

Alberi Decisionali

Fondamenti teorici e applicazione a un problema di classificazione

Prof. Fedeli Massimo IIS Fermi Sacconi Cipa

1 Introduzione

Gli alberi decisionali sono modelli di apprendimento supervisionato ampiamente utilizzati in ambito di classificazione e regressione. La loro diffusione è dovuta principalmente alla semplicità concettuale, all'elevata interpretabilità e alla capacità di modellare relazioni non lineari tra le variabili.

In questo documento vengono illustrati:

- il funzionamento teorico degli alberi decisionali;
- i criteri di costruzione del modello;
- l'applicazione concreta a un problema di classificazione delle richieste di assistenza tecnica.

2 L'albero decisionale

Un albero decisionale è una struttura gerarchica composta da:

- un nodo radice;
- nodi interni di decisione;
- nodi foglia che rappresentano l'output del modello.

Ogni nodo interno effettua un test su una variabile di input, mentre ogni ramo rappresenta l'esito possibile del test. Il percorso dalla radice a una foglia costituisce una regola di decisione.

3 Apprendimento supervisionato

Nel contesto dell'apprendimento supervisionato, l'albero decisionale viene addestrato su un insieme di esempi etichettati:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

dove:

- x_i è il vettore delle caratteristiche (feature);

- y_i è la classe associata.

L'obiettivo dell'algoritmo è apprendere una funzione:

$$f : X \rightarrow Y$$

che consenta di predire correttamente la classe di nuovi esempi non visti.

4 Criteri di suddivisione

Durante la costruzione dell'albero, l'algoritmo seleziona a ogni nodo la variabile che permette la migliore separazione delle classi.

I criteri più comuni sono:

- Indice di Gini;
- Entropia e Information Gain.

Nel caso dell'indice di Gini, l'impurità di un nodo è definita come:

$$G = 1 - \sum_{i=1}^k p_i^2$$

dove p_i è la probabilità della classe i nel nodo.

L'obiettivo è minimizzare l'impurità dei nodi figli.

5 Costruzione e arresto dell'albero

La crescita dell'albero può proseguire fino a ottenere nodi puri, ma questo porta spesso a overfitting. Per questo motivo vengono introdotti criteri di arresto, tra cui:

- profondità massima dell'albero;
- numero minimo di campioni per nodo;
- riduzione minima dell'impurità.

Nel nostro esercizio è stata impostata una profondità massima per favorire l'interpretabilità del modello.

6 Descrizione del problema applicativo

Il problema affrontato consiste nella classificazione automatica delle richieste di assistenza tecnica in base alla loro priorità.

Ogni richiesta è descritta dalle seguenti variabili:

- Tipo_problema (software, hardware, rete);
- Numero_utenti_coinvolti (1, 2–5, > 5);
- Impatto_servizio (basso, medio, alto);

- Urgenza_ dichiarata (bassa, media, alta).

La variabile target è:

- Priorità (bassa, media, alta).

Il problema è quindi un problema di classificazione multiclasse.

7 Preparazione dei dati

Poiché gli alberi decisionali implementati nelle librerie di Machine Learning lavorano su dati numerici, le variabili categoriche vengono codificate mediante tecniche di encoding.

Nel programma Python sviluppato:

- a ciascuna variabile categoriale viene associato un encoder;
- gli encoder vengono salvati insieme al modello;
- in fase di previsione viene applicata esclusivamente la trasformazione.

Questo garantisce coerenza tra fase di addestramento e fase di inferenza.

8 Addestramento del modello

Il dataset viene suddiviso in:

- insieme di addestramento;
- insieme di test.

L'albero decisionale viene addestrato sull'insieme di training e valutato sul test set tramite l'accuratezza.

Il modello appreso rappresenta una collezione di regole decisionali interpretabili, ad esempio:

Se l'impatto è alto e l'urgenza è alta, allora la priorità è alta.

9 Fase di previsione

In fase di utilizzo, l'utente inserisce i dati di una nuova richiesta. Il sistema:

1. codifica gli input tramite gli encoder salvati;
2. applica il modello addestrato;
3. restituisce la priorità prevista in forma testuale.

Questo separa chiaramente la fase di apprendimento dalla fase di utilizzo operativo.

10 Vantaggi e limiti

I principali vantaggi degli alberi decisionali sono:

- elevata interpretabilità;
- semplicità concettuale;
- assenza di assunzioni forti sui dati.

I principali limiti includono:

- tendenza all'overfitting;
- instabilità rispetto a piccole variazioni dei dati;
- prestazioni inferiori rispetto a modelli ensemble in problemi complessi.

11 Conclusioni

L'albero decisionale rappresenta una soluzione efficace e didatticamente significativa per il problema analizzato. L'esercizio consente di collegare teoria, implementazione e interpretazione del modello, fornendo una visione completa del ciclo di vita di un sistema di Machine Learning supervisionato.