UNIVERSITÀ DEGLI STUDI DI SALERNO

Project Work

ADVANCED STATISTICAL LEARNING I

Analisi Comparativa di Bagging, Boosting e Random Forest attraverso Simulazione Monte Carlo

Studentessa:

Carmela Pia Senatore

Matr. 0522501721

Anno Accademico 2023/2024

Analisi Comparativa di Bagging, Boosting e Random Forest attraverso Simulazione Monte Carlo

Project Work- Carmela Pia Senatore

1. Introduzione

Lo scopo di questo project work è valutare le performance di tre differenti classificatori, ovvero bagging, boosting e random forest, mediante l'uso del metodo Monte Carlo. I modelli saranno confrontati utilizzando un dataset simulato attraverso la funzione *simulate_data* in R, e i risultati saranno analizzati in termini di tasso di errore e Monte Carlo Standard Error (MCSE).

2. Definizione del dataset

Il dataset simulato consiste in una variabile di outcome binaria (Y) e variabili predittive continue e categoriche. La simulazione è stata effettuata tenendo conto dello schema di campionamento e generando per alcune variabili campioni multivariati normali con medie specificate per ciascuna classe di Y. La funzione **simulate_data**, (analizzata nello specifico in R) definita sulla base del modello descritto all'interno della traccia, permette di campionare i dati come richiesto. Questa prende in input parametri obbligatori o opzionali:

- *N*, obbligatorio, numero di osservazioni da campionare;
- *Media_Y_A, Media_Y_B, Sig,* parametri opzionali che rappresentano rispettivamente: il vettore medio relativo al gruppo A per la variabile Y per le variabili (X_1, X_2, X_3); il vettore medio relativo al gruppo B per la variabile Y per le variabili (X_1, X_2, X_3); la matrice di varianza e covarianza.

Dalle specifiche fornite, possiamo anticipare alcune caratteristiche chiave che ci aspettiamo dall'implementazione dello schema di campionamento:

- Poiché è stata specificata una ***dimensione del dataset*** di n = 250, ci aspettiamo che il dataset simulato contenga 250 osservazioni.

-***VARIABILE Y:*** Poiché la variabile di risposta Y è campionata casualmente da un insieme {A, B}, ci aspettiamo una distribuzione uniforme di Y all'interno del dataset

simulato o quantomeno una frequenza abbastanza equilibrata tra le categorie “A” e “B”. Chiaramente è possibile modificare lo sbilanciamento.

-***VARIABILI*** $\boldsymbol{X_1, X_2, X_3}$ **:** Le variabili sono delle variabili continue con distribuzione congiunta normale , questo implica che: le variabili $X_1, X_2\ e\ X_3$ formano un insieme di variabili casuali che sono continue (cioè, possono assumere un numero infinito di valori all'interno di un intervallo). Nel contesto applicativo questo implica che sono stocasticamente legate tra loro in modo tale che la loro distribuzione complessiva è definita come una distribuzione normale multivariata. Inoltre, per costruzione, ci aspettiamo di vedere differenze statisticamente significative nei valori medi di $X_1, X_2\ e\ X_3$ tra le categorie A e B di Y. Questo è dovuto al fatto che la distribuzione congiunta di queste variabili ha vettori medi diversi per le due categorie di Y. La matrice di covarianza è semidefinita positiva e simmetrica oltreché diagonale indicando che le variabili sono incorrelate.

-***VARIABILI*** $\boldsymbol{X_4, X_5}$**:** Le variabili categoriche non numeriche $X_4\ e\ X_5$ sono campionate casualmente da insiemi di categorie discrete. Ci aspettiamo quindi una diversità di categorie all'interno di ciascuna variabile. Oltreché queste variabili non dipendono da Y, per cui ci aspettiamo che le loro distribuzioni siano simili tra le categorie di Y.

-***VARIABILI*** $\boldsymbol{X_6, X_7, X_8, X_9, X_{10}}$: Le variabili numeriche, sia discrete che continue, sono generate casualmente. Ci aspettiamo quindi una varietà di valori all'interno di ciascuna di queste variabili. Poiché queste variabili numeriche non dipendono da Y, ci aspettiamo che le loro distribuzioni siano simili tra le categorie di Y. Tuttavia, le statistiche descrittive (media, mediana, deviazione standard, ecc.) potrebbero differire in base alla natura delle variabili.

In generale, la simulazione di dati iid implica che:

1. Campioni identicamente distribuiti: ciascuna osservazione nel campione è estratta dalla stessa distribuzione di probabilità. Non importa quale osservazione è stata estratta in precedenza; ogni osservazione ha la stessa probabilità di essere estratta dalla distribuzione.

2. Indipendenti: Le osservazioni nel campione sono indipendenti l'una dall'altra. La realizzazione di un'osservazione non influisce sulla probabilità di realizzare un'altra osservazione.

Quando si campionano campioni iid da questo schema di campionamento, ci aspettiamo che le osservazioni nel campione siano rappresentative della popolazione sottostante; ovvero, il campione dovrebbe riflettere accuratamente la variabilità e la struttura della popolazione da cui è stato estratto. Questo è essenziale per garantire che le conclusioni e le previsioni basate sul campione siano generalizzabili alla popolazione di cui non si ha conoscenza.

3. ANALISI MONTE CARLO

a. Definizione del metodo

L'obiettivo della simulazione è confrontare le prestazioni dei diversi modelli attraverso una serie di possibili realizzazioni del fenomeno in esame, utilizzando poi la distribuzione dei risultati così ottenuti. Utilizzando repliche Monte Carlo, possiamo tener conto della variabilità intrinseca nei dati e stimare l'incertezza nelle stime delle prestazioni del modello. Per quel che concerne l'analisi Monte Carlo sono state analizzate le performance, in termini di tasso di errore, per i seguenti tre classificatori:

1. Bagging: fa parte dei metodi ensemble e ha l'obiettivo principale di ridurre l'instabilità del previsore associata alla definizione intrinseca dei classification trees. Il bagging permette di ridurre la varianza del previsore basandosi sul boostrap esplorando le diverse variabilità campionarie e aggregando tutti i previsori facendo media sulle stranezze.
2. Random forest: è un bagging volutamente reso distorto per risolvere la problematica legata al fatto che nel bagging le previsioni vengono fatte sullo stesso dataset. In particolar modo rf si basa sull'esplorazione dei diversi dataset settando un numero q di features producendo alberi completamente diversi. Le previsioni vengono poi aggregate facendo media.
3. Boosting: si diversifica rispetto ai due classificatori appena enunciati in quanto lavora esclusivamente sullo stesso dataset utilizzando dei sistemi di pesaggio che permettono al classificatore implementato di apprendere dagli errori svolti.

Si giustifica l'utilizzo dell'analisi Monte Carlo in quanto:

1. I dataset reali possono essere afflitti da variabilità. La performance di un modello può dipendere dalla specifica suddivisione del dataset in training e test set. Eseguendo il modello su un singolo split potrebbe non fornire una rappresentazione completa delle sue performance. L'utilizzo di simulazioni Monte Carlo, che coinvolgono la generazione di molteplici dataset di training e test in modo casuale generando tanti insiemi simili, che vanno, per così dire, a approssimare la popolazione, consente di valutare le prestazioni del modello in diverse configurazioni di dati.
2. La simulazione Monte Carlo consente di stimare la variabilità nei risultati. Calcolando il Monte Carlo standard error (MCSE), è possibile ottenere una misura della precisione delle stime del modello.
3. Ancora, permette di ottenere stime dell'aspettativa e della varianza del tasso di errore. Queste stime possono essere utili per comprendere non solo la media delle performance del modello, ma anche quanto tali performance variano in diverse situazioni.

Lo scenario applicativo è stato definito in questo modo:

1. Riproducibilità

L'algoritmo inizia impostando un seed per garantire che la simulazione sia riproducibile. Questo significa che, in caso di esecuzione più volte, si otterranno sempre gli stessi risultati.

2. Numero di Repliche Monte Carlo (R):

Viene definito il numero di repliche Monte Carlo, indicato con R, che rappresenta quante volte verranno eseguiti i passaggi successivi per ottenere stime più robuste delle performance dei modelli. In particolare è stato fissato a 50, 100 , 300.

3. Inizializzazione dei Vettori per i Risultati:

Vengono inizializzati vettori vuoti per salvare i risultati dei tassi di errore per ciascun metodo di classificazione (Bagging, Random Forest, Boosting).

4. Simulazione Monte Carlo:

Vengono generati i dati della replica utilizzando la funzione simulate_data. Il dataset viene formattato correttamente, convertendo la variabile di outcome Y.

Per ciascuna delle R repliche Monte Carlo:

a. Si effettua la suddivisione del dataset in set di addestramento e test: il training set contiene l'80% delle osservazioni e il test set il restante 20%.
b. I dati vengono adattati per differenti formattazioni della variabile di outcome Y.
c. Per ciascun metodo di classificazione (Bagging, Random Forest, Boosting):
    1. Si addestra il modello utilizzando il set di addestramento.
    2. Si effettuano previsioni sul set di test.
    3. Si calcola il tasso di errore e si salva nei vettori dei risultati.

5. Calcolo delle Statistiche di Interesse:

Dopo la simulazione Monte Carlo, vengono calcolati il valore atteso del tasso di errore e il corrispondente Monte Carlo Standard Error (MCSE) per ciascun metodo. Queste statistiche offrono la stima delle performance medie e la loro variabilità su diverse repliche.

6. Stampa dei Risultati:

I risultati vengono quindi stampati a video e in una tabella.

L'algoritmo per l'esperimento è articolato come segue:

• Generazione di un set di dati di dimensione *N* (N = 250);

• Ciascuno step definito precedentemente verrà ripetuto per un numero di repliche pari a *R* (R=50, 100, 300).

b. Analisi dei risultati

Innanzitutto, è stato calcolato il valore atteso del tasso di errore per ciascun classificatore. Questa metrica indica la media teorica degli errori commessi dai modelli su nuovi dati. La scelta di stimare l'errore attraverso la media degli errori è sostenuta dalla teoria statistica, in quanto la legge dei grandi numeri ci assicura che all'aumentare del numero di campioni, questa stima convergerà al valore atteso teorico del tasso di errore. Pertanto, calcolare la media degli errori su molte repliche fornisce un'approssimazione affidabile del valore atteso del tasso di errore. Nell'ambito dell' analisi Monte Carlo, si è proceduto simulando ripetutamente campioni da un modello complesso, introducendo così la variabilità intrinseca al fenomeno che si sta cercando di modellare. Attraverso

questo processo, si è stati in grado di ottenere stime del tasso di errore per ciascuna replica Monte Carlo. In *tabella 1* sono riportati i valori dell'approssimazione del valore atteso del tasso di errore per il processo basilare di 250 osservazioni e 50 repliche Montecarlo. Si è osservato che il Bagging ha raggiunto una stima del valore atteso del tasso di errore del 8.28% suggerendo che l'aggregazione di modelli addestrati su sottoinsiemi casuali del dataset ha contribuito a ridurre la varianza e migliorare la stabilità, mentre Random Forest e Boosting hanno ottenuto rispettivamente il 9.0% 8.64%. Per cui, in media, Bagging mostra le prestazioni migliori in termini di errore. Si è ,quindi, valutato il Monte Carlo standard error, che rappresenta la dispersione o la variabilità nei risultati delle repliche Monte Carlo. Infatti, l'errore ottenuto è una quantità stocastica ed è utilizzato per valutarne la veridicità. Un errore standard più elevato indica una maggiore variabilità nei risultati delle repliche Monte Carlo. Notiamo che il Random Forest ha mostrato una variabilità dell'errore inferiore rispetto a Bagging e simile a quella del Boosting, indicando una maggiore coerenza nelle sue prestazioni.

Tabella 1: Stima errore e MCSE

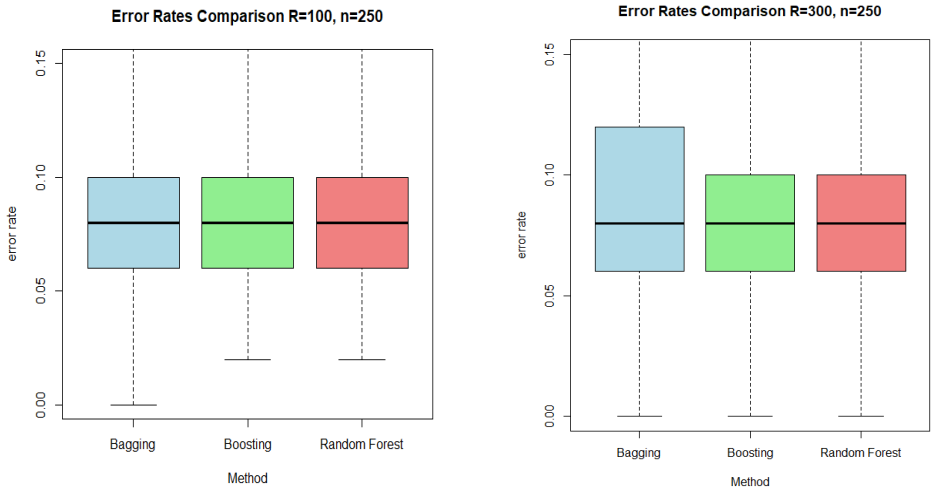
Classificatore	Stima dell'errore	MCSE
"Bagging"	0.0828	0.00481
"Random Forest"	0.09	0.00436
"Boosting"	0.0864	0.00486

Vengono forniti in aggiunta , nella *Figura 1* , i boxplot per le distribuzioni campionarie degli errori per i tre modelli ottenuti con il Monte Carlo. L'analisi dei grafici permette di interpretare con maggior grado di dettaglio i risultati della *tabella 1*. Il random forest è il classificatore la cui distribuzione degli errori è meno compatta, ma è anche l'unico che presenta molti valori con tassi di misclassificazione superiore al 15%, è l'espressione di medie aritmetiche quindi per un numero elevato di osservazioni sta entrando in definizione il teorema centrale del limite ovvero distribuzioni approssimativamente normali. Il Boosting è invece il classificatore il cui tasso medio d'errore è superiore rispetto agli altri presentando una evidente asimmetria negativa nella distribuzione.

Figura 1: Distribuzioni Monte Carlo relative agli errori in funzione del classificatore

Si è successivamente passato a indagare il comportamento del tasso di errore per un aumento del numero di repliche. Si valuta quindi numero di repliche uguali a (100, 300).

Figura 2: Distribuzioni Monte Carlo relative agli errori in funzione del classificatore per repliche N=(100,300)



Dalla *Figura 2* è evidente come al crescere del numero di repliche la distribuzione del tasso di errore diventa sempre più approssimativamente normale per tutti e tre i classificatori. Per numero di repliche pari a 300 la media del tasso di errore si setta quasi sullo stesso valore per tutti e tre i classificatori. Da 100 repliche in poi, boosting si distingue rispetto agli altri classificatori per stima del tasso di errore minore oltre che minor MCSE.

Tabella 2: Stima errore e MCSE per diverse Repliche

Classificatore	N Repliche	Stima dell'errore	MCSE
"Bagging"	100	0.0856	0.0034
"Random Forest"	100	0.086	0.0033
"Boosting"	100	0.0836	0.0033
"Bagging"	300	0.0874	0.0021
"Random Forest"	300	0.083	0.0020
"Boosting"	300	0.081	0.0019

4. CONCLUSIONI

Infine, è stato fatto un resoconto delle analisi. La *Figura 3* illustra graficamente le stime dell'errore medio per i tre diversi classificatori al variare del numero di repliche (R). Le linee rappresentano le tendenze delle stime dell'errore per ciascun classificatore, mentre le barre di errore indicano il Monte Carlo Standard Error (MCSE) associato a ciascuna stima. Per il classificatore "Bagging", si osserva un aumento graduale dell'errore medio all'aumentare del numero di repliche, con un MCSE che si riduce progressivamente : Il bagging crea diversi campioni di addestramento utilizzando il bootstrap e costruisce modelli indipendenti. Aumentare il numero di repliche potrebbe portare a una maggiore varietà nei modelli, ma a volte può anche causare overfitting al training set originale. Per quel che concerne il classificatore "Random Forest" mostra una diminuzione graduale del tasso di errore con stime dell'errore generalmente più basse. La motivazione è effettivamente giustificata dal fatto il Random Forest è una variazione del bagging che introduce ulteriore casualità creando alberi decisionali in modo casuale; l'aumento del numero di repliche può portare a una diminuzione dell'overfitting e ad una maggiore generalizzazione, riducendo così il tasso di errore medio. Per il classificatore "Boosting", si evidenzia una significativa diminuzione dell'errore medio all'aumentare del numero di repliche, con un MCSE che diminuisce drasticamente. Tenendo conto che il Boosting costruisce una sequenza di modelli in cui ogni modello cerca di correggere gli errori del modello precedente, questo può portare a una riduzione costante dell'errore poiché ogni nuovo modello è progettato per affrontare specifici casi difficili per il modello precedente.

In conclusione, l'andamento delle stime dell'errore con l'aumentare delle repliche suggerisce che l'incertezza nelle stime si riduce, fornendo una maggiore affidabilità nei

risultati predittivi dei modelli. Il confronto tra i tre classificatori indica che il nostro problema di classificazione beneficia particolarmente dal classificatore "Boosting" che si mostra con prestazioni leggermente superiori rispetto agli altri due, con una tendenza a mantenere un errore medio più basso.

Figura 3: Grafico conclusivo