Data Mining e Bioimmagini

Classificazione ruled-based, copertura e accuratezza

Prof. Pierangelo Veltri – 18/10/2023- Autori: Panarello, Saeedzadeh - Revisionatori: Panarello, Saeedzadeh

Per studiare meglio questo argomento, il professore ha consigliato di leggere le slide e il libro (capitolo 5) che ha condiviso sul canale Teams.

Con l'albero decisionale si effettua:

- **Splitting:** Questo implica la divisione dell'insieme già rappresentato in parti più piccole. La suddivisione comporta l'identificazione di attributi e la decisione più efficace nel separare i dati in gruppi; più un attributo è efficace, più sarà posizionato in alto nella gerarchia. Questa operazione nel contesto dei database è nota come **ottimizzazione**, cioè si effettua un filtraggio dei dati per attributi.
- **Ordering:** ossia scegliere quale elemento mettere prima e quale mettere dopo, stabilendo così l'ordine.

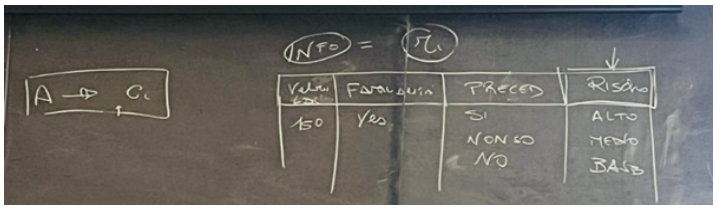
Gli alberi decisionali funzionano bene come classificatori quando l'insieme dei record può essere ragionevolmente diviso o organizzato geometricamente in settori distinti. Volendo rappresentare uno spazio in due o tre dimensioni come zone geometriche, l'idea è che queste zone siano omogenee tra loro. Più queste suddivisioni sono numerose, più l'albero decisionale avrà molte foglie, permettendo di identificare diversi casi distinti.

Il problema si verifica spesso quando si ha a che fare con rappresentazioni che contengono nuvole di punti. Inizialmente, i dati potrebbero non essere facilmente divisibili in categorie chiare. Questo significa che, quando arriva un nuovo dato, anche se si trova nelle foglie dell'albero, l'accuratezza in termini di correttezza potrebbe non essere molto affidabile. Ciò può essere dovuto a due fattori: in primo luogo, i dati possono essere sbilanciati, il che significa che mancano elementi distintivi per una classificazione efficace. In secondo luogo, si potrebbe avere un numero così elevato di attributi che sembrano distinguere con precisione tutte le classi durante la fase di addestramento, ma potrebbero non essere altrettanto efficaci nella classificazione in situazioni reali; così, quando si applica il modello a un set di test, potresti sperimentare difficoltà nell'interpretazione dell'albero e alla fine ottenere risultati che non sono sempre affidabili. Questo concetto è collegato a problemi noti come overfitting e underfitting.

*In statistica e in informatica, si parla di **overfitting** o **sovradattamento** (oppure adattamento eccessivo) quando un modello statistico molto complesso si adatta ai dati osservati (il campione) perché ha un numero eccessivo di parametri rispetto al numero di osservazioni. L'underfitting è uno scenario nella data science in cui un modello di dati non è in grado di cogliere la relazione tra le variabili di input e di output in modo accurato, generando un elevato tasso di errore sia sul set di addestramento che sui dati non visti. Si verifica quando un modello è troppo semplice, che può essere il risultato di un modello che ha bisogno di più tempo di addestramento, più caratteristiche di input e meno regolarizzazione. Come succede per l'overfitting, un modello soggetto ad underfitting non può stabilire la tendenza dominante all'interno dei dati, da cui ne conseguono errori di addestramento e scarse prestazioni del modello. Se non è in grado di eseguire bene la generalizzazione sui nuovi dati, un modello non può essere sfruttato per le attività di classificazione o previsione. La generalizzazione di un modello sui nuovi dati è in definitiva ciò che ci consente di utilizzare gli algoritmi di machine learning ogni giorno per fare delle previsioni e classificare i dati. (Approfondimento preso da internet)*

RULE BASED: È una tecnica in cui un insieme di informazioni (le condizioni) implicano un'altra informazione (la regola). Poiché stiamo parlando di classificazione delle informazioni, l'idea rimane la stessa: ho un insieme di attributi nel mio dataset che considero come elementi discreti, cioè come informazioni che devono guidarmi verso l'appartenenza a un cluster.

Esempio
Qui abbiamo il valore di LDL, mentre l'attributo 'rischio' appartiene a un dominio definito come alto, medio e basso. Nel contesto di un albero decisionale, il concetto di dataset significa utilizzare i dati degli attributi che non sono considerati nella classe come informazioni utilizzabili per associare il dato all'appartenenza a una classe o meno. In altre parole, si utilizzano gli attributi non direttamente collegati alla classe per effettuare una decisione di classificazione.



Valore <150 ^ (familiarità=yes v preced=yes) ➔ rischio= medio
Valore < 150 ^ (familiarità=no ^ preced=no) ➔ rischio= basso

Rule-based significa creare regole utilizzando le informazioni disponibili, dove ciascuna regola ha una forma congiuntiva. Ad esempio, se il valore LDL è inferiore a 150 e familiarità è uguale a 'sì' o precedenti è 'sì', allora il rischio viene classificato come 'medio'. Altrimenti, se il valore LDL è inferiore a 150 e familiarità è 'no' e precedenti è 'no', allora il rischio viene classificato come 'basso'. Queste regole sono formulate utilizzando le informazioni presenti nel dataset di partenza.

- ✓ Il punto di partenza è sempre lo stesso: trovare un modello che sia in grado di identificare l'appartenenza a una classe. Ovviamente, per farlo, è necessario utilizzare un training set e un test set.
- ✓ Data una collezione di record (training set) ogni record è composto da un insieme di attributi, di cui uno esprime la classe di appartenenza del record.
- ✓ Trova un modello per l'attributo di classe che esprima il valore dell'attributo in funzione dei valori degli altri attributi.
- ✓ Obiettivo: record non noti devono essere assegnati a una classe nel modo più accurato possibile. Viene utilizzato un test set per determinare l'accuratezza del modello. Normalmente, il data set fornito è suddiviso in training set e test set. Il primo è utilizzato per costruire il modello, il secondo per validarlo.
- ✓ I classificatori possono essere utilizzati sia a scopo descrittivo sia a scopo predittivo.
- ✓ Sono più adatti ad attributi nominali (binari o discreti) poiché faticano a sfruttare le relazioni implicite presenti negli attributi ordinali, numerici o in presenza di gerarchie di concetti (es. scimmie e uomini sono primati).

Classificatori basati su regole
Classificano i record utilizzando insiemi di regole del tipo "if..., then…"
Una regola ha la forma:

- (Condizione) y [y è l'etichetta di classe]
- LHS (left hand side): antecedente, è una congiunzione di predicati su attributi
- RHS(right hand side): conseguente

Esempi di regole:

- (Blood Type = Warm) (Lay Eggs =Yes) Class = Birds

La regola indica che se un animale ha il tipo di sangue caldo (Warm) e depone le uova (Lay Eggs = Yes), allora viene classificato come un uccello (Class = Birds).

- (Taxable Income < 50K) (Refund = Yes) Evade=No

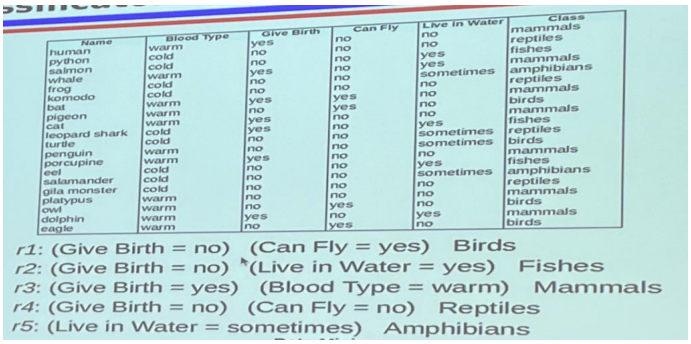
La regola indica che se il reddito tassabile (Taxable Income) di una persona è inferiore a 50.000 e ha diritto a un rimborso (Refund = Yes), allora non evaderà le tasse (Evade = No).
Costituire un modello rule-based significa partire da un dataset, dividerlo in un training set e un test set, e identificare delle regole. Questa volta, le regole non si basano sugli attributi singolarmente, ma prendono gli attributi e li combinano tra loro in diverse combinazioni.

I vantaggi dei modelli rule-based includono:

- Facili da interpretare
- Facili da generare
- Rapidi nella classificazione di nuove istanze

Svantaggi:

- Il costo di costruzione non scala al crescere del training set
- Risentono fortemente del rumore sui dati



La condizione rule based non si stabilisce per ogni attributo, ma si crea definendo delle espressioni che combinano gli attributi.

Esempio
La regola 1 dice che se l'animale non partorisce (Give Birth = no) e può volare (Can Fly = yes), allora è classificato come un uccello.
r2 dice che se l'animale non partorisce (Give Birth = no) e vive in acqua (Live in Water = yes), allora è classificato come un pesce (Fishes).
r3 dichiara che se l'animale partorisce (Give Birth = yes) e ha il tipo di sangue caldo (Blood Type = warm), allora è classificato come un mammifero (Mammals).
Se prendo la regola R1, ho tre tuple che rispondono a questa regola. In realtà, il problema dei modelli rule-based è che le loro caratteristiche consentono di avere più tuple o record che rispettano le regole, cioè si verifica una sovrapposizione nel senso che le regole spesso non sono completamente esclusive. È possibile avere condizioni e regole che si sovrappongono tra loro.
** Attenzione, le regole che abbiamo notato sono basate su una parte degli attributi e non su tutti
Una regola r copre un'istanza x se i valori di x soddisfano l'antecedente di r
Significa che ho trovato le tuple che sono perfettamente rappresentate dalla regola stessa. Quindi, le regole che ho definito precedentemente sono regole che coprono alcune tuple successive del mio dataset.

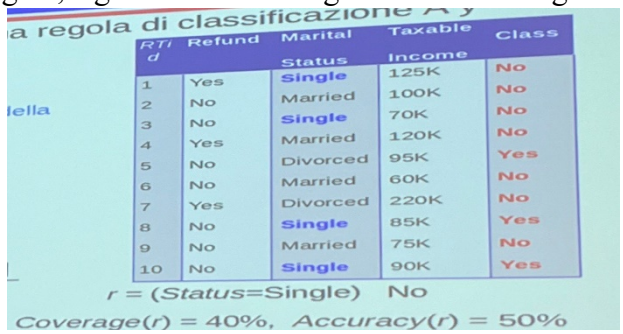
COPERTURA E ACCURATEZZA
Dato un dataset D e una regola di classificazione A y definiamo:

- **Copertura:** frazione dei record che soddisfano l'antecedente (A) della regola sul totale (D)
- **Accuratezza:** frazione dei record che soddisfano il conseguente della regola (A ∩ y) rispetto la frazione dei record che soddisfano l'antecedente (A)

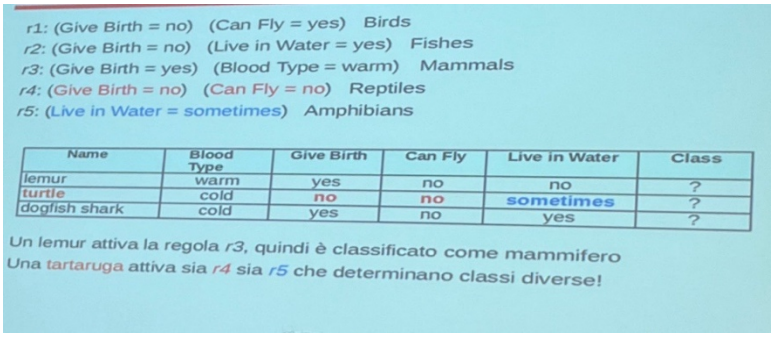
$$\text{Coverage}(r) = \frac{|A|}{|D|}$$

$$\text{Accuracy}(r) = \frac{|A \cap y|}{|A|},$$

La copertura serve per misurare il grado di corrispondenza delle regole rispetto al dataset, ossia quanto le regole scelte sono efficaci nel coprire le tuple o i record del dataset. Per l'accuratezza, prendo le regole e considero tutte le istanze del mio dataset che le rispettano, fino a quando si verificano le regole stesse. Quindi, l'accuratezza indica sostanzialmente il numero di regole che sono adeguatamente ottimali da essere soddisfatte su un totale di regole. Per esempio, se ho un basso numero di regole soddisfatte rispetto al numero totale di regole, significa che ho un gran numero di regole che alla fine non riescono a coprire il dataset (non sono applicabili). Di conseguenza, l'accuratezza sarà bassa. Nella tabella vediamo che il valore di copertura è del 40% perché ci sono 4 casi single su un totale di 10 persone. Invece, il valore dell'accuratezza è del 50% perché tra i 4 casi singoli abbiamo due classificati come 'no' e due classificati come 'yes'.



Come funzionano le regole? Sul mio test set, una regola può attivare un dato. In questo caso, se verifico le mie regole e vado a vedere le istanze, trovo che le regole attivano delle istanze, o che alcune istanze rispondono alle regole. Ad esempio, verifico se una stessa istanza può essere influenzata da più regole.



Regole mutuamente esclusive: un insieme di regole R è detto mutuamente esclusivo se nessuna coppia di regole può essere attivata dallo stesso record

- ✓ Ogni record è coperto al più da una regola

Regole esaustive: un insieme di regole R ha una copertura esaustiva se esiste una regola per ogni combinazione di valori degli attributi

- ✓ Ogni record è coperto da almeno una regola

**Non sempre è possibile determinare un insieme di regole esaustive e mutualmente esclusive.

Mancanza di mutua esclusività

Un record può attivare più regole dando vita a classificazioni alternative

Soluzione:

- ✓ Definire un ordine di attivazione delle regole, si parla in questo caso di liste di decisione.
- ✓ Assegnare il record alla classe per la quale vengono attivate più regole (voto)

Mancanza di esaustività

Un record può non attivare nessuna regola

Soluzione:

- ✓ Utilizzare una classe di default ("altro") a cui il record viene associato in caso di non attivazione delle regole

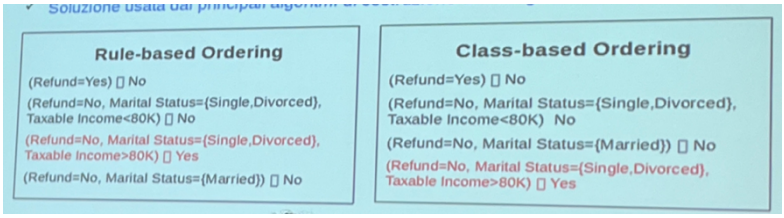
Regole ordinate

- Le regole sono ordinate secondo una data priorita

*un insieme ordinato di regole è anche chiamato lista di decisione

- Quando un record è sottomesso al classificatore
  *è assegnato alla classe della prima regola che viene attivata
  *se nessuna regola è attivata, è assegnato a una classe di default

Modalita di ordinamento

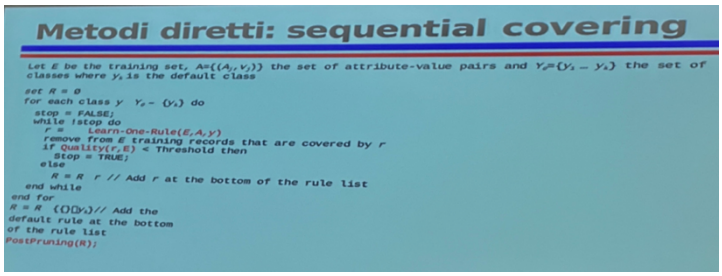


- **Ordinamento Rule-based:** le singole regole sono ordinate in base alla loro qualità
- **Ordinamento Class-based:** gruppi di regole che determinano la stessa classe compaiono di seguito nella lista.

L'ordinamento rilevante diventa quello tra le classi che può dipendere dall'importanza della classe o dalla gravità di commettere un errore di classificazione per quella classe.
Il rischio con questa soluzione è che una regola di buona qualità sia superata da una regola di qualità inferiore ma appartenente a una classe importante. Tale soluzione è usata dai principali algoritmi di costruzione delle regole (RIPPER, C4.5 rules)

Costruzione delle regole

- *Metodi diretti:* estraggono le regole direttamente dai dati (es. RIPPER, CN2,Holte's 1R)
- *Metodi indiretti*: estraggono le regole dal risultato di altri metodi di classificazione come, ad esempio, i decision tree (C4.5 rules)



Metodo Diretto: Sequential Covering

- Si inizia da una regola vuota
- Implementa la regola usando la funzione «Learn-One-Rule»
- Rimuovi i record di training coperti dalla regola
- Ripeti i passaggi (2) e (3) fino a quando il criterio di arresto non viene soddisfatto

L'ordinamento delle regole è determinato dalle classi. Punti di attenzione:

1) Modalità di creazione delle regole (Learn-One-Rule)
2) Eliminazione delle istanze
3) Criterio di stop

Se rappresento le classi come positive e negative, il meccanismo di definizione delle regole consiste principalmente nell'identificare le zone di appartenenza. Questo esempio è interessante perché, come ha affermato il professore all'inizio della lezione, se volessi prendere uno spazio e rappresentare le classi, in questo caso due classi (positiva e negativa), avendo la distribuzione del mio oggetto in uno spazio tridimensionale, che è piuttosto evidente con una nuvola di punti, la rappresentazione con un albero binario risulta utile quando i dati sono distribuiti geometricamente in modo netto. Tuttavia, l'albero decisionale potrebbe non riuscire a essere sufficientemente discriminante nonostante l'addestramento.