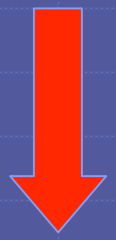
Che cosa è un classificatore

Un classificatore o una regola di classificazione rappresenta un modo sistematico per predire la classe di appartenenza di un dato sperimentale



Un classificatore è una funzione **d** che associa una classe di appartenenza **c**, dello spazio ***C*** di tutte le classi, ad ogni vettore **x** dello spazio ***X*** costituito da tutti i dati sperimentali

$$\mathbf{d}(\mathbf{x}) = \mathbf{c} \qquad \mathbf{x} \in \boldsymbol{X},\ \mathbf{c} \in \boldsymbol{C}$$

Gli scopi di un'analisi di classificazione

Acquisire una maggiore conoscenza delle variabili che governano il fenomeno oggetto di studio:

- Stabilire una gerarchia di fattori sulla base della loro importanza
- Eliminare i fattori non rilevanti ai fini della classificazione

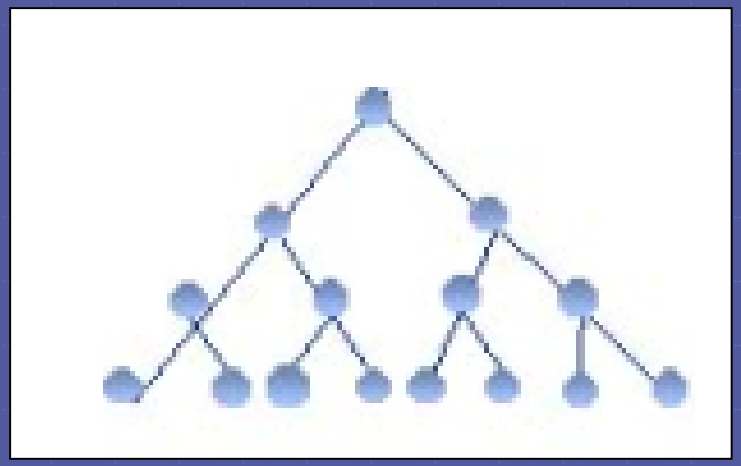
Creare una struttura predittiva in grado di determinare la classe di appartenenza dei nuovi dati sperimentali

Ciascuna unità percorre l'albero dalla *radice* (il nodo più alto dell'albero) verso il basso, procedendo ad una classificazione per ciascun nodo che trova sul suo cammino, sino al raggiungimento di un nodo terminale o *foglia*

Classification and Regression Trees

Metodologia statistica introdotta da Breiman *et al.* nel 1984

- produce classificazioni nella forma di diagrammi ad albero rovesciato



- produce segmentazioni di tipo binario
- produce alberi di classificazione se i dati sono qualitativi, alberi di regressione se i dati sono quantitativi

Problema affrontato

Un collettivo è descritto da variabili rilevate con scala di misura eterogenea

Una di queste variabili ha il ruolo di risposta per l'impatto delle altre (regressori)

Si ipotizzano relazioni non lineari o indirette tra la risposta e regressori

E' necessario prevedere il valore della risposta di una unità non ancora analizzata (Regression)

Ovvero stabilire la classe più plausibile di appartenza della nuova unità (Classification)

Idea guida del Cart

Passo_1: si sceglie un regressore X e si bipartiscono le unità in DUE gruppi.

Se il regressore è su scala metrica oppure è su scala ordinale, la suddivisione avviene per un valore di soglia X_s :

unità con $X<X_s$ e unità con $X>X_s$

Se il regressore è binario o politomo, la suddivisione avviene rispetto ad una delle cagegorie A_s

unità con $X=A_s$ e unità con $X=A_s$

La stessa variabile può essere utilizzata più volte in livelli differenti dell'albero

Idea guida del Cart/2

Passo_2: Ognuno dei due gruppi è a sua volta suddiviso in due sottogruppi con la stessa procedura.

La suddivisione ricorsiva termina quando l'ultima suddivisione ha raggiunto un grado di omogeneità (o purezza) soddisfacente nei due gruppi finali.

Oppure termina quando gli annidamenti sono troppi per essere realmente utili.

Come classifica il CART

- Suddivide ogni nodo dell'albero in altri due nodi dopo aver individuato lo *split* migliore mediante:
  - suddivisioni binarie per ogni variabile per varie soglie
  - un criterio di bontà dello split ad ogni nodo

- decide quando un nodo è da considerarsi terminale (regole di stop)

- assegna una classe di appartenenza
- ad ogni nodo terminale (ad esempio,
- la moda)

$$\pi_i = \frac{\sum_{l=1}^{n_m} y_{lm}}{n_m}$$

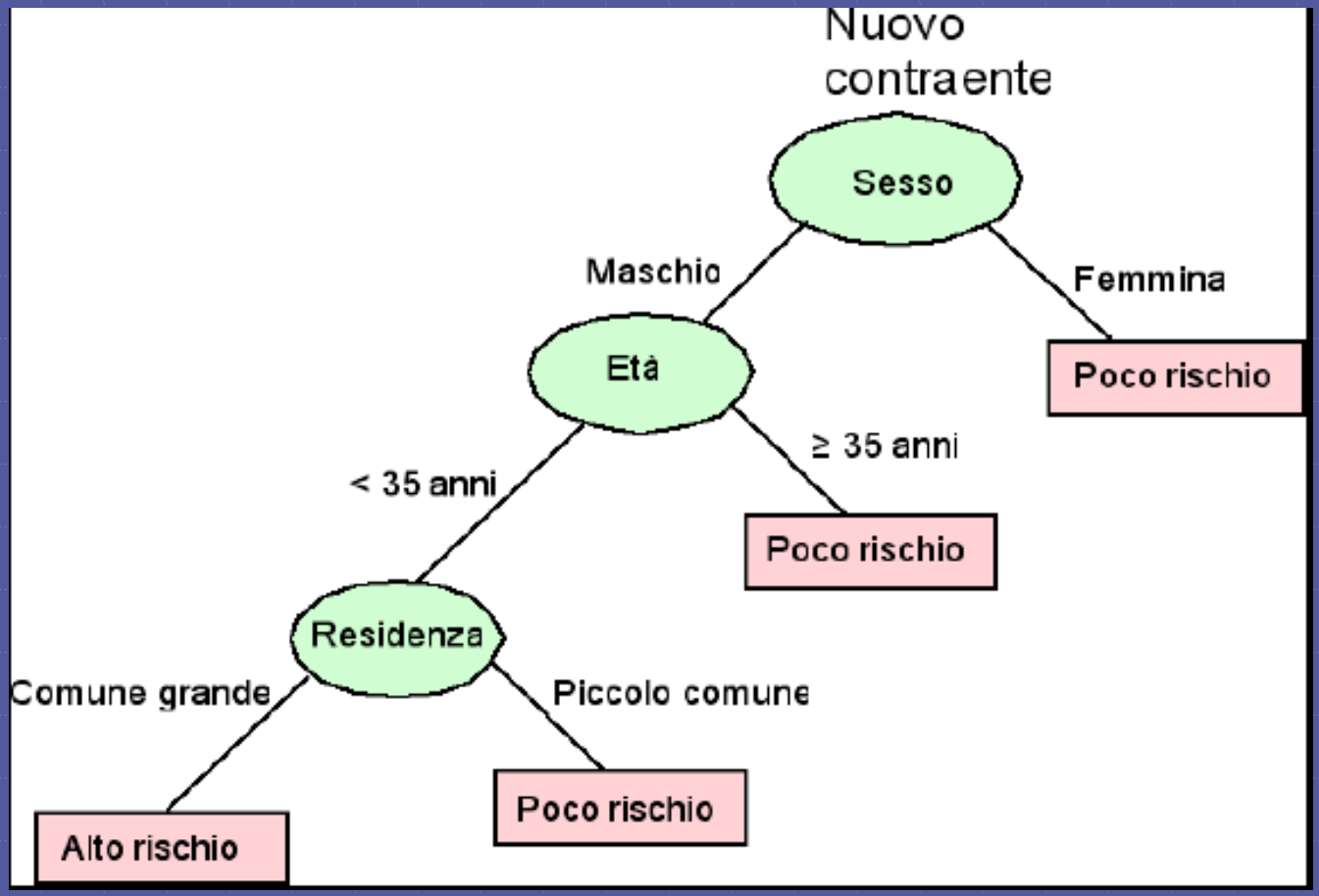
- ovvero assegna un valore della risposta ad
- ogni nodo terminale. Ad esempio la media
- dei valori della risposta nel gruppo terminale.

$$\hat{y}_i = \frac{\sum_{l=1}^{n_m} y_{lm}}{n_m}$$

Esempio

- si realizzano delle osservazioni su alcuni assicurati sui quali si rilevano le variabili di interesse.
- Si analizzano i dati così raccolti per costruire delle procedure stilizzate di individuazione delle polizze (percorsi).
- Al presentarsi del nuovo contraente si acquisiscono i dati solo sulle variabili che sono state effettivamente coinvolte in almeno in un percorso.
- Si segue il percorso e si colloca l'assicurando al livello di rischiosità che compare alla fine del percorso.

Esempio/2

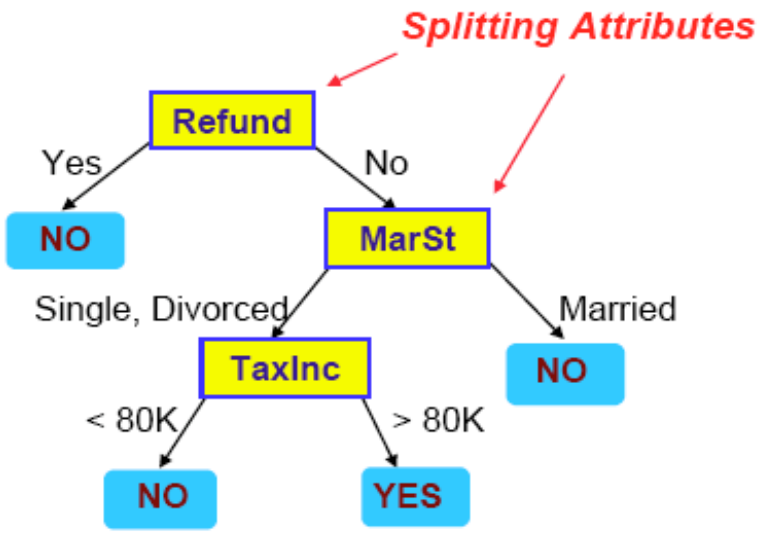


- I nuovi contraenti sono divisi in due distinte categorie di rischio:
- BASSO (donne, maschi di età ≥ 35, maschi di età <35 residenti in piccoli comuni)
- ALTO: (Maschi di età inferiore a 35 anni residenti in comuni grandi).

Esempio

Esempio di albero di classificazione

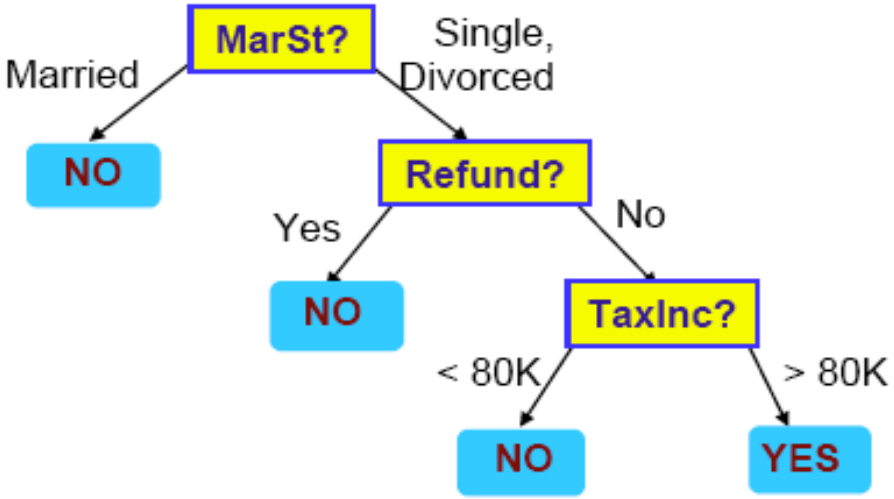
Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Esempio/2

Un altro albero di classificazione

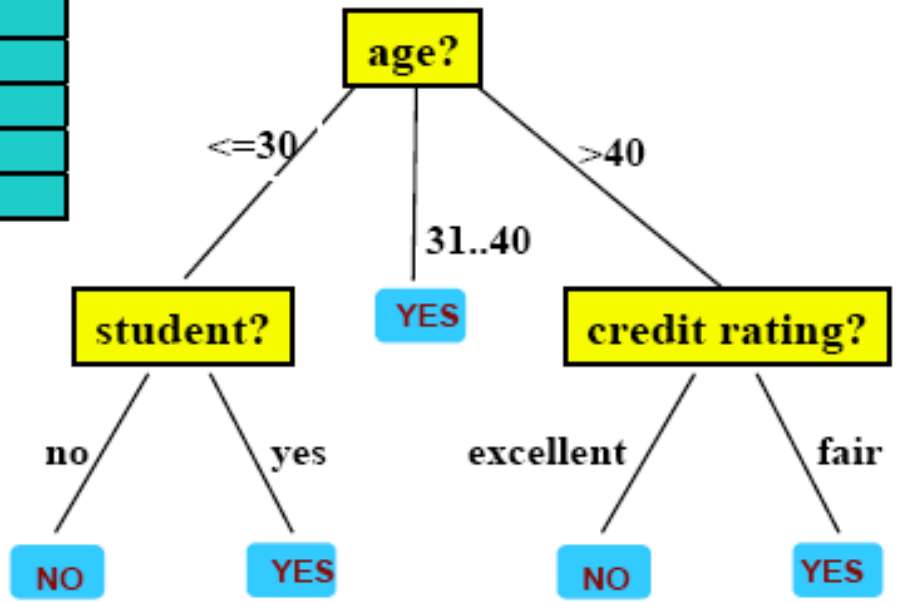
	categorical	categorical	continuous	class
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Si possono derivare più alberi dagli stessi dati!

Altro esempio

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



I vantaggi del CART

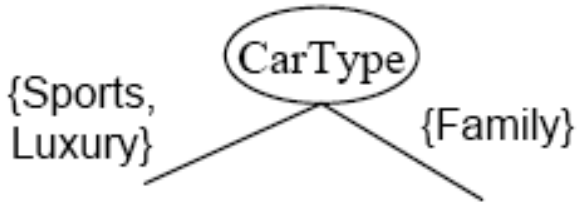
- La classificazione finale ha una forma semplice e di facile interpretazione
- non è necessario fare ipotesi sulla distribuzione statistica dei dati (tecnica non parametrica)
- **Le variabili possono essere sia qualitative che quantitative**
- si possono trattare set di dati molto ampi e con un numero molto alto di variabili
- i risultati sono invarianti per trasformazioni monotone delle variabili (logaritmi, elevamento a potenza positiva...)
- **è possibile processare casi con dati mancanti**

Splitting attributi nominali

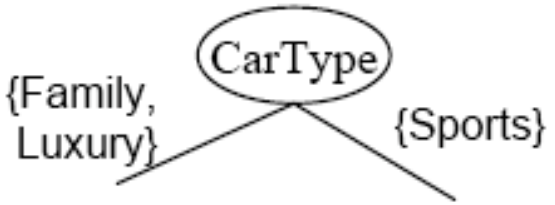
- Ciascuna partizione è caratterizzato da un sottoinsieme di valori.
- **Multi-way split:** Usa tanti ramificazioni dello split quanti sono i valori distinti.



- **Binary split:** Divisi i valori in due sottoinsiemi.
  Bisogna individuare un partizionamento ottimale.

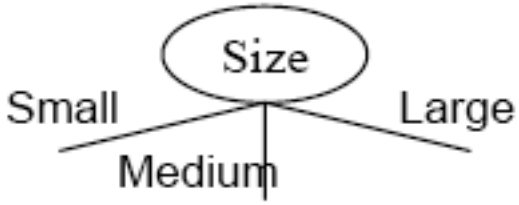


oppure

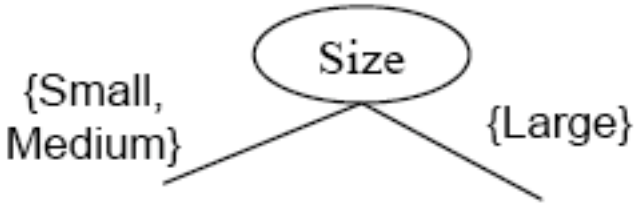


Splitting attributi ordinali

- **Ciascuna partizione è caratterizzato da un sottoinsieme di valori.**
- **Multi-way split: Usa tanti ramificazioni dello split quanti sono i valori distinti.**



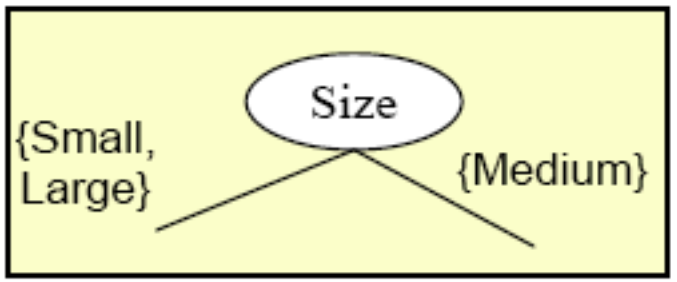
- **Binary split: Divisi i valori in due sottoinsiemi.**
  Bisogna individuare un partizionamento ottimale.



oppure



- **Questo partizionamento potrebbe essere possibile?**



Splitting attributi metrici

- **Differenti metodi**
  - **Binary Decision: ($A < v$) or ($A \geq v$)**
    - considera tutti i possibili split e individua il miglior taglio
    - Può risultare molto costo computazionalmente, anche se esistono dei metodi basati sull'ordinamento
  - **Discretizzazione per formare un attributo categorico (ordinale)**
    - Statico – discretizzato una volta per tutte all'inizio
    - Dinamico – le suddivisioni possono essere trovati tramite *equal interval partitioning*, *equal frequency partitioning*, o *partitioning*.

Omogeneità nei gruppi

Il meccanismo di bipartizione è governato dalla qualità dei due gruppi ottenuti.

Ogni volta si deve decidere se la suddivisione ha raggiunto un grado significativamente maggiore di omogeneità nei nuovi gruppi ottenuti.

Esistono molti modi di definire delle *misure di impurità* e quindi dei criteri da ottimizzare come guida per la segmentazione. In genere si sceglie la suddivisione che rende massimo il valore del criterio

Per una risposta con k categorie sia $p(j|t)$ la frequenza relativa della j-esima categoria nel t-esimo nodo per $j=1,\ldots,k$:

Indice di Gini

- In corrispondenza di un certo nodo *t* esiste un certo dataset
- Il Gini Index in corrispondenza del nodo *t* :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTA: $p(j \mid t)$ è la frequenza relativa della classe *j* al nodo *t*).

– Misura l'impurità/disordine del dataset corrispondente a *t*.
  - Massimo valore ($1 - 1/n_c$) quando i record sono equamente distribuiti tra tutte le classi ⇒ informazione meno interessante
  - Minimo valore (0.0) quando tutti i record appartengono ad una sola classe ⇒ informazione più interessante

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Applicazione del criterio

- Criterio di Splitting: Minimizza il Gini Index della suddivisione.
- Quando un nodo *t* è suddiviso in *k* partizioni (figli), la qualità della suddivisione è calcolata come:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

dove, n_i = numero di record della partizione (figlio) *i*,
n = numero di record del dataset al nodo *t*.

n_i / n costituisce il peso dei vari $GINI(i)$

- Dato il dataset in corrispondenza del nodo *t*, si sceglie l'attributo che fornisce il più piccolo $GINI_{split}(t)$ per partizionare il dataset
  - *È necessario enumerare tutti i possibili punti di splitting per ciascun attributo*

Entropia della suddivisione

- In corrispondenza di un certo nodo *t* esiste un certo dataset
- Information Gain in corrispondenza del nodo *t* :

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

(NOTA: $p(j \mid t)$ è la frequenza relativa della classe *j* al nodo *t*).

- Misura l'omogeneità/ordine di un nodo.
  - Massimo (log n_c) quando i record sono equamente distribuiti tra tutte le classi ⇒ implica meno informazione
  - Minimo valore (0.0) quando tutti i record appartengono ad una sola classe ⇒ implica più informazione
- Computazioni basate sull'Entropia sono simili a quelle basate sul GINI index

Uso del criterio di entropia

- Quando un nodo *t* è suddiviso in *k* partizioni (figli), la qualità della suddivisione è calcolata come:

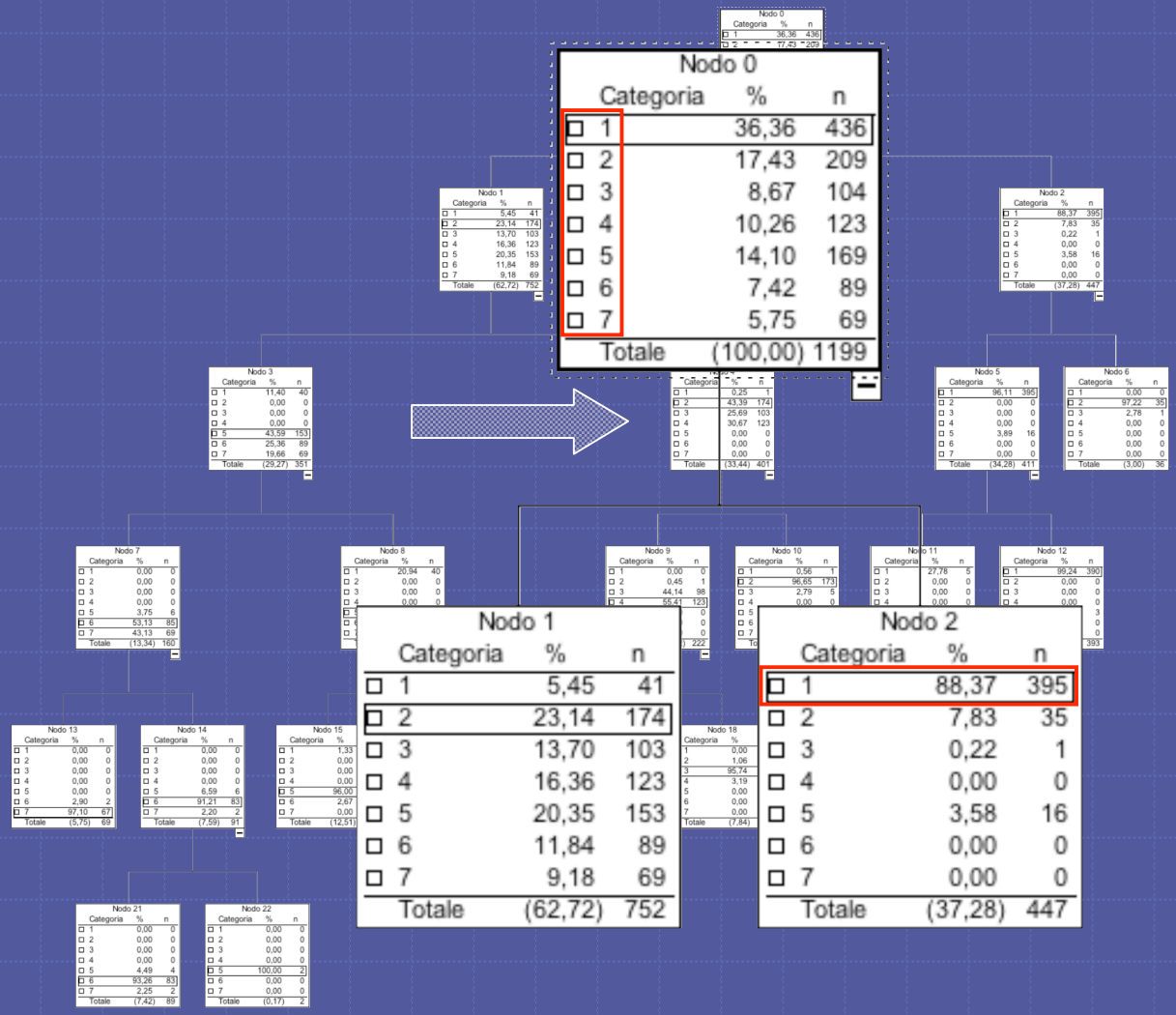
$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

dove, n_i = numero di record della partizione (figlio) *i*,
n = numero di record del dataset al nodo *t*.

n_i / n costituisce il peso dei vari *Entropy(i)*

- Misura la riduzione nell'Entropia in conseguenza dello split. Scegli lo split che raggiunge la riduzione maggiore (massimizza il GAIN)

Esempio applicativo



La potatura (pruning)

Interrompere i frazionamenti successivi non è semplice:

Alberi con molte foglie rischiano l'overfitting cioè adattano un modello troppo complesso per essere interpretabile.

Alberi con poche foglie possono mancare aspetti importanti delle unità da classificare.

Questi problemi si possono affronate con la potatura o pruning: si costruisce un albero molto frondoso per poi potare, con regole opportune, le ramificazioni meno interessanti.

La potatura /2

Si minimizza una misura dell'albero che tiene conto dell'errore di classificazione e della complessità dello stesso:

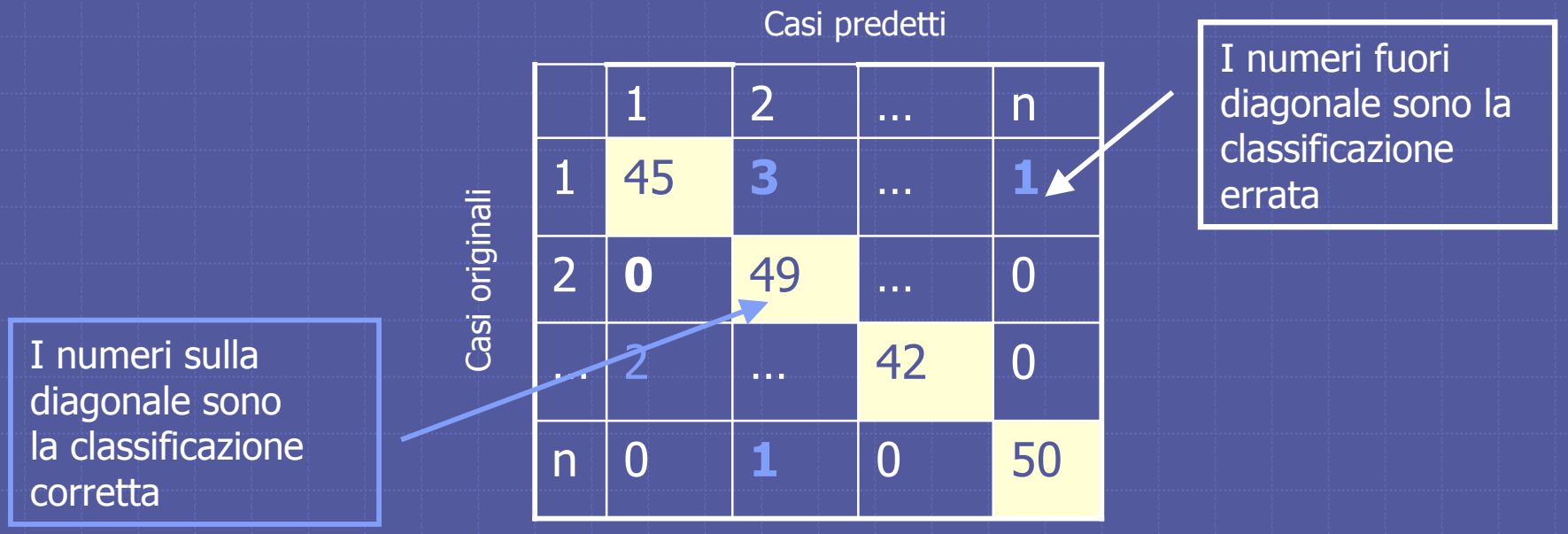
$$R_{\alpha}(T) = R(T) + \alpha N(T)$$

L'errore di classificazione $R(T)$ per un certo albero T corrisponde alla percentuale di osservazioni classificate erroneamente (che è zero per l'albero di dimensione massima che classifica correttamente tutte le osservazioni)

$N(T)$ indica il numero di foglie o nodi terminali, mentre a è un coefficiente che penalizza la complessità dell'albero

Un punto di contatto…

- Matrici di confusione: entrambe le metodologie statistiche creano come output delle matrici, nelle quali si hanno evidenziati i casi classificati correttamente e quelli mal classificati



	1	2	…	n
1	45	**3**	…	**1**
2	**0**	49	…	0
…	2	…	42	0
n	0	**1**	0	50

Criticità…

- Overfitting e selection bias per gli indicatori con molte possibili suddivisioni.
- Non è in grado di distinguere tra miglioramenti significativi e banali
- Occorre coinvolgere le distribuzioni statistiche nelle suddivisioni