DOMANDA 1

a)

Immagine che contiene testo, schermata, linea, diagramma

Descrizione generata automaticamente

L’algoritmo Decision Tree posiziona sempre l’attributo più selettivo nel nodo radice dell’albero, cioè nel nostro caso “node-caps”

b) L’altezza di un albero è determinata dalla lunghezza massima tra il nodo radice e un qualsiasi nodo foglia. L’albero generato da questo dataset è di altezza 6

c)

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Un partizionamento puro si verifica quando uno split su un valore di un attributo dell’albero produce come risultato una partizione in cui tutti gli elementi appartengono alla stessa classe. Nel nostro caso possiamo vedere come la partizione ottenuta dal nodo “breast” scegliendo il ramo dello split “left” sia un partizionamento puro poiché tutti i suoi elementi appartengono solo alla classe rossa

DOMANDA 2

Il valore del parametro “minimal-gain” stabilisce la soglia per cui l’albero deve continuare ad effettuare uno split nei nodi o meno. Uno split viene effettuato se il suo valore di “gain” è maggiore del valore del parametro “minimal-gain”. Questo parametro determina sia il numero di foglie dell’albero sia la sua altezza poiché, se viene impostato un valore basso di “minimal-gain” allora l’albero tenderà a splittare più nodi e sarà più profondo e più fitto, d’altra parte con alti valori di “minimal-gain” l’albero avrà meno foglie e sarà più basso. Invece il valore del parametro “maximal-depth” stabilisce il valore massimo di altezza che l’albero può raggiungere, se è minore dell’altezza dell’albero esso verrà tagliato.

Proviamo a variare i valori di “minimal-gain” e “maximal-depth” per vedere il risultato.

-Prima modifichiamo solo il valore di “minimal-gain” che di default è impostato ad 0.01. Per valori di poco più alti rispetto al default, es. 0.04, otterremo un albero meno fitto e alto rispetto a quello di default

Immagine che contiene linea, schermata, diagramma, testo

Descrizione generata automaticamente

- Mentre per valori molto più alti, es. 0.1, l’albero verrà compresso tutto in una foglia dato che non riesce a calcolare il gain necessario per splittarsi

Immagine che contiene testo, schermata, Carattere, Elementi grafici

Descrizione generata automaticamente

-Se invece impostiamo un valore minore di 0.01 non noteremo alcun cambiamento nell’albero generato

Immagine che contiene linea, testo, diagramma, Parallelo

Descrizione generata automaticamente

-Modifichiamo ora solo il valore di “maximal-depth”. Impostiamo prima il suo valore a 10, in modo tale che sia maggiore dell’altezza effettiva dell’albero e quindi possiamo osservare alcun cambiamento all’albero

Immagine che contiene linea, testo, diagramma, Parallelo

Descrizione generata automaticamente

-Mentre se si imposta un valore minore, es. 5, otterremo un taglio dell’albero che andrà ad eleminare tutti i nodi con altezza maggiore di 5

Immagine che contiene testo, linea, schermata, diagramma

Descrizione generata automaticamente

-Infine, modifichiamo entrambi i valori di “minimal-gain” e “maximal-depth”, impostandoli rispettivamente a: minimal-gain:0.04 e maximal-depth:4, in cui l’albero potrebbe avere più nodi in profondità poiché il minimal-gain lo permetterebbe ma il maximal-depth dli impedisce di avere un’altezza superiore a 4

Immagine che contiene schermata, testo, diagramma, linea

Descrizione generata automaticamente

DOMANDA 3

In generale, riducendo il valore del minimal gain e aumentando la maximal depth si genera un modello di classificazione più dettagliato e quindi più accurato. Tuttavia, sulla base dei risultati ottenuti in precedenza otteniamo in alcuni casi l’effetto denominato “overfitting”, ovvero il modello risulta troppo “focalizzato” sui dati di train per classificare in modo accurato nuovi dati di test.

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Minimal gain:0.1, maximal depth 10

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Minimal gain:0.04, maximal depth:10

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Minimal gain:0.01, maximal depth:5

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Minimal gain:0.01, maximal depth:7

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Minimal gain:0.04, maximal depth: 4

DOMANDA 4

Incrementando il valore di K, nell’algoritmo K-NN, il classificatore considera un numero maggiore di dati di train “vicini” al dato di test e quindi l’accuratezza media cresce: 73.77% con K=5, 74.51% con K=8, 75.20% con K=10, 70.26% con K=13, 73.79% con K=20. Considerando un numero molto elevato di record di train “vicini”, nel nostro caso quando K>10, la presenza di dati rumorosi comincia ad inficiare le performance di classificazione e dunque l’accuratezza media di classificazione diminuisce leggermente. L’accuratezza di Naïve Bayes risulta essere inferiore al K-NN con K=5, quindi per valori di K superiori a 5 conviene utilizzare l’algoritmo K-NN per ottenere un’accuratezza maggiore

K=5

**Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente**

K=8

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

K=10

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

K=13

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

K=20

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Mentre utilizzando Naïve Bayes

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

DOMANDA 5

La figura sottostante mostra la matrice di correlazione ottenuta dal dataset analizzato. Essa riporta la correlazione mutua e simmetrica tra coppie di attributi. La coppia di attributi “irradiant” e “inv-nodes” è la coppia col più alto valore di correlazione. L’ipotesi d’indipendenza Naïve risulta essere irrealistica per il dataset analizzato, in quanto i dati del dataset presentano una dipendenza tra di loro non trascurabile, e ciò non si può conciliare con l’ipotesi Naïve.

Immagine che contiene testo, schermata, quadrato, numero

Descrizione generata automaticamente