**Quaderno #2** *Data mining*

*Carlo MIGLIACCIO* matricola: 272489

1. **Decision tree** (**minimal gain**: 0.01, **maximal depth**: 10)

Immagine che contiene testo, bianco, algebra

Descrizione generata automaticamente

Figura - Albero generato con i parametri sopra (le foglie sono le etichette di classe)

* 1. L’attributo considerato dall’algoritmo il più selettivo ai fini dell’analisi è **node-caps** che è un booleano associato al fatto che il tumore resti arginato alla *capsula del linfonodo* (come si può leggere dalla descrizione in <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer>: *“Node caps: if the cancer does metastasise to a lymph node, although outside the original site of the tumor it may remain “contained” by the capsule of the lymph node. However, over time, and with more aggressive disease, the tumor may replace the lymph node and then penetrate the capsule, allowing it to invade the surrounding tissues“)*
  2. Come si può vedere dall’immagine seguente l’albero riportato al passo precedente ha **altezza** pari a 6.

Immagine che contiene linea

Descrizione generata automaticamente

Figura - Esempio di cammino radice-foglia

* 1. Di seguito si riporta un esempio di partizionamento **puro.** La purezza è associata al fatto che c’è una distribuzione netta delle etichette di classe. Esempio: (NA: 0, NB: 15), oppure (NA: 15, NB: 0) dove A e B sono le classi e NA e NB sono il numero di record associati alle rispettive classi.

Immagine che contiene testo, schermata, software, Icona del computer

Descrizione generata automaticamente

Figura - Esempio di partizione pura

Essendo il dataset di dimensione ridotta (circa 300 record) e avendo fatto pre-pruning dell’albero, impostando un minimal gain non si riescono ad osservare casi in cui tutte le foglie arrivano da partizioni pure.

1. **Decision tree**: impatto del **minimal gain** e del **maximal depth**

Si riportano di seguito alcune immagini che mostrano come i due parametri di minimal gain e maximal depth influiscono sulle caratteristiche dell’albero generato (numero di nodi, altezza...).

**minimal gain**: 0.01 **maximal depth**: 6

Immagine che contiene bianco, algebra

Descrizione generata automaticamente

Figura - Albero generato (mg=0.01, md=6)

**minimal gain**: 0.03 **maximal depth**: 10

Immagine che contiene linea, diagramma, appendiabiti

Descrizione generata automaticamente

Figura - Albero generato con mg=0.03, md=10)

**minimal gain**: 0.02 **maximal depth**: 5

Immagine che contiene linea, diagramma, Diagramma, Carattere

Descrizione generata automaticamente

Figura - Albero generato con mg=0.02, md=5

**minimal gain**: 0.03 **maximal depth**: 5

Immagine che contiene linea, diagramma

Descrizione generata automaticamente

Figura - Albero generato con mg=0.03, md=5

**minimal gain**: 0.03 **maximal depth**: 4

Immagine che contiene linea, diagramma, design

Descrizione generata automaticamente

Figura - Albero generato con mg=0.03, md=4

Nelle immagini riportate prima per praticità e chiarezza di visualizzazioe si sono state nascoste le etichette agli archi e ai nodi. Un esempio di albero completo di queste informazioni in precedenza omesse è riportato di seguito:

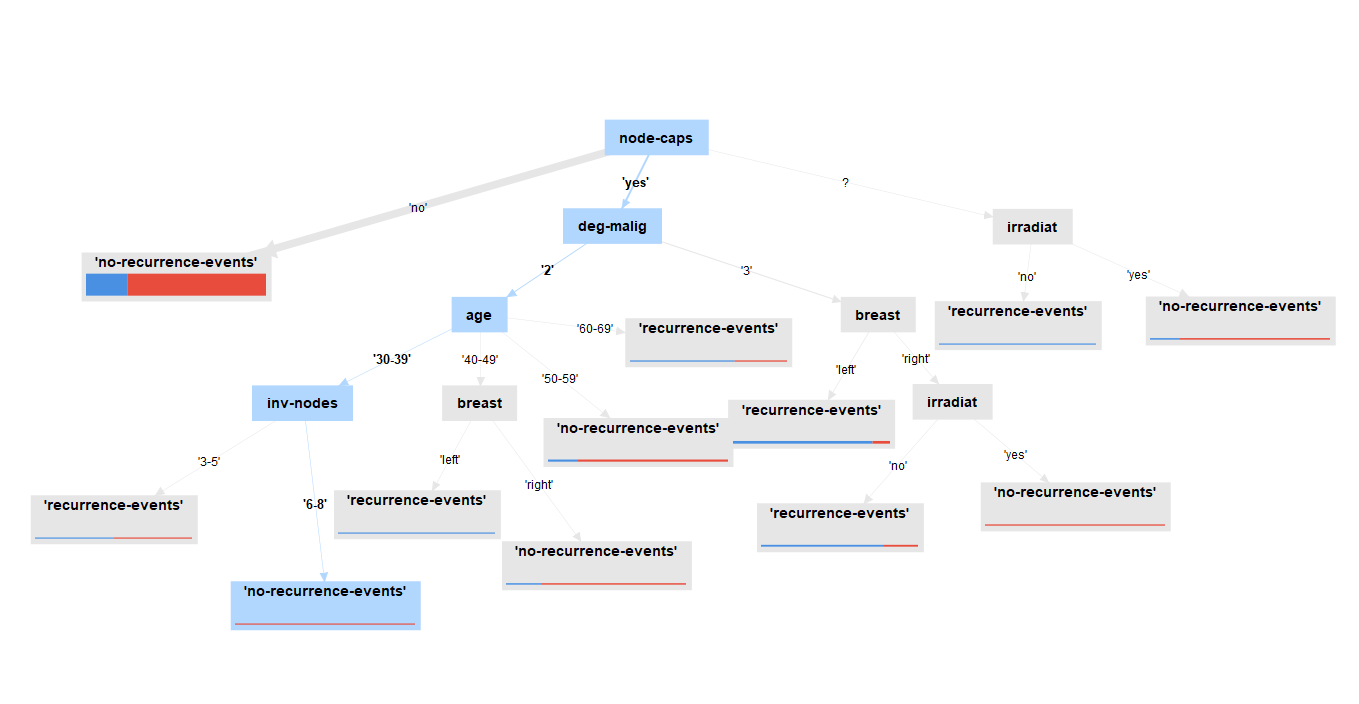


Figura - Albero generato con mg=0.04, md=5 (completo di tutte le informazioni)

Ad ogni passo di un cammino (vedi nodi in azzurro) che porta dalla radice (**node-caps**) fino alla foglia (classe **no-recurrence-events**) si fa una scelta sul valore di un certo attributo.

1. **Decision tree:** validazione del modello tramite *10-fold Stratified Cross-Validation*

La tecnica di validazione ***k-fold Stratified Cross-Validation*** si applica soprattutto quando si hanno a disposizione dataset di dimensione ridotta. I record disponibili vengono divisi in **k insieme disgiunti** attraverso un *campionamento stratificato*[[1]](#footnote-1), successivamente, k-1 insiemi vengono usati per addestrare il modello (training set) mentre la partizione rimanente viene utilizzata per la verifica del modello stesso (test set). Nel caso analizzato k=10.

Immagine che contiene bianco, algebra

Descrizione generata automaticamenteIl tuning di entrambi i parametri, minimal gain e maximal depth, che danno vincoli sulla crescita dell’albero, influenzano le metriche di performance ricavabili dalle matrici di confusione. In particolare un modello troppo specifico, con minimal gain basso e profondità elevata tende ad essere affetto dal problema dell’overfitting, si ha quindi in generale una carenza sulle performance di generalizzazione del modello risultante.

**minimal gain**: 0.01 **maximal depth**: 6

Immagine che contiene linea, diagramma, appendiabiti

Descrizione generata automaticamenteImmagine che contiene testo, Carattere, linea, numero

Descrizione generata automaticamente

Tabella - Matrice di confusione (parametri albero mg 0.01 - md 6)

**minimal gain**: 0.03 **maximal depth**: 10

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Tabella - Matrice di confusione (parametri albero mg 0.03 - md 10)

Immagine che contiene linea, diagramma, Diagramma, Carattere

Descrizione generata automaticamente**minimal gain**: 0.02 **maximal depth**: 5

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Immagine che contiene linea, diagramma

Descrizione generata automaticamenteTabella - Matrice di confusione (parametri dell'albero mg 0.02 - md 5)

**minimal gain**: 0.03 **maximal depth**: 5

Immagine che contiene linea, diagramma, design

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Tabella - Matrice di confusione (parametri dell'albero mg 0.03 - md 5)

**minimal gain**: 0.03 **maximal depth**: 4

Immagine che contiene testo, Carattere, linea, numero

Descrizione generata automaticamente

Tabella - Matrice di confusione (parametri dell'albero mg 0.03 - md 4)

1. **K-Nearest Neighbour (KNN) e Naïve Bayes:** validazione e confronto
   1. ***KNN: 10-fold Stratified Cross validation, tuning del parametro K***

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamenteCome mostrano le matrici di confusione ottenute, facendo variare il parametro K da 2 a 10 l’accuratezza media del modello cresce, da 11 in poi comincia a diminuire in quanto probabilmente sono inclusi punti nell’intorno con etichetta di classe diversa da quella effettiva, che inficiano quindi la bontà della classificazione.

**K=2**

Immagine che contiene testo, Carattere, numero, linea

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente**K=6**

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente**K=8**

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente**K=10**

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente**K=13**

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

* 1. **Naïve Bayes*: 10-fold Stratified Cross validation, confronto con KNN***

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Il classificatore *Naïve Bayes* ottiene prestazioni inferiori al KNN (per certi parametri di K), ma migliori rispetto alle prestazioni ottenute con alberi decisionali con i parametri scelti. Si ricorda che questo tipo di classificatore ha alla base il Teorema di Bayes e un’ipotesi (molto forte) di indipendenza delle varie features del dataset. Al passo seguente si valuta quanto corretta può essere questa ipotesi.

1. **Matrice di correlazione**

Immagine che contiene schermata, testo, quadrato, Rettangolo

Descrizione generata automaticamenteQuella riportata a fianco è la **matrice di correlazione** ricavata dal dataset analizzato. L’ipotesi di indipendenza tra gli attributi, utilizzata nel Naive Bayes, non risulta valida soprattutto per alcune coppie di feature che mostrano una tendenza di correlazione, in particolare correlazione negativa.

La coppia di attributi **maggiormente (negativamente) correlata** è

(node-caps, inv-nodes).

Nella matrice sono stati inoltre riquadrati altri valori di correlazione significativi.

1. Con questa tecnica di campionamento si assicura che ogni campione abbia le etichette di classe distribuite in modo equo [↑](#footnote-ref-1)