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 – Data Integration.

Ulteriore fase della pipeline di analisi dei dati è quella di integrazione dei dati. Come anticipato, quando si addestra un modello è possibile utilizzare dati provenienti da sorgenti diverse; ciò implica verosimilmente che i dati presenteranno una struttura diversa, rendendo complicato l'addestramento del modello stesso.

Per risolvere tale problematica, nella fase di integrazione ci si occupa di uniformare la struttura dei dati, applicando eventuali trasformazioni, e di gestire eventuali inconsistenze. Con quest'ultime s'identificano dati che, sebbene associati ad uno stesso fenomeno, presentano valori tra loro discordi.

– 1.2 – Data integration.

Nella precedente sezione si è fatto a delle trasformazioni applicabili ai dati. Queste, atte a migliorare la qualità delle informazioni, si distinguono in

- *smoothing:* si tratta di una tecnica che applica un'ulteriore pulizia ai dati, così da rimuovere ulteriore rumore;
- *aggregazione:* i dati sono combinati tra loro così da permettere la descrizione di concetti più generali;
- *normalizzazione:* si riduce il range di valori assunti.

Per quanto riguarda la normalizzazione: esisto varie soluzioni, tra le più utilizzate: la *scalatura decimale*, con la quale si riduce il range tra (0, 1); e la *z-core*, con la quale si sottrae ai dati la media degli stessi, e li si divide per la rispettiva deviazione standard.

– 1.3 – Data reduction.

Poiché generalmente quando si opera con i dati si ha a che fare con quantità dell'ordine dei TB, sebbene ciò possa sembrare un lato positivo, in quanto il modello avrebbe più dati da cui apprendere; ciò risulta eccessivamente lento. Per questa e altre ragioni nella maggior parte dei casi si preferisce ridurre la mole dei dati (*Sezione* **??**).

– 1.4 – 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 qualche relazione (**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.