Coding e Big Data

2024-2025

Vincenzo Nardelli

vincenzo.nardelli@unicatt.it

Cos'è un Albero di Classificazione

- Un albero di classificazione è un modello di apprendimento automatico usato per classificare osservazioni in categorie (classi).
- Costruisce una serie di domande (split) su variabili indipendenti per predire una variabile target.
- Ogni nodo rappresenta una domanda, ogni ramo una risposta, e ogni foglia (nodo terminale) una previsione di classe.

Differenza con Algoritmo Deterministico

- Un algoritmo deterministico produce sempre lo stesso output per un dato input (es. una funzione matematica).
- Gli alberi di classificazione non sono deterministici: le divisioni (split) possono cambiare a seconda del dataset, del campionamento e di altri fattori.
- Gli alberi di classificazione si basano su criteri probabilistici e la loro struttura può cambiare con un nuovo campione di dati.

Deviance: Definizione e Formula

- La deviance misura l'eterogeneità dei dati in un nodo dell'albero di classificazione. Una deviance più bassa indica un nodo più omogeneo rispetto alla variabile target.
- Formula della deviance (classificazione binaria):

$$\text{Deviance} = -2 \sum_{i=1}^{n} \left(y_i \log(p) + (1 - y_i) \log(1 - p) \right)$$

Dove:

- y_i è la classe dell'osservazione (1 per la classe positiva, 0 per la negativa),
- p è la probabilità stimata di appartenere alla classe positiva,
- n è il numero di osservazioni.

Gradi di Libertà

- I gradi di libertà rappresentano il numero di valori indipendenti che possono variare in un'analisi statistica, dato un certo numero di vincoli.
- Nel contesto della deviance residua di un modello, i gradi di libertà si calcolano come:

  $$\text{Gradi di libertà} = n - p$$

  dove:

  - n è il numero totale di osservazioni,
  - p è il numero di parametri stimati (o nodi terminali nell'albero di classificazione).
- Un numero maggiore di gradi di libertà indica più libertà nell'adattamento del modello ai dati.

Esempio di Riduzione della Deviance con uno Split

- Consideriamo un nodo con 100 osservazioni, di cui 40 appartengono alla classe "Yes" e 60 alla classe "No".
- Deviance iniziale del nodo:

$$\text{Deviance} = -2\left(40 \cdot \log(0.4) + 60 \cdot \log(0.6)\right) \approx 91.61$$

Esempio di Riduzione della Deviance con uno Split

- Consideriamo un nodo con 100 osservazioni, di cui 40 appartengono alla classe "Yes" e 60 alla classe "No".
- Deviance iniziale del nodo:

$$\text{Deviance} = -2\left(40 \cdot \log(0.4) + 60 \cdot \log(0.6)\right) \approx 91.61$$

- Effettuiamo uno split in base a una variabile, creando due nuovi nodi:
  - Nodo 1: 30 osservazioni (10 Yes, 20 No)
    Deviance $= -2(10 \cdot \log(0.33) + 20 \cdot \log(0.67)) \approx 31.62$
  - Nodo 2: 70 osservazioni (30 Yes, 40 No)
    Deviance $= -2(30 \cdot \log(0.43) + 40 \cdot \log(0.57)) \approx 51.35$
- Deviance totale dopo lo split: $31.62 + 51.35 = 82.97$
- La deviance si è ridotta da 91.61 a 82.97, segnalando un miglioramento nell'omogeneità dei nodi.

Esempio di Output di un Albero di Classificazione

- Nodo radice: Rappresenta l'intero dataset, con la probabilità per ciascuna classe.
- Split: Condizione che divide il dataset in due sottoinsiemi.
- Nodo terminale: Nodo senza ulteriori divisioni, dove si assegna una classe.

Interpretazione Output

`1) root 400 541.50 No ( 0.5900 0.4100 )`
- Nodo radice con 400 osservazioni, deviance = 541.5. Classe predetta: `No` (59% No, 41% Yes).

Interpretazione della Deviance

- La deviance misura l'eterogeneità dei dati nel nodo.
- La deviance del nodo radice è alta poiché contiene l'intero dataset.
- La deviance si riduce con ogni split, creando nodi più omogenei.
- Deviance residua media (1.102) = deviance totale (435.2) / gradi di libertà (395).

Misclassification Error Rate

- Indica la percentuale di osservazioni classificate in modo errato dal modello.
- Nell'esempio:
  Misclassification error rate = $\frac{105}{400}$ = 0.2625 (26.25%).
- Precisione del modello: 73.75%, con 295 osservazioni classificate correttamente.

Confusion Matrix

- La matrice di confusione aiuta a valutare le prestazioni del modello:

	No	Yes
No	170	39
Yes	66	125

- True Positives (Yes predetto correttamente): 125
- True Negatives (No predetto correttamente): 170
- False Positives: 66, False Negatives: 39

Conclusioni

- Gli alberi di classificazione sono modelli interpretabili, che permettono di prevedere la classe di un'osservazione basandosi su variabili esplicative.
- Il tasso di errore e la deviance sono metriche chiave per valutare l'efficacia del modello.
- Il modello ha un'accuratezza del 73.75%, con un buon equilibrio tra le classi predette.