Metodi di classificazione basati su alberi

Leonardo Michi

Laurea in Statistica

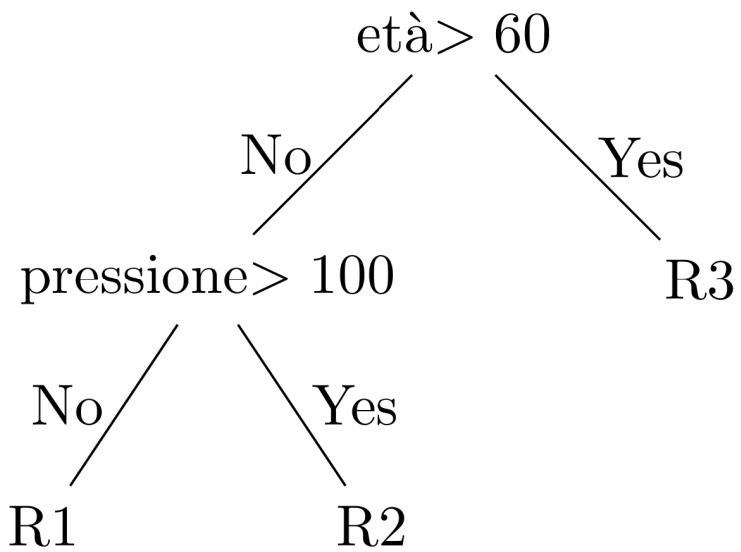
A.A. 2022/2023

Alberi decisionali

Gli alberi decisionali sono un particolare tipo di modelli predittivi non lineari. L'idea è quella di considerare una partizione ricorsiva dello spazio dei predittori e applicare un *modello semplice* per ogni regione della partizione.La partizione ricorsiva è rappresentata attraverso un albero.

Gli alberi di classificazione sono metodi che partizionano lo spazio dei predittori in regioni disgiunte e classificano le osservazioni sulla base della regione in cui l'osservazione cade.

Un esempio di albero decisionale



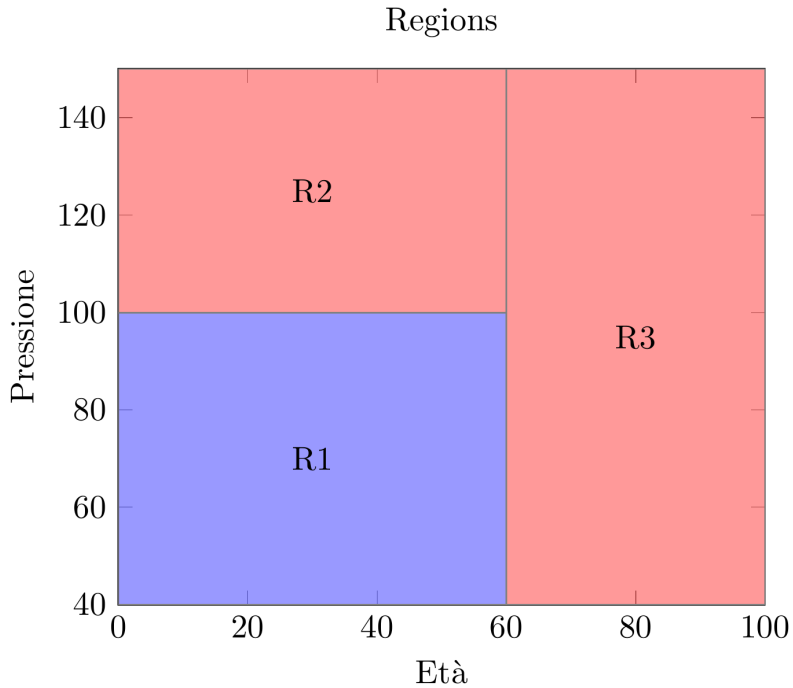


Figura 1: Esempio di albero di classificazione sinistra e partizione risultante destra

Alberi di classificazione

Dato il *training set* $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)$, con $\boldsymbol{x}_i \in \mathbb{R}^p$ e $y_i \in \{1, \ldots, K\}$, per un generico valore $\boldsymbol{x} = (x_1, \ldots, x_p)$ dello spazio dei predittori e un valore soglia s si individua la partizione binaria

$$R_1 = \{\boldsymbol{x} | x_j < s\}, \quad R_2 = \{\boldsymbol{x} | x_j > s\}$$

e si ripete il processo sulle sotto regioni finché non è soddisfatto lo *stopping criterion*.

Individuate le regioni nodi terminali $R_1, \ldots, R_M$, ciascuna contente n_m osservazioni, $m \in \{1, \ldots, M\}$,si ottiene la stima della probabilità della classe k

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{\boldsymbol{x}_i \in R_m} \mathbb{1}_{\{y_i = k\}}$$

e, in un problema di classificazione con K classi, si definisce la classe

$$k(m) = \underset{k \in \{1, \ldots, K\}}{\text{argmax}} \ \hat{p}_{mk}.$$

Misure di impurità

La partizione binaria è definita minimizzando una misura di *impurità*. Possibili misure di impurità sono:

- *Mis-classification error*

$$E = 1 - \max_k (\hat{p}_{mk})$$

- Indice di *Gini* :

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- Indice di *Entropia*:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Tree pruning

Le procedure di tree pruning hanno lo scopo di evitare l'overfitting. Considerando l'albero più grande T_0 il tree pruning mira a individuare dei sotto alberi adeguatamente selezionati. L'approccio che si basa sulla complessità dei costi, si sceglie l'albero $T \subset T_0$ tale che sia minima la quantità:

$$C_\alpha(T) = \sum_{m=1}^{|T|} n_m Q_m(T) + \alpha |T|$$

Dove:

- $|T|$ è il numero di nodi terminali in T
- $Q_m(T) = 1 - \frac{1}{n_m} \sum_{\boldsymbol{x}_i \in R_m} \mathbb{1}_{\{y_i = k(m)\}}$
- α è un parametro di tuning da scegliere opportunamente.

Bagging e Random Forest

Il bagging o bootstrap aggregation consiste in una procedura generale che consente di migliorare l'accuratezza della classificazione rispetto a un singolo albero. A partire dal training set, si ottengono B campioni bootstrap e su ciascuno di essi si costruisce un albero di classificazione. Indicata con $p_k(\boldsymbol{x})$ la frazione di *alberi bagging* che classificano $\boldsymbol{x}$ nella classe k, la classificazione finale si ottiene come:

$$\hat{k}_{\text{bag}}(\boldsymbol{x}) = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \{p_1(\boldsymbol{x}), \ldots, p_K(\boldsymbol{x})\}$$

In presenza di predittori forti il bagging potrebbe non fornire un miglioramento sostanziale dell'accuratezza. Le random forest rappresentano una variante del bagging in cui ad ogni split si considera un sottoinsieme di $q < p$ variabili ad ogni split. (usualmente $q = \sqrt{p}$)

Boosting

Il Boosting è una tecnica che si basa sulla costruzione *sequenziale* di alberi. L'obiettivo è quello di applicare sequenzialmente l'algoritmo di classificazione debole a versioni dei dati ripetutamente modificate in modo da avere una sequenza di classificatori deboli $T^{(b)}$ con $b = 1, \ldots, B$.
Il risultato finale della classificazione si ottiene come :

$$\hat{k}_{boost}(x) = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \sum_{b=1}^{B} \alpha_b \mathbb{1}_{\{T^{(b)}(\boldsymbol{x})=k\}}$$

dove gli α_b sono pesi costruiti ad ogni passo in modo da tener conto delle osservazioni classificate non correttamente.

Misure di importanza delle variabili

Alcune misure di importanza delle variabili sono:

- La *Mean decrease Gini* che si ottiene considerando una misura costruita come ammontare della decrescita dell'indice di Gini che risulta dagli split su una variabile (media sugli alberi).
- La *Mean decrease accuracy* che si basa sull'ammontare della decrescita dell'accuratezza quando una certa variabile è permutata (media sugli alberi).

Model assessment

		previsto	
	Popolazione totale (PT)	positivo	negativo
Effettivo	positivo	Vero positivo (VP)	Falso negativo (FN)
	negativo	Falso positivo (FP)	Vero negativo (VN)

$$Acc = \frac{VP + VN}{PT} \quad spec = \frac{VN}{VN + FP} \quad sens = \frac{VP}{VP + FN}$$

$$prec = \frac{VP}{VP + FP} \quad F1 = \frac{2VP}{2VP + FP + FN}$$

Classificazione dei tumori tramite alberi decisionali

Il dataset è composto da 569 osservazioni di 30 variabili per caratterizzare forma e dimensioni del tumore. 357 osservazioni sono della classe benigni e 212 maligni. Si utilizzano 250 osservazioni come *training set* e le restanti come *test set*.

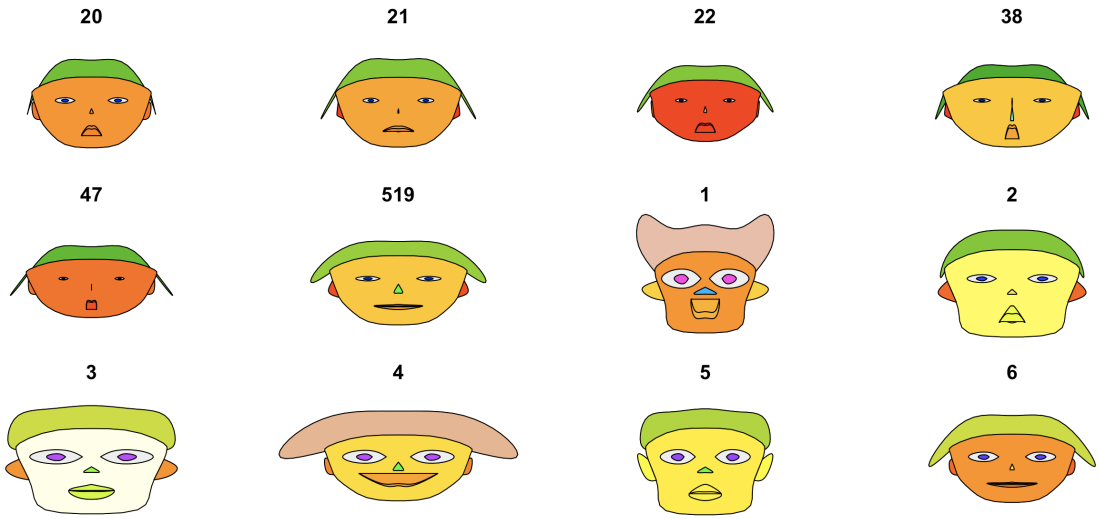


Figura 2: Facce di Chernoff

Classificazione dei tumori tramite alberi decisionali 2

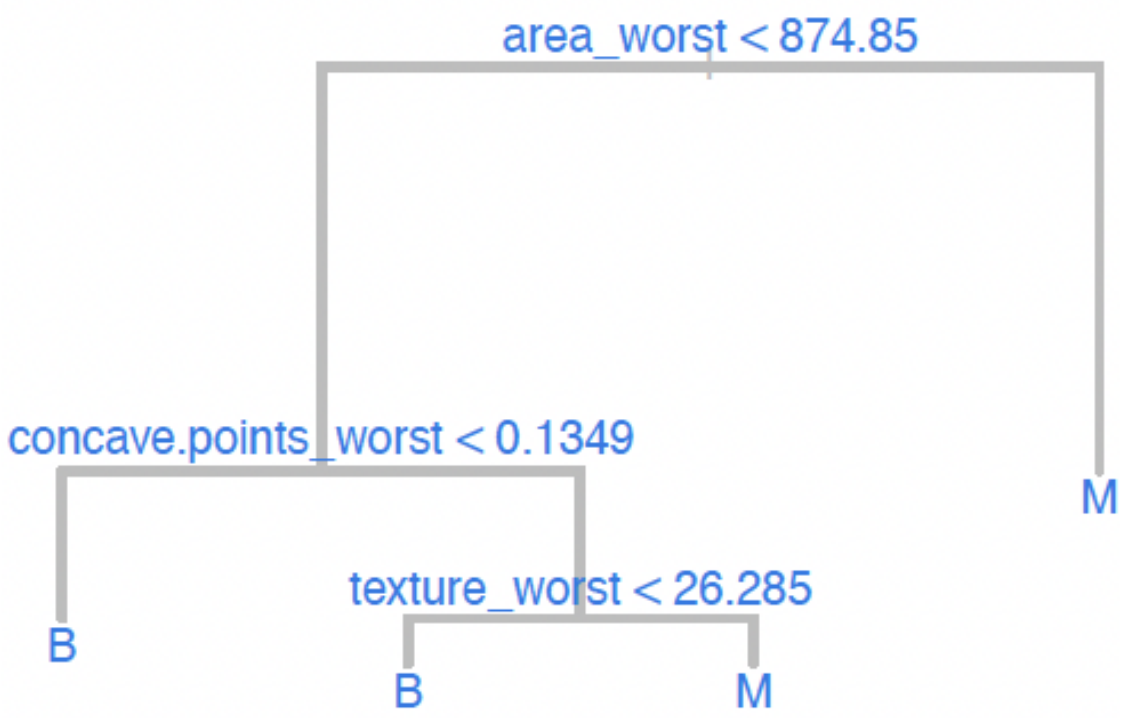


Figura 3: Esempio di singolo albero di classificazione dopo il pruning con accuratezza pari a 95.3%

Classificazione dei tumori tramite alberi decisionali 3

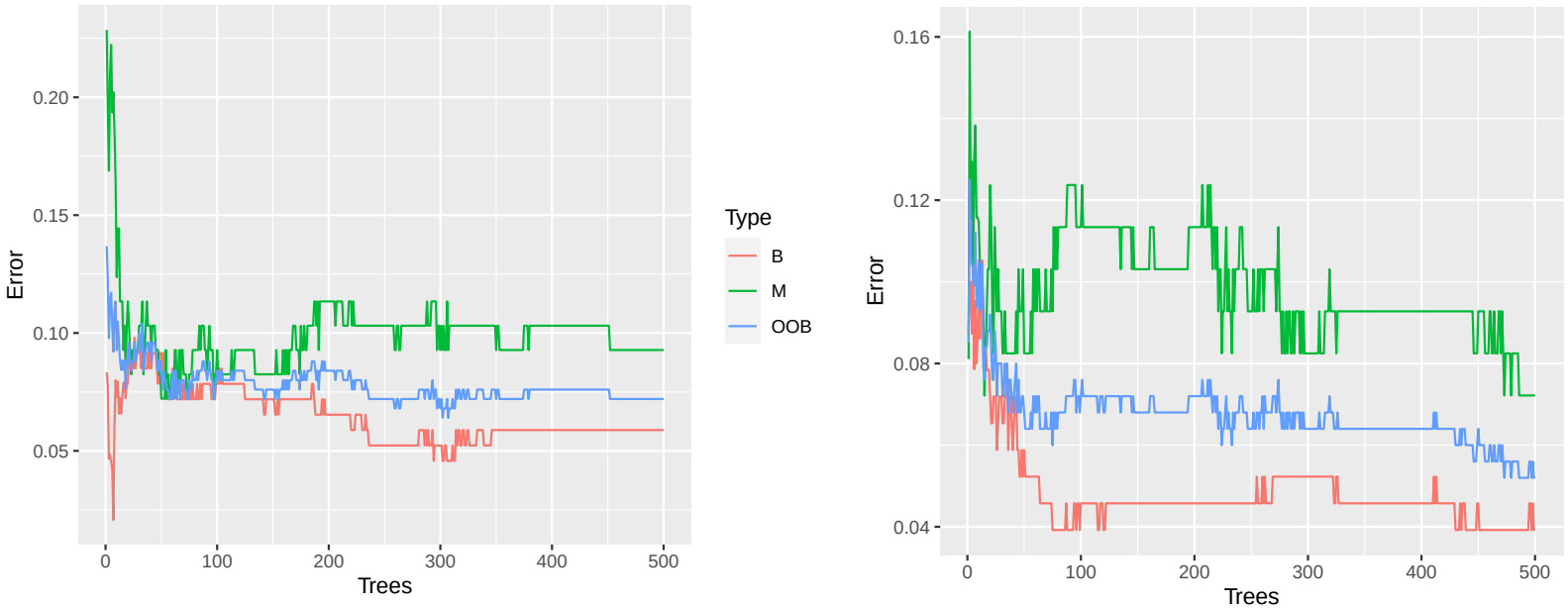


Figura 4: Errori di classificazione per bagging sinistra e random forest destra

Classificazione dei tumori tramite alberi decisionali 4

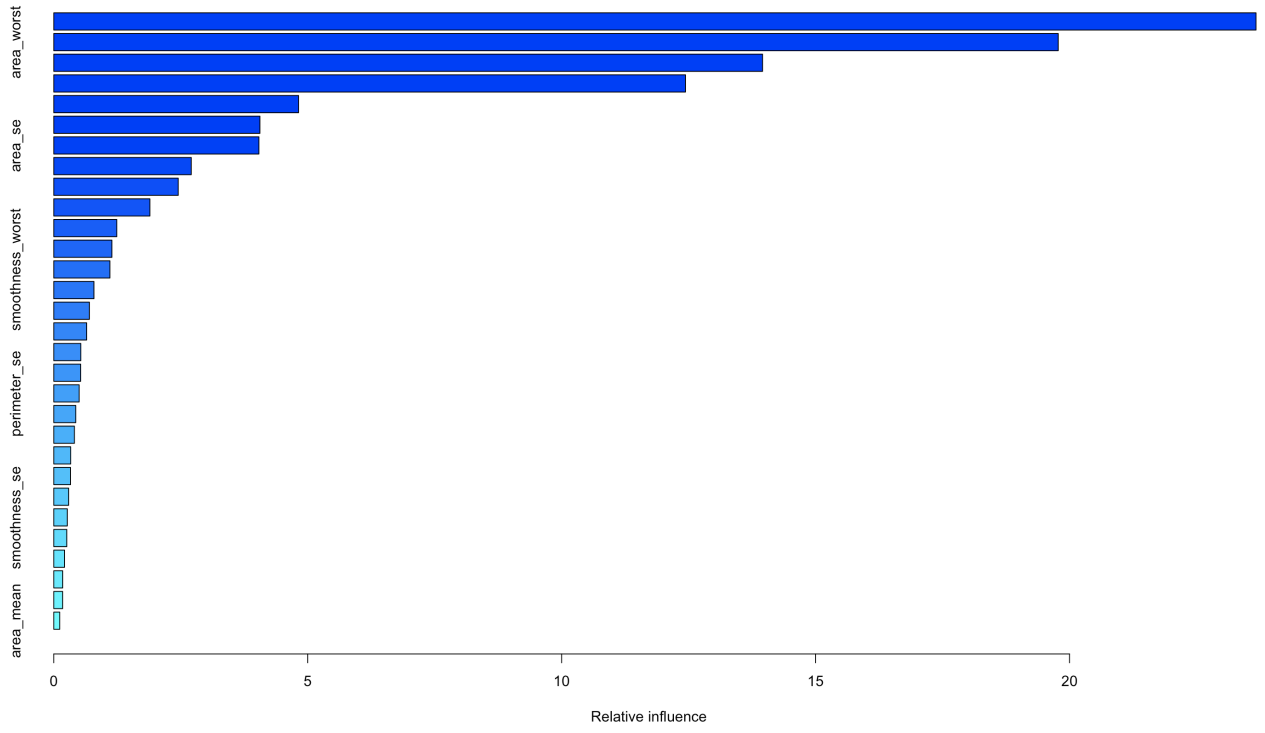


Figura 5: influenza della variabili nel boosting

	Accuratezza	Sensibilità	Specificità	Precisione	F1
DecisionTree	0.9529781	0.9608	0.9391	0.9655	0.9631
Bagging	0.9623824	0.9657	0.9565	0.9752	0.9704
RandomForest	0.9686520	0.9853	0.9391	0.9663	0.9757
Boosting	0.9717868	0.9755	0.9652	0.9803	0.9779

Some references

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction tostatistical learning (Second Edition). New York: Springer.
- Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).
- Biau, G., Scornet, E. A random forest guided tour. TEST 25, 197–227 (2016).
- Introduzione ai Modelli Statistici Giovanni M. Marchetti Dipartimento diStatistica, Informatica, Applicazioni, Firenze 2013, 2019 rev. 1
- Wasserman, L. (2004). All of statistics. Springer-Verlag, New York.
- Friedman, J., Hastie, T., & Tibshirani, R. (2013). The elements of statistical learning. Second edition. Springer, Berlin: Springer series in statistics.