FONDAMENTI DI SCIENZA DEI DATI

APPUNTI A CURA DI: RICCARDO LO IACONO

Università degli studi di Palermo

a.a. 2023-2024

Indice.

– 1 – Pipeline di analisi dei dati.

Il processo di analisi dei dati parte banalmente dall'ottenimento stesso dei dati, cosa che può essere effettuata da database, file di log, o altre fonti. In ogni caso quando si lavora con i dati si deve prestare attenzione a due problemi, uno legato alla struttura dei dati, l'altro alla numerosità degli stessi.

In generale comunque, prima di operare con i dati si procede ad una fase di pulizia. Questa è atta a gestire

- *valori nulli o erronei:* è ovvio che la presenza di valori errati possa influire negativamente sul modello;
- *valori duplicati:* se non gestiti potrebbero portare ad un cattivo addestramento;
- *struttura dei dati:* è opportuno uniformare la struttura con cui i dati sono rappresentati.

Una volta effettuata la fase di pulizia, si può passare alla parte principale dell'analisi dei dati: l'estrazione delle feature. Dal un punto di vista strettamente formale, queste rappresentano caratteristiche valorizzabili dei dati, utili alla risoluzione di un qualche problema. Una volta estratte le feature, queste sono adattate per essere utilizzate da un modello di machine learning, ed eventualmente combinate per definire caratteristiche supplementari.

– 1.1 – Integrazione dei dati.

Fase essenziale della analisi dei dati, *l'integrazione* è atta alla gestione di dati provenienti da fonti diverse. In generale, in questa fase ci si occupa di

Si assume che dati provenienti da fonti diverse, avranno anche una rappresentazione diversa.

- uniformare la struttura dei dati;
- gestire le inconsistenze: ossia dati che, sebbene riferiti ad uno stesso evento, sono discordi tra loro.

– 1.2 – Trasformazione dei dati.

Una volta eseguita la fase di integrazione, i dati sono sottoposti ad operazioni di trasformazione utili ad ottimizzare l'addestramento del modello. Tra queste si considerano

- *lo smoothing:* si tratta di un'ulteriore fase di pulizia dei dati, con l'obbiettivo di eliminare del rumore eventualmente permasto;
- *l'aggregazione:* i dati sono tra loro combinati, per poter descrivere concetti più generici;
- *la normalizzazione:* per ridurre il numero di valori assunti dai dati, si procede a ridurre l'intervallo che questi possono assumere.

– 1.2.1 – Normalizzazione.

Sebbene esistano decine di normalizzazione diverse, di interesse risultano

- *la normalizzazione max-min:* si tratta di una normalizzazione dipendente dai valore di massimo e minimo dei dati. Nello specifico, assunti v i dati, v' i dati normalizzati, MAX_0 e min_0 rispettivamente i valori di massimo e minimo iniziali, e assunti inoltre MAX e min gli estremi del nuovo range, segue

$$v' = \frac{v - min_0}{(MAX_0 - min_0)}(MAX - min) + min$$

In generale, poiché si preferisce operare con valori in $[0,1]$, $Max = 1, min = 0$.

- *la normalizzazione z-core:* è una normalizzazione dipendente dalla media e dalla deviazione standard dei dati. Nello specifico, posti v i dati, v' i dati normalizzati, μ la media e σ la deviazione standard, si ha

$$v' = \frac{v - \mu}{\sigma}$$

- *la normalizzazione decimale:* si normalizza dividendo per una potenza di 10 tale che, a seguito della normalizzazione, i valori cadano nel range $(0,1)$. Formalmente

$$v' = \frac{v}{10^k}, k = \min_{k \in \mathbb{N}} \{||v'|| \leq 1\}$$

– 1.3 – Riduzione della dimensionalità.

In generale, quando si opera con i dati, questi risultano essere di grandi quantità. Poiché gestirli tutti risulta complesso, si preferisce molto spesso ridurre al necessario i dati da utilizzare. Per far ciò si utilizza una delle diverse tecniche di riduzione della dimensionalità, tra queste

- *l'aggregazione:* simile a quella descritta nella fase di trasformazione;
- *la selezione degli attributi minimi:* si procede ad estrarre un sottoinsieme dei dati, tali che questi siano sufficienti a rappresentare l'intero insieme. Di questa tecnica si distinguono
  - *la selezione in avanti:* a partire dall'intero set di dati S, si costruisce un sottoinsieme S' estraendo di passo in passo un elemento da S e inserendolo in S', se e solo se questi migliora la qualità dei dati in S'.
  - *la selezione all'indietro:* da S si rimuove un elemento, se a seguito di una sua eventualmente estrazione, la qualità degli elementi in S risulta massimizzata.

– 2 – Stimatori e predittori.

L'intero processo di machine learning è basato sulla *teoria dell'apprendimento statistico*. Secondo tale teoria, i dati che rappresentano un qualunque fenomeno, possono essere visti come appartenenti ad una qualche distribuzione di probabilità ignota; e per tale motivo si può assumere che essi siano tra loro indipendenti ed equiprobabili, concludendo pertanto che il valore atteso dei dati di addestramento e quello dei dati di training coincida. A seconda del tipo di machine learning adoperato (si veda *Sezione 3*), l'intero processo fa uso di *stimatori* o di *predittori*.

Definizione: uno stimatore, dal punto di vista statistico, è una funzione che, in funzione dei dati, permette di stimare una quantità/funzione interessante dei dati.

Di questi, supposto Θ_n uno stimatore, Θ il valore da stimare, si distinguono

- *gli stimatori polarizzati:* ossia stimatori tali che

$$\mathbb{E}(\Theta_n) - \Theta \neq 0.$$

  Cioè, lo stimatore commette un certo errore nell'approssimare Θ;

- *gli stimatori non-polarizzati:* si tratta di stimatori tali che

$$\mathbb{E}(\Theta_n) - \Theta = 0.$$

  Ossia, nell'approssimareΘnon si commette alcun errore.

Osservazione. Se si verifica che

$$\lim_{n\to\infty} \mathbb{E}(\Theta_n) = \Theta$$

si dirà che Θ_n è uno stimatore asintoticamente non-polarizzato.
Ulteriori caratteristiche degli stimatori sono la correttezza e la coerenza, nello specifico

$$\Theta_n \text{ corretto} \iff \mathbb{E}(\Theta_n) = \Theta$$
$$\Theta_n \text{ coerente} \iff \lim_{n\to\infty} \mathrm{Var}(\Theta_n) = 0$$

Ovviamente un buon stimatore sarà sia corretto che coerente.

Definizione: un predittore è un algoritmo della forma

$$y = f(x) + \varepsilon$$

che permette di stimare la funzione f sulla base dei dati di addestramento x.

Segue dalla definizione che, affinché y sia buona, l'errore ε deve essere minimo.

– 2.1 – MSE e bias-variance decomposition.

Si è parlato di stimatori e predittori, rimane dunque da discutere come determinare la "bonta" degli stessi. Sebbene ne esistano altre, la tecnica che risulta di interesse è l'*MSE*. Questi nel caso degli stimatori è definito come

$$MSE(\Theta_n) = \mathbb{E}\left[(\Theta_n - \Theta)^2\right] = \underbrace{\mathbb{E}(\Theta_n)^2 + \Theta^2 - 2\,\mathbb{E}(\Theta_n)\,\Theta}_{\text{Bias}(\Theta_n)^2} + \underbrace{\mathbb{E}\left(\Theta_n^2\right) - \mathbb{E}(\Theta_n)^2}_{\text{Var}(\Theta_n)}$$

Osservazione. Tale decomposizione prende il nome di *bias-variance decomposition.*

Supposto infine $y = f(x) + \varepsilon$ un predittore, e $\tilde{f}$ l'approssimazione di f sulla base dei dati, si dimostra che

$$MSE(y) = \mathbb{E}\left[\left(y - \tilde{f}\right)^2\right] = \cdots = \text{Var}\left(f(x)\right) + \text{Var}\left(\varepsilon\right) + \text{Bias}\left(f(x)\right)^2$$

– 3 – Tipologie di machine learning.

Sia assunto

$$y = f(x, z)$$

un algoritmo di machine learning. Da un punto di vista matematico, y deve fornire una valutazione statistica della relazione tra l'input e l'output dell'algoritmo. Formalmente, f è un modello di ML e (x, z) i suoi parametri.

– 3.1 – Apprendimento supervisionato.

Si tratta di una tipologia di machine learning in cui i dati di addestramento sono "etichettati": cioè presentano un campo che descrive la classe di appartenenza degli stessi. Tra questi si distinguono

Entrambe le categorie saranno discusse nelle sezioni a seguire.

- gli algoritmi di regressione;
- gli algoritmi di classificazione.

In generale, si utilizzano quando l'obbiettivo è quello di separare in classi i dati. Si osserva però che affinche l'addestramento possa definirsi "buono", l'algoritmo deve minimizzare l'errore relativo i dati di addestramento: il cosiddetto *training error*, e l'errore relativo i dati di test: il cosiddetto *test/generalization error*. In fine, per quanto detto e quanto discusso in *Sezione 2*, si deve fare attenzione

- all'*over-fitting:* ossia un fenomeno per cui il modello si è troppo adattato ai dati, non riuscendo dunque a generalizzare;
- all'*under-fitting:* fenomeno opposto all'over-fitting, è una condizione in cui il modello ha appreso poco dai dati, pertanto non ha le capacità sufficienti a generalizzare.

– 3.2 – Apprendimento non-supervisionato.

È un caso di machine learning in cui i dati non sono "etichettati". Tra questi, gli algoritmi di interesse sono

Come per gli algoritmi di ML supervisionato, entrambe le categorie saranno discusse nelle sezioni a seguire.

- quelli di riduzione della dimensionalità;
- quelli di clustering.

In tal senso, l'obbiettivo è quello di determinare similarità tra i dati.