FONDAMENTI DI SCIENZA DEI DATI

APPUNTI A CURA DI: RICCARDO LO IACONO

Università degli studi di Palermo

a.a. 2022-2023

Indice.

– 1 – Introduzione: pipeline di data processing.

Ogni progetto di scienza dei dati segue una pipeline ben definita. Questa parte dall'ottenimento dei dati, dati che possono essere acquisiti da sorgenti diverse: sensori, file di log, ecc. Indipendentemente dall'origine, ai dati sono legati due problematiche, quali

- gestione dei dati strutturati;
- gestione del volume dei dati.

Assunto di aver gestito queste problematiche, si può procedere alle fasi di data-cleaning e di estrazione delle feature. Qui con *data-cleaning* si fa riferimento ad un'insieme di tecniche atte a migliorare la qualità dei dati; in tale fase

- si gestiscono i valori nulli: questi se non gestiti porterebbero a un cattivo addestramento del modello;
- si eliminano eventuali duplicati: ciò è atto a prevenire un bias nel modello;
- si ricercano e rimuovono eventuali outliers: è ovvio che se non si procedesse a gestirli questi influenzerebbero negativamente il modello;
- si standardizzano i dati: se si utilizzano dati provenienti da fonti diverse, è necessario che questi abbiano una stessa struttura.

Successivamente si può procedere alla fase di *feature-extraction*: ossia la fase in cui si selezionano quelle caratteristiche valorizzabili dei dati, che possono essere utilizzate per addestrare il modello, ed eventualmente combinate per definire concetti più generali.

– 1.1 – Tipologie di dati.

I dati porterebbero essere suddivisi per diversi aspetti, è però di interesse la distinzione tra dati *interdipendenti* e *non-interdipendenti*. Sostanzialmente la differenza è la seguente: nel caso di dati interdipendenti, questi hanno una relazione di qualche tipo (**eg:** *peso-altezza*); segue che informazioni in uno o più dati, dipendono o influenzano altri dati. Per i dati non-interdipendenti non si hanno tali relazioni.
Osservazione. È giusto puntualizzare che il concetto di interdipendenza è fortemente legato al problema in esame.