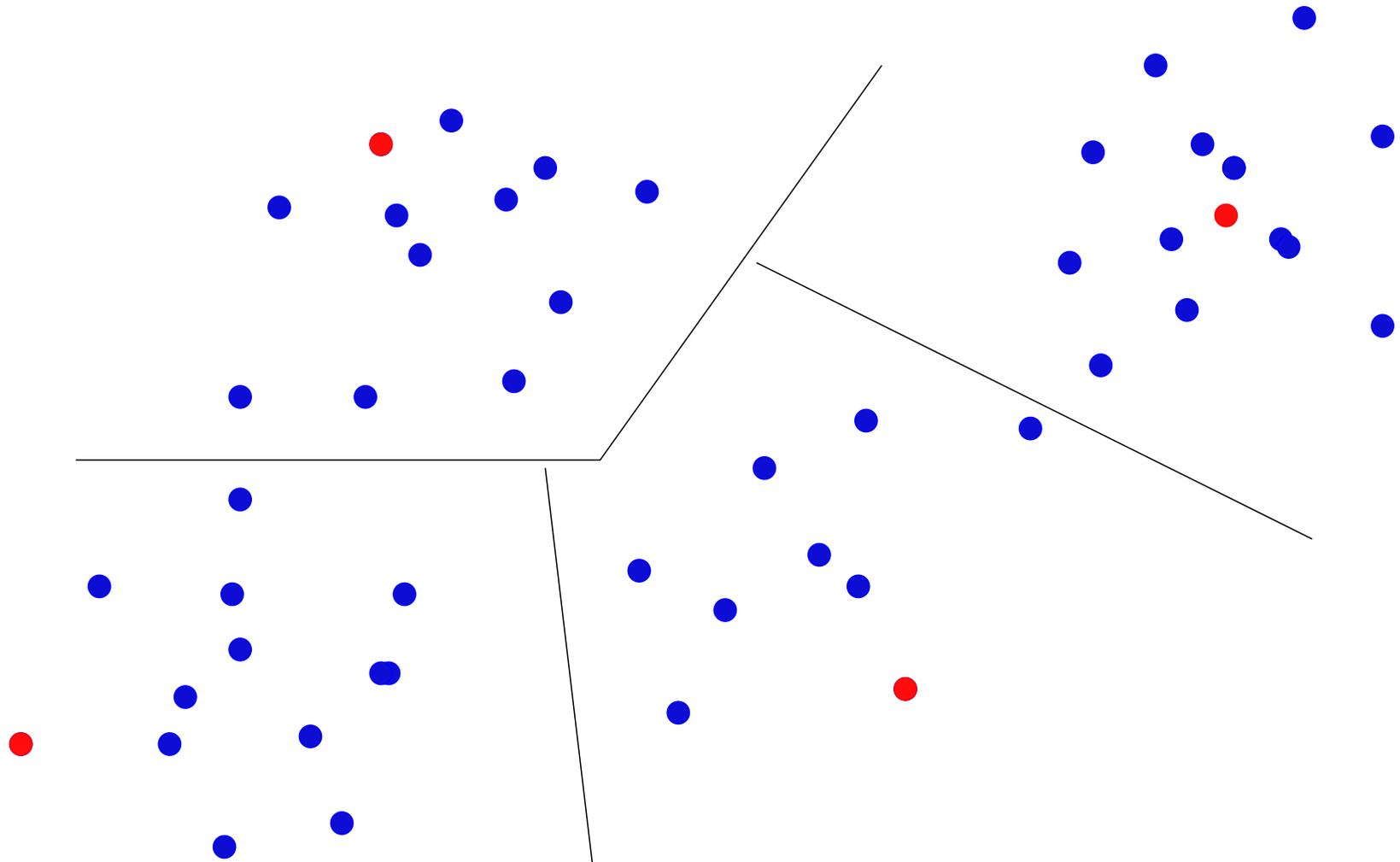
FPF con un esempio



FPF con un esempio



FPF con un esempio



K-means



Algoritmo k-means

- Assegna ogni punto al centroide più vicino
- Aggiorna il centroide
 1. Per ogni dimensione del punto calcola il valore intermedio

- In uno spazio euclideo equivale al baricentro

K-means



Inizializzazione

- Punti casuali
- Centri calcolati con FPF
- One pass k-means (McQueen)
 1. Centroidi inizialmente scelti casualmente
 2. Aggiungo un punto al cluster con il centroide più vicino
 3. Aggiorno il centroide
 4. Ricomincio dal punt 2 finche non finisco i punti da assegnare

Quanti cluster fare?

- Dipende dal problema in esame
 - Se sono pochi il contenuto non è omogeneo
 - Se sono troppi si rischia di “spezzare” cluster omogenei
- Dunque?
 - Soluzioni dipendenti dal problema
 - Possono non essere possibili
 - Approcci teorici
 - Sempre disponibili
 - Indipendenti dal problema



Esempio: uso del clustering per il problema del carrello (I)

- Problema: analizzare le abitudini di acquisto dei clienti trovando associazioni tra prodotti comprati
Acquisto di un cliente modellato come vettore di booleani in cui ad ogni elemento corrisponde un prodotto
- Itemset: insieme di oggetti venduti insieme
- Metodo:
 - Cercare itemset frequenti
 - Derivare regole di associazione

Esempio: uso del clustering per il problema del carrello (2)

- Per il clustering gli itemset possono essere rappresentati come vettori nello spazio di tutti i prodotti

Pane	Carne	Latte	Biscotti	Uova	Farina	Zucchero
	i			i		l
		l	l			l
l		l	l			l
		l	l		l	l

- Mediante clustering si possono raggruppare itemset non identici, ma con prodotti condivisi

Analisi associativa

- Permette di identificare condizioni che si verificano contemporaneamente con elevata frequenza
- Rileva pattern che si e ne deriva regole di implicazione del tipo $A \Rightarrow B$
 - Regola nella forma: Attributo (Soggetto, valore)
- **Esempio:**
 - Compra (X, biscotti) Compra (X, zucchero) \Rightarrow Compra (X, latte)
- Elementi interessanti sono coppie che co-occorrono frequentemente (supporto)
 - $supporto(A \Rightarrow B) = Prob(A \cup B)$



Classificazione



Classificazione e predizione

- **Costruzione di modelli per**
 - Predire gli eventi futuri
 - Stimare il valore di elementi non noti
- **Classificazione**
 - Definizione di criteri che permettono di assegnare un soggetto ad una classe
- **Predizione**
 - Calcolo di funzioni di tendenza continue tramite l'interpolazione dei dati noti



Classificazione e predizione

- Costruzione “basata su esempi”
 - Il modello deriva da un sottoinsieme significativo dei dati esistenti (**training set**)
 - L’efficacia viene testata su un sottoinsieme diverso (disgiunto) dei dati (**test set**)
 - Se il modello si rivela efficace può essere usato come ‘predittore’.
- Applicazioni
 - Propensione all’acquisto dei clienti
 - Qualità dei fornitori
 - Affidabilità dei prodotti



Classificazione

- Permette di indicare l'appartenenza di un elemento ad una certa classe
- Diversi tipi di modelli
 - Analisi statistiche
 - Regole associative
 - **Alberi di decisione**
 - Reti bayesiane (Naïve Bayes ne è un caso)
 - Reti neurali

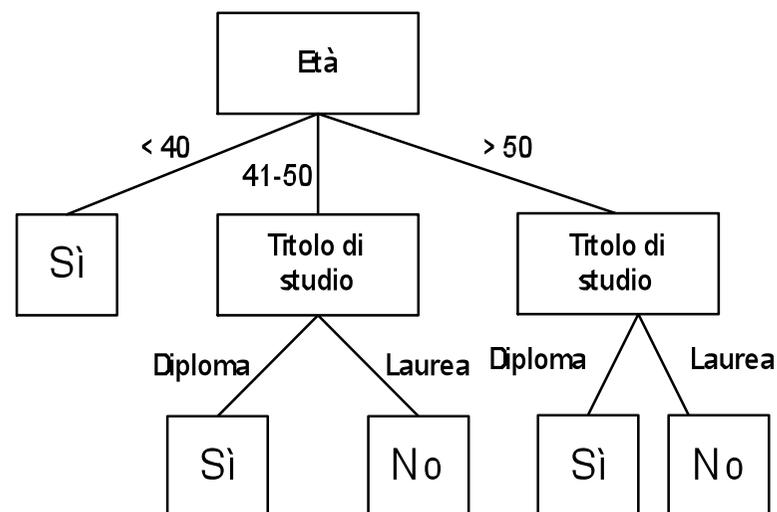


Alberi di decisione

- Rappresentano un insieme di regole che permettono di fare la predizione automaticamente
- Sono costruiti automaticamente usando i dati disponibili
- Struttura di classificazione basata sulla valutazione di condizioni del tipo if-then-else

Alberi di decisione

- Nodi interni
 - Attributi del soggetto da classificare
- Archi in uscita
 - Etichettati con i valori che l'attributo può assumere
- Nodi foglia
 - Classi
- La classificazione avviene seguendo un percorso guidato dai valori assunti dagli attributi dell'elemento da classificare



Esempio di albero di decisione che suddivide in due classi

Alberi di decisione: creazione

- La realizzazione dell'albero si basa su due fasi
 - Costruzione
 - Raffinamento
- Costruzione
 - Si cerca un buon criterio C per dividere i dati in due sottoinsiemi: $D1$, $D2$
 - un buon criterio dovrebbe minimizzare la profondità dell'albero
 - Si costruisce un nodo che usa il criterio C e si applica ricorsivamente l'algoritmo a $D1$ e $D2$
- Raffinamento (pruning)
 - L'albero costruito viene semplificato eliminando i rami meno significativi



Valutazione piattaforme di data mining

- Sia i sistemi di clustering che quelli di classificazione devono essere valutati prima di affidare le decisioni strategiche aziendali al sistema
- La bontà del sistema dipende:
 - Dal modello
 - Dagli algoritmi usati
 - Dal numero di cluster (solo clustering)
 - Dal training (solo classificazione)

Strategie di valutazione

Misure interne



- Modelli matematici
- Sempre disponibili
- Non correlati al problema

Misure esterne



- Confronto con il ground truth
- Spesso non disponibili
- Correlate al problema

Misure interne: omogeneità e separazione

- Gruppi molto omogenei indicano buon clustering/classificazione
- Gruppi molto separati indicano buon clustering/classificazione

$$\text{Homogeneity} = \frac{1}{\# \text{mates}} \sum_{o_i, o_j \text{ mates}, i < j} \text{Sim}(o_i, o_j)$$

$$\text{Separation} = \frac{2}{n(n-1) - 2\# \text{mates}} \sum_{o_i, o_j \text{ non-mates}, i < j} \text{Sim}(o_i, o_j)$$

- Due oggetti sono mates se appartengono allo stesso gruppo
- Alta omogeneità, bassa separazione indicano migliori risultati

Misure esterne: precision e recall per raggruppamenti

- Dato un raggruppamento c_j ed una classe GT_i si definiscono

$$precision(GT_i, c_j) = \frac{|GT_i \cap c_j|}{|c_j|}$$

$$recall(GT_i, c_j) = \frac{|GT_i \cap c_j|}{|GT_i|}$$

- La precision misura la probabilità che un elemento della classe GT_i sia anche nel gruppo c_j
- La recall misura la probabilità che un elemento nel gruppo c_j sia anche della classe GT_i

F-measure

- Dato un clustering/classificazione ed il suo ground truth, precision e recall possono essere combinate insieme:
- F-measure per una classe

$$F(GT_i, c_j) = 2 \frac{\text{precision}(iGT, c_j) \text{recall}(GT_i, c_j)}{\text{precision}(GT_i, c_j) + \text{recall}(GT_i, c_j)}$$

- Estensione all'intero raggruppamento

$$F = \sum_i \frac{|GT_i|}{n} \max_j (F(GT_i, c_j)),$$