Explainability - Che fare
Carlo Batini

Tecniche per la explainability

- Prima di entrare nel merito sulle tecniche che abilitano la interpretabilità, occorre chiedersi anzitutto, seguendo [Guidotti 2018], cosa si intenda per interpretabilità e per i collegati concetti di esistenza di una spiegazione (o spiegabilità) e comprensibilità.

Opacità - 1

- [Burrel 2016] introduce tre tipi di opacità. La prima forma di opacità è la deliberata forma di auto-protezione da parte dell'owner del dato con il fine di proteggere segreti commerciali per vantaggio competitivo. Ben note alternative ai modelli cosiddetti proprietari sono il software open source, e i dati open. Non ci occuperemo nel seguito di questa forma di opacità.

Opacità - 2

- La seconda forma di opacità è quella introdotta in precedenza con gli esempi dei programmi; scrivere programmi è una competenza caratteristica di programmatori, che hanno skill specializzati. La comprensione dei programmi resta inaccessibile alla maggioranza degli utenti. Le metodologie della ingegneria del software enfatizzano l'importanza della scrittura di programmi chiari, eleganti e comprensibili.
- Per contrastare questa forma di opacità, citiamo il movimento che a livello mondiale porta avanti l'idea del pensiero computazionale a tutti i livelli della formazione;: pensiamo solo alla importanza del pensiero computazionale nella professione del giornalista per il quale le tradizionali tecniche di indagine sono sempre più spesso arricchite o sostituite da un lavoro di ricerca di dati descrittivi di fatti sul Web, dati che, come abbiamo notato nel capitolo xx, sono per loro natura non verificati e opachi.

Opacità - 3

- La terza forma di opacità è quella che ci interessa, e riguarda i modelli abilitati dal machine learning.
- In questo caso la opacità non riguarda solo la comprensibilità del programma o del collegato algoritmo, ma, piuttosto, l'essere capaci di comprendere l'algoritmo in azione, mentre apprende dai dati il modello di learning.
- Anche se si possono concepire tecniche di machine learning facilmente comprensibili, è difficile che la comprensibilità si concili con la utilità. Modelli costruiti con tecniche di apprendimento che siano effettivamente utili, ed in particolare, accurati, nel senso introdotto in precedenza) posseggono un grado di inevitabile complessità.

L’interpretabilità

- L’interpretabilità si occupa di rendere esplicite le interazioni tra la tecnica di apprendimento e i dati su cui essa opera. Essa è rilevante sia quando un modello decisionale è investigato per scoprire bias e discriminazioni sistematiche, sia quando si vuole spiegare una decisione che riguarda un singolo individuo.
- Supponiamo per esempio che un modello decisionale produca una graduatoria per accedere a un servizio. Se un individuo inserisce i suoi dati e riceve come risultato un punteggio, questo numero da solo non fornisce alcuna informazione sul perché sia stato assegnato tale punteggio e sul perché della posizione comparativa rispetto agli altri partecipanti.

La comprensibilità

- Interpretare significa dare o fornire il significato o spiegare e presentare in termini comprensibili dei concetti. Pertanto, nel data mining e nel machine learning l' interpretabilità è la capacità di spiegare o fornire significato in termini comprensibili per un essere umano. In sostanza, una spiegazione è una interfaccia tra esseri umani e un decisore, spiegazione che è allo stesso tempo una approssimazione accurata dell'azione svolta dal decisore che comprensibile agli umani.
- Una importante caratteristica della interpretabilità è la comprensibilità, cioè lo sforzo cognitivo necessario all'essere umano per interpretare il modello di apprendimento. Potrebbe accadere che siamo riusciti a fornire una spiegazione che rende il modello di apprendimento interpretabile, ma che questa spiegazione sia troppo complessa per essere compresa dall'umano.

Distinzione tra interpretabilità globale e locale

- Una ulteriore precisazione che dobbiamo fare riguardo alla interpretabilità è la distinzione tra interpretabilità globale e locale
- Un modello è globalmente interpretabile se siamo in grado di comprendere la logica complessiva del modello e seguire l'intero ragionamento che porta ai differenti possibili risultati della classificazione, è localmente interpretabile se siamo in grado di comprendere le motivazioni per uno specifica decisione/predizione.

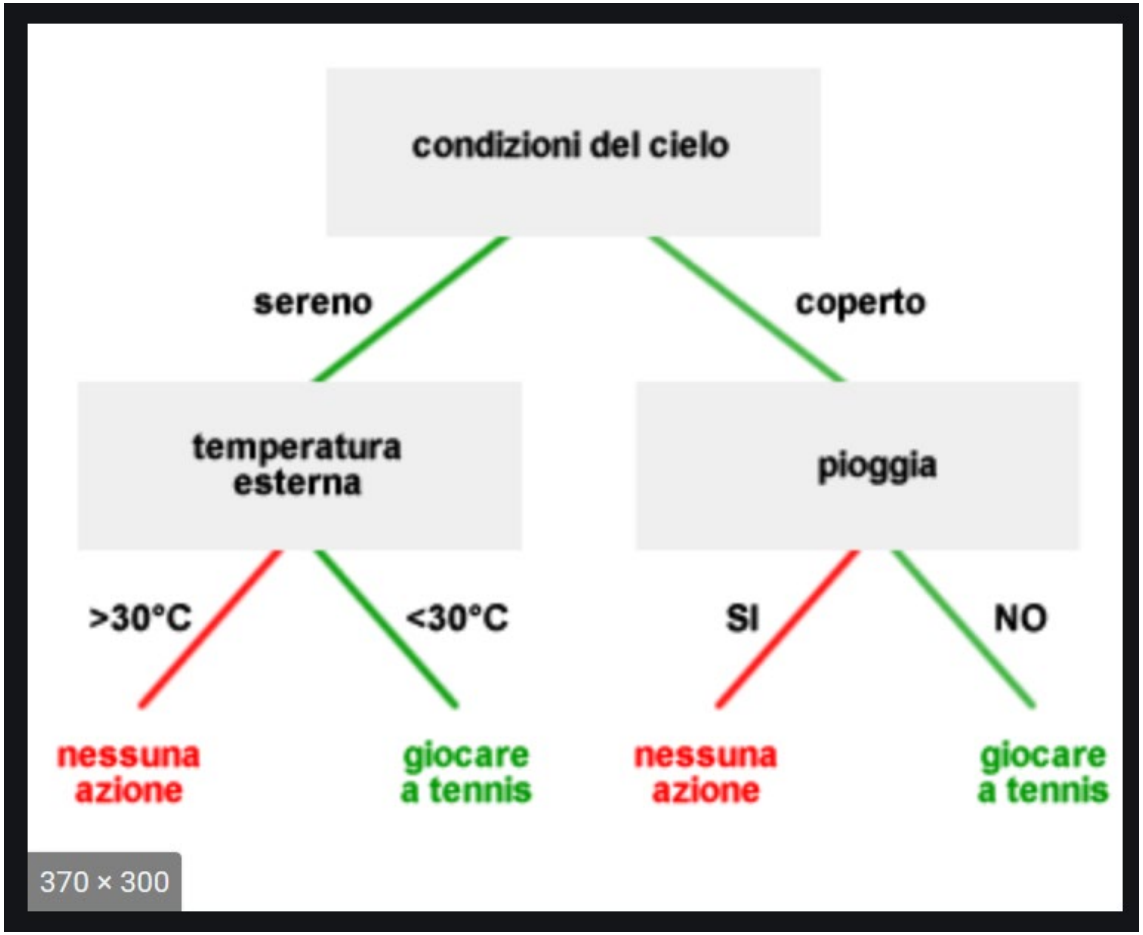
Modelli interpretabili

- Allo stato dell'arte, i modelli considerati interpretabili, cioè la cui procedura decisionale è comprensibile, sono tre:

1. gli alberi di decisione,
2. i sistemi a regole,
3. i modelli lineari.

- Consideriamo le prime due. Gli alberi di decisione sono stati introdotti nel Capitolo xx, ad esso rimandiamo per le definizioni e gli esempi. Anche in virtù della rappresentazione grafica (un grafico spiega più di mille parole…) possiamo convenire che essi sono una tecnica comprensibile anche senza molte conoscenze di machine learning. I sistemi a regole esprimono il procedimento decisionale/classificatorio per mezzo di formule logiche del tipo:

Albero di decisione



I sistemi a regole

- Esempio di regola - Se la febbre è superiore a 38 **and** la gola è rossa **and** il paziente starnutisce frequentemente **allora** il paziente ha una influenza

- Quando applichiamo un sistema a regole, dobbiamo applicare l'antecedente della regola ai dati in input al sistema (detti anche osservazione), e se essi rispettano la formula logica, allora possiamo associare a tali dati il valore di classificazione che compare a destra della regola (nel nostro caso, "ha una influenza"). Anche se i sistemi a regole non hanno naturalmente associata una rappresentazione grafica, essi esprimono un tipo di ragionamento logico che ci è piuttosto usuale.

Modelli candidati a esprimere le spiegazioni

Individuati negli alberi di decisione e nei sistemi a regole i modelli che per la loro natura comprensibile sono candidati a esprimere le spiegazioni connesse alla interpretabilità, vediamo ora si precisare un po' meglio il perimetro della interpretabilità, distinguendo due aspetti in cui è coinvolta una spiegazione; essa può riguardare:

- il modello a scatola nera (black box) nel suo complesso:
- uno specifico risultato del modello, relativo ad uno specifico valore di input.
- dando luogo ai seguenti due problemi.

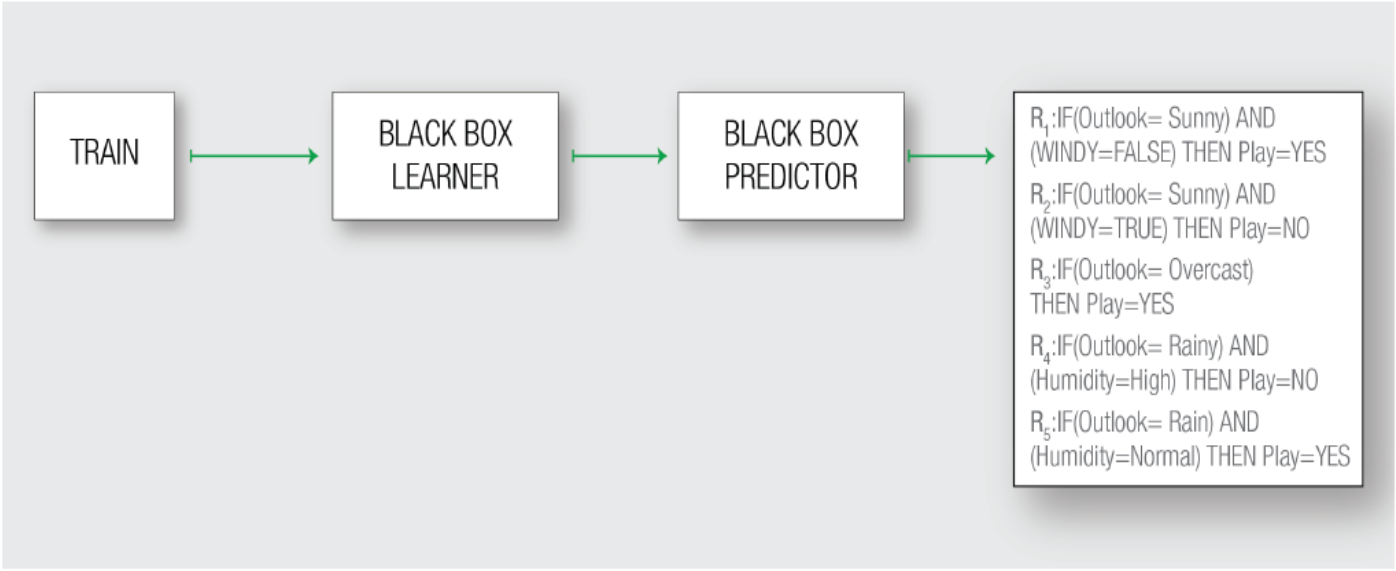
Problema 1 - Trovare una spiegazione per il modello a scatola nera

- Possiamo definire il problema in questo modo: dato un modello M1 a scatola nera che risolve un problema di classificazione, trovare una spiegazione per il modello M1 consiste nel costruire un nuovo modello M2 tra quelli considerati nativamente interpretabili (albero o regole), che imita il comportamento del modello non nativamente interpretabile e che è anche globalmente interpretabile (cioè è in grado di interpretare tutto il modello e non solo casi particolari).

Problema 1 - Trovare una spiegazione per il modello a scatola nera

- M2 deve inoltre auspicabilmente essere accurato, cioè fornire risultati di classificazione con qualità simile a quella di M1.
- Per quanto riguarda quest'ultimo punto, è chiaro che non basta spiegare, la spiegazione non deve divergere rispetto al modello originario fornendo risultati "sballati".

Esempio di spiegazione per il Problema 1



Problema 2 - Trovare una spiegazione per uno specifico risultato prodotto dal modello

- In questo caso il problema consiste nel fornire un risultato interpretabile; in altre parole, il modello interpretabile deve fornire un modello predittivo, insieme alle ragioni di tale predizione per un particolare valore di input.
- Chiaramente in questo caso la predizione è solo localmente interpretabile. Non è necessario spiegare l'intera logica del modello.

Esempio di spiegazione per il problema 2