Università degli Studi di Palermo

dipartimento di ingegneria unipa

Introduzione al Data Processing

CORSO DI BIG DATA

a.a. 2022/2023

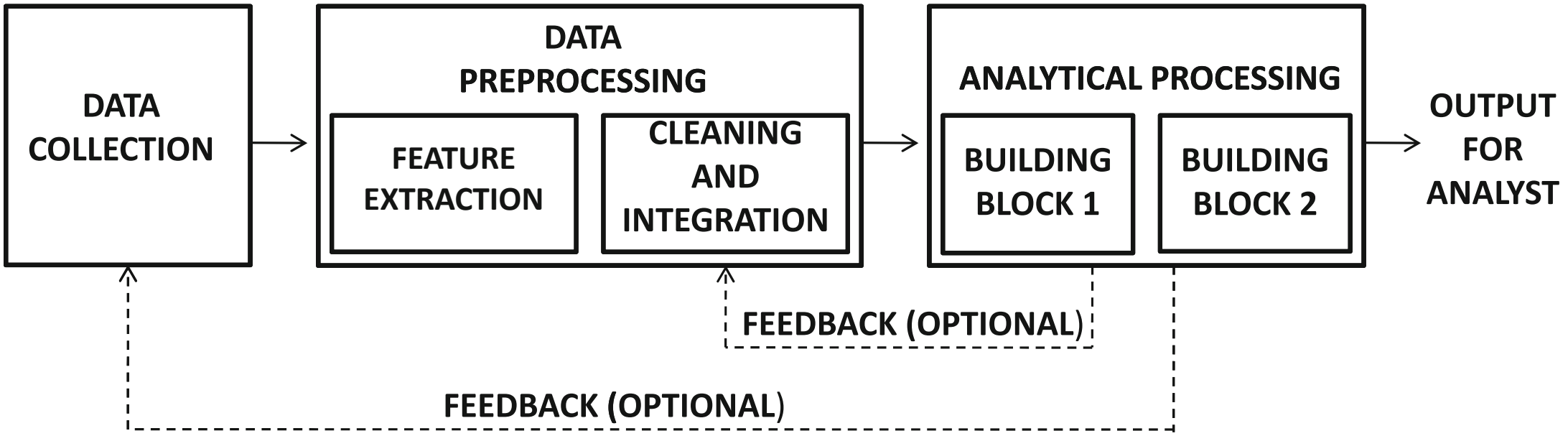
Prof. Roberto Pirrone

LABORATORIO DI INTERAZIONE UOMO-MACCHINA
CHILAB

Sommario

- Pipeline di data processing
- Tipologie generali di analisi dei dati
- Principali tipi di dati

Pipeline di data processing



Pipeline di data processing

- Data collection
  - I dati possono essere estratti da sorgenti eterogenee: sensori, log internet, corpora documentali, dati biomedici …
  - Problematiche di gestione di dati strutturati o parzialmente strutturati
  - Volume e velocità dei dati
  - Database, datawarehouse, data lake, HDFS e database NoSQL

Pipeline di data processing

- Feature extraction e data cleaning
  - Trasformazione dei dati grezzi in formati compatibili con gli algoritmi di analisi:
    - Vettori multidimensionali
    - Serie temporali
    - Dati binari o categorici
    - …
  - Feature: *una qualunque proprietà valorizzabile del dato* (es. i campi di un record)

Pipeline di data processing

- Feature extraction e data cleaning
  - Scegliere le feature più significative per il problema!!
  - Data cleaning: gestione dei dati mancanti o erronei
    - Stima dei dati mancanti
    - Correzione degli errori
    - *Dipende dalla conoscenza sul problema!!*
  - I dati «curati» (feature significative, gestione degli errori e selezione del formato) sono di nuovo conservati in un database (non necessariamente relazionale)

Pipeline di data processing

- Analisi dei dati
  - Si considerino i dati organizzati in un «database» $\mathcal{D}$ con *n* record e *d* attributi che può essere rappresentato da da una matrice di dati *X* composta da *n* vettori riga $\mathbf{x} \in \mathbb{R}^d$
  - In senso molto generale possiamo ricercare:
    - Strutture ricorrenti tra le feature *all'interno di ogni singolo dato* che lo correlano con un'ulteriore feature che è l'obiettivo dell'analisi e che può non essere nota nel problema reale (relazioni tra le colonne di *X*)
    - Similarità *tra diverse istanze dei dati* che li rendono *più simili tra loro rispetto agli altri* secondo qualche criterio di analisi (relazioni tra le righe di *X*)

Tipologie generali di analisi dei dati

- Classificazione
  - Predizione di una *etichetta* (label) discreta da associare a ciascun dato per categorizzarlo in una *classe di appartenenza*
  - *Supervisionato*: le etichette sono note per il data set disponibile per l'analisi, ma lo scopo dell'analisi è *predire* le etichette per nuovi dati in arrivo

- Clustering
  - Ricerca di similarità tra i dati per l'individuazione di *gruppi* (cluster) tra di essi
  - *Non supervisionato*: i gruppi sono non noti sia per numero sia per struttura nel data set disponibile per l'analisi

Tipologie generali di analisi dei dati

- Outlier analysis
  - Il clustering può essere utilizzato in forma duale per individuare degli elementi anomali o *outlier* che non si armonizzano con i dati analizzati
  - Outlier: «Un'osservazione che devia così tanto dalle altre da sollevare il sospetto sia stata generata da un meccanismo differente»
  - Intrusion detection
  - Rilevamento di frodi finanziarie o su carte di credito
  - Pattern anomali da rilevamento di sensori (prevenzione guasti)
  - Pattern inusuali nel medical imaging (malattie)
  - Previsioni meteo e ambientali

Tipologie generali di analisi dei dati

- Frequent pattern mining (Association pattern mining)
  - Si assuma che i dati siano binari che *X* sia molto sparsa
    - Database delle transazioni commerciali ottenuti dagli scontrini
    - Ogni record è un acquisto: solo gli item acquistati valgono 1
  - Supporto supp(*P*) di un pattern *P* (insieme di item o *itemset*) - frequenza relativa del pattern in *X*
    - Association pattern mining: non si usa la frequenza del pattern, ma altre misure di significatività statistica come il χ^2

Tipologie generali di analisi dei dati

- Association pattern mining (Frequent pattern mining)
  - Regola di associazione $A \Rightarrow B$: associazione tra pattern frequenti che ha una certa *confidenza*:
    - Frazione delle righe contenenti *A*, che contiene anche *B* $\mathrm{conf}(A \Rightarrow B) = \dfrac{\mathrm{supp}(A \cup B)}{\mathrm{supp}(A)}$
  - Mining di regole di associazione con supporto minimo *s* e confidenza *c*:

$$\mathrm{supp}(A) \geqslant s$$

$$\mathrm{conf}(A \Rightarrow B) \geqslant c$$

Principali tipi di dati

- Dati interdipendenti e non interdipendenti
  - L'interdipendenza è legata alla natura del fenomeno descritto dai dati per cui le singole istanze dipendono tra di loro
    - Dipendenze *implicite* ed *esplicite*
    - Serie temporali (letture successive di sensori, parole in un testo …)
    - Grafi (social network analysis, database molecolari …)
    - Dati geospaziali
  - Dati non interdipendenti:
    - Dati multidimensionali conservati in un database

Dati non interdipendenti

- Dati multidimensionali

$$\mathcal{D} = \{\bar{X}_i,\ i = 1, \ldots, n : \bar{X}_i = (x_i^1, \ldots, x_i^d)\}$$

Name	Age	Gender	Race	ZIP code
John S.	45	M	African American	05139
Manyona L.	31	F	Native American	10598
Sayani A.	11	F	East Indian	10547
Jack M.	56	M	Caucasian	10562
Wei L.	63	M	Asian	90210

Dati non interdipendenti

- Dati multidimensionali

$$\mathcal{D} = \{\bar{X}_i,\ \ i = 1, \ldots, n : \bar{X}_i = (x_i^1, \ldots, x_i^d)\}$$

- X_i
  - Data point
  - Istanza
  - Esempio
  - Tupla
  - Oggetto
  - Record
  - Transazione
- x_i^k
  - Campo
  - Attributo
  - Dimensione
  - Feature

Dati non interdipendenti

- Dati numerici
- Dati categorici: assumono un insieme *discreto e non ordinato* di valori
- Dati misti: numerici e categorici
- Binari: dati categorici a due valori o dati di appartenenza ad un insieme
- Dati testuali
  - *Vector Space Representation*: creazione di uno spazio euclideo in cui parole/frasi/documenti possano essere giudicati simili attraverso misure di distanza
  - Latent Semantic Analysis (LSA)

Dati interdipendenti

- Serie temporali
  - Attributi contestuali: descrivono l'implicita interdipendenza tra i dati
    - Time stamp
    - Indice di posizione
  - Attributi comportamentali: descrivono l'effettivo contenuto informativo della serie

  - In genere si tratta di dati multivariati continui o discreti
    - Sequenze discrete: stringhe, sequenze proteiche, dati di log …

$$(t_1; \bar{Y}_1), (t_2; \bar{Y}_2), \ldots, (t_n; \bar{Y}_n)$$

$$\bar{Y}_i = \left(y_i^1 \ldots y_i^d\right)$$

Dati interdipendenti

- Dati spaziali
  - Simili alle serie temporali: la marca è una geolocalizzazione L_i

$$(L_1; \bar{X}_1), (L_2; \bar{X}_2), \ldots, (L_n; \bar{X}_n)$$
$$\bar{X}_i = \left(x_i^1 \ldots x_i^d\right)$$

  - Si possono cercare regolarità su dati legati a marche spaziali adiacenti

Dati interdipendenti

- Dati spazio-temporali
  - Spazio e tempo sono entrambi attributi contestuali
    - Variazioni di parametri atmosferici/ambientali
  - Il tempo è un attributo contestuale, ma lo spazio è un attributo comportamentale
    - Analisi di traiettorie
    - Una «traiettoria» più in generale può essere una correlazione tra due attributi comportamentali di una serie temporale multivariata che vengono accoppiati ad ogni time stamp

Dati interdipendenti

- Dati di rete e grafi
  - I dati sono esplicitamente organizzati secondo un grafo

$$G = (V, E)$$
$$\forall v_i \in V,\ v_i \mapsto \bar{X}_i$$
$$\forall e_{i,j} \in E,\ e_{i,j} \mapsto \bar{Y}_{i,j}$$

Dati interdipendenti

- Dati testuali: word embedding
  - Rappresentazioni vettoriali delle parole in un testo in cui si cerca di catturare la dipendenza di una parola dalle altre
  - Ottenuti attraverso reti neurali
  - WE non contestuali: associazione univoca parola ←→ vettore
  - WE contestuali: si addestrano particolari reti neurali dette *language model* in grado di predire la parola successiva dopo aver osservato un certo numero di parole nel testo