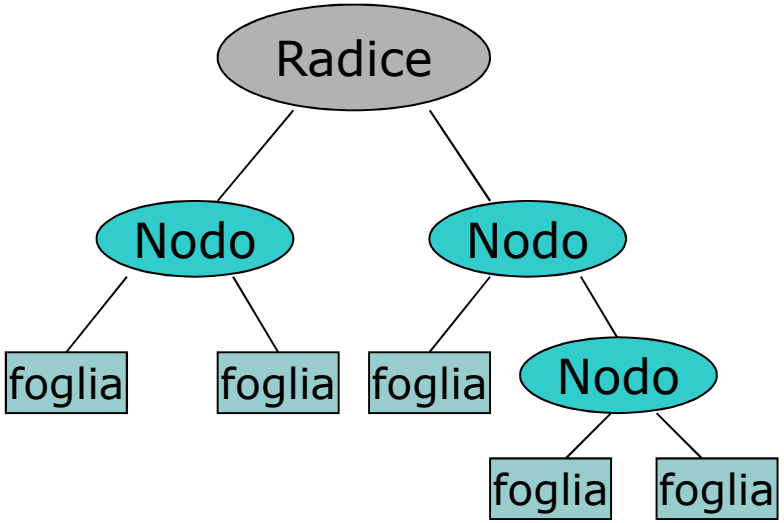
Gli alberi di decisione

Simona Balbi

Cosa sono gli alberi di decisione

- Gli alberi decisionali sono prodotti da procedure di segmentazione
- Quello in figura è prodotto da una segmentazione binaria
- Sono tipicamente utilizzati per problemi di classificazione e previsione



Un esempio classico: giochiamo a tennis?

I dati:

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

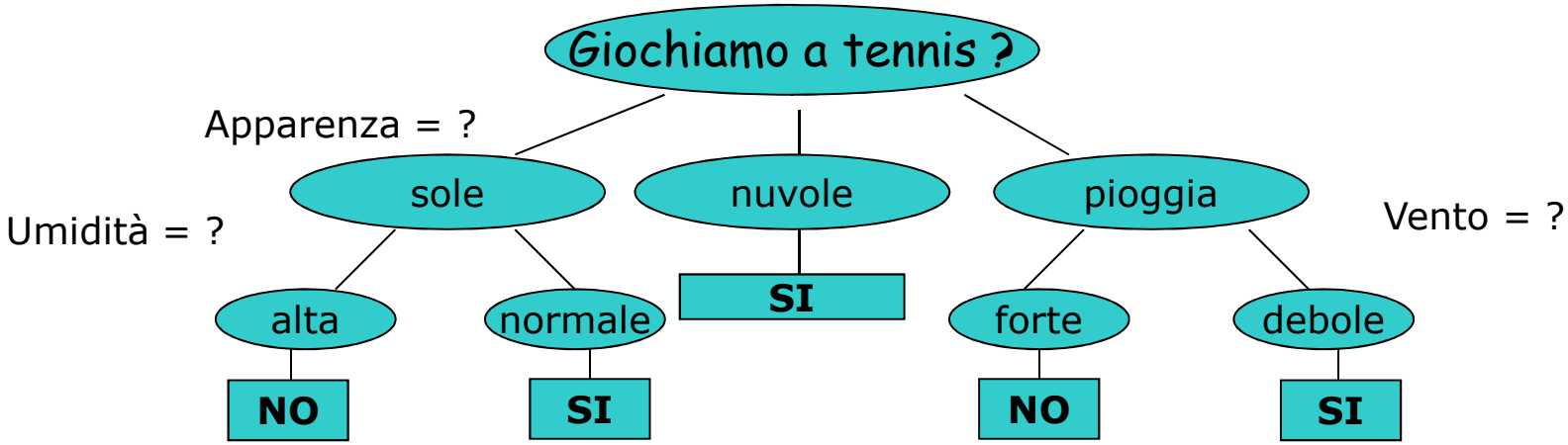
Un esempio classico: giochiamo a tennis?

I **nodi** sono test per i valori di attributi:

uguaglianza (Esempio: "Vento = ?"), diseguaglianza, ogni altro test possibile (>, <, …)

I **rami** (cammini) rappresentano valori di attributi, come corrispondenza uno-ad-uno ("Vento = forte", "Vento = Leggero")

Le **foglie** rappresentano la classificazione assegnata



La segmentazione binaria (sb)

Compito:

Classificare un collettivo di n oggetti in classi omogenee al loro interno e differenziate fra loro, mediante una successione di partizioni dicotomiche (partizione *recursiva*)

La struttura dei dati:

Una matrice (n, p+1):
individui;
p variabili esplicative X_j, $j = 1, \ldots, p$ categoriche o categorizzate;
1 variabile dipendente Y (o criterio, o risposta), quantitativa, ordinale, nominale

Obiettivo esplicativo:

spiegare la variabile di risposta sulla base delle variabili esplicative

Obiettivo decisionale:

sfruttare la regola di classificazione per classificare nuovi casi

Un esempio

Compito:

Classificare di 323 clienti di un istituto di credito secondo la regolarità nella restituzione del prestito, sulla base delle variabili esplicative:
periodicità della retribuzione (settimanale/mensile) e dell'età

La struttura dei dati:

				CREDIT RANKING		Totale
				Affidabile	Non affidabile	
PAGA	sett	ETA'	<=35	15	143	158
			>35	/	0	/
	mens	ETA'	<=25	25	24	49
			>25	108	1	109
Totale				155	168	323

(Zani, 2000 - I dati sono tratti dal file credit.sav collocato nella directory di AnswerTree del package SPSS)

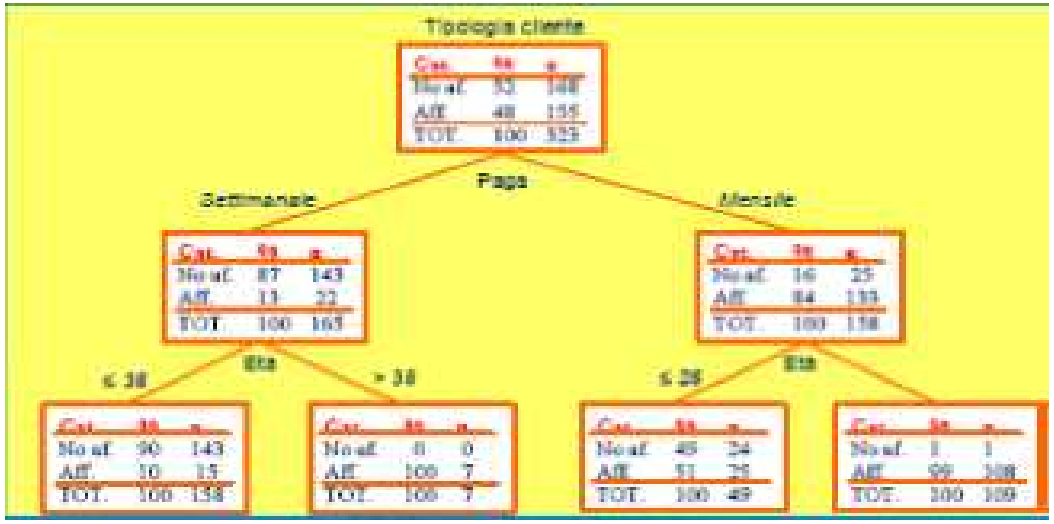
Obiettivo esplicativo:

spiegare la regolarità nei pagamenti sulla base delle caratteristiche note dei clienti

Obiettivo decisionale:

sfruttare la regola di classificazione per decidere di nuovi finanziamenti

La rappresentazione ad albero



L'esempio si riferisce ad una esperienza statunitense. La paga settimanale è di regola indice di una situazione economica precaria. Evidentemente anche l'età gioca il suo ruolo. L'albero di segmentazione consente di affermare che:

✓ i giovani (fino ai 36 anni) con paga settimanale sono decisamente inaffidabili

✓ anche i giovani con paga mensile sono poco affidabili, ma lo sono un po' di più e soprattutto cambia la definizione di "giovane" (fino a 26 anni)

✓ i più anziani sono più affidabili, ricordando però le diverse classi di età

Nota – Le implicazioni decisionali sono EVIDENTI

Un po' di vocabolario

Si dice, infatti, che si tratta di un albero ROVESCIATO

Il rettangolo superiore è la **radice (R)**

L'albero è costituito da un insieme finito di elementi, **i nodi**

Ogni **nodo** è un gruppo di unità a diversi stadi del processo di classificazione

Il **nodo radice** è un nodo disomogeneo al suo interno rispetto alla variabile obiettivo perché racchiude tutti gli individui considerati

L'insieme dei nodi (ad eccezione della radice) può essere suddiviso

in insiemi distinti: **i sottoalberi** del nodo R

Un nodo viene chiamato

- *padre* rispetto ai nodi che esso genera
- *figlio* rispetto al nodo da cui discende

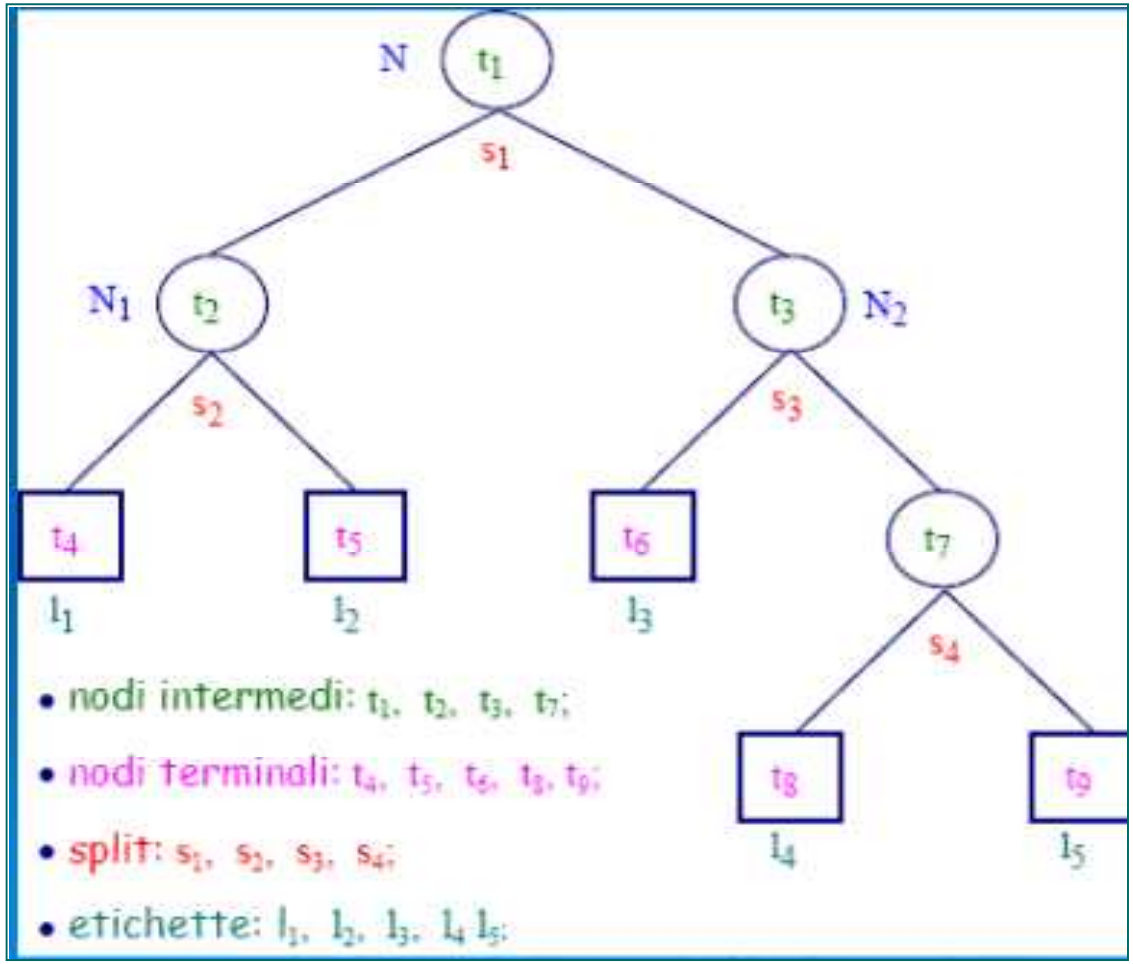
I **valori di soglia** di una variabile che dividono le unità di un determinato nodo sono chiamati **split**

I rami sono le condizioni che hanno determinato la suddivisione

L'insieme di tutti i nodi terminali di un albero viene indicato con il simbolo $\tilde{T}$

Le **foglie** sono i nodi terminali per i quali non si ritieneutile una ulteriore suddivisione

L'albero di decisione binaria



Le fasi di una procedura di s.b.

1. **Un insieme di domande binarie:** stabilire, per ciascun nodo, l'insieme delle divisioni ammissibili
2. **Un criterio di split:** definire un criterio per selezionare la migliore divisione di un nodo
3. **Una regola di arresto:** definire una regola per dichiarare un nodo come terminale o intermedio
4. **Una regola di assegnazione:** ad ogni nodo terminale viene assegnata una delle J classi della variabile risposta nominale o un valore della variabile di risposta continua
5. **La valutazione della qualità della regola di decisione:** stimare il rischio di errore di classificazione o di previsione associato

Fase 1: un insieme di domande binarie a)

stabilire, per ciascun nodo, l'insieme delle divisioni ammissibili

Si tratta per ciascun nodo di stabilire l'insieme delle divisioni ammissibili

Natura del predittore	Numero di modalità	Numero di split
variabile continua	N	N-1
variabile binaria	2	1
variabile ordinale	m	m-1
variabile nominale	m	2^{m-1}-1

Es. di variabile nominale, (m modalità; 2^{m-1} - 1 split)

"colori della confezione":

rosso, ***blu***, ***verde***

3 modalità , 3 split 3

rosso ***vs*** ***blu***, ***verde***

rosso, ***blu*** ***vs*** ***verde***

rosso, ***verde*** ***vs*** ***blu***

Fase 1: un insieme di domande binarie b)

stabilire, per ciascun nodo, l'insieme delle divisioni ammissibili

Es. di variabile ordinale (m modalità; (m-1) split)

"titolo di studio" :

laurea, diploma, licenza media inferiore, licenza elementare, senza titolo

m = 5 *split = 4*

laurea *vs.* diploma, licenza media inferiore, licenza elementare, senza titolo

laurea, diploma *vs.* licenza media inferiore, licenza elementare, senza titolo

laurea, diploma, licenza media inferiore *vs.* licenza elementare, senza titolo

laurea, diploma, licenza media inferiore, licenza elementare *vs.* senza titolo

Fase 2: un criterio di split :

definire un criterio per selezionare la migliore divisione di un nodo

Le tecniche di segmentazione si differenziano per il **criterio di split** adottato

Un **criterio di split** è un indice statistico che consente di selezionare la partizione migliore fra tutte le possibili di ogni variabile esplicativa

Fra tutte le variabili esplicative viene selezionata la migliore in relazione al **criterio di eterogeneità prescelto**

L'insieme iniziale deve essere suddiviso in gruppi il più possibile omogenei al loro interno ed il più possibile eterogenei tra loro

Fase 3: una regola di arresto

definire una regola per dichiarare un nodo come terminale o intermedio

La ripartizione ricorsiva di un insieme di unità statistiche si arresta quando i nodi terminali contengono solo individui appartenenti alla stessa classe della variabile dipendente, o una % predefinita

Scelta della regola:

Fra due regole di arresto si sceglie quella che fornisce l'albero di taglia minore (*proprietà della semplicità – Rasoio di Occam**)

Fra due regole di arresto si sceglie quella che consente di distinguere nel modo più efficace possibile unità statistiche appartenenti a classi differenti (*potere discriminatorio*)

"A parità di fattori, la spiegazione più semplice tende ad essere quella esatta" :

- *« Entia non sunt multiplicanda praeter necessitatem»*
- *« Pluralitas non est ponenda sine necessitate»*
- *« Frustra fit per plura quod fieri potest per pauciora. » William of Ockham*

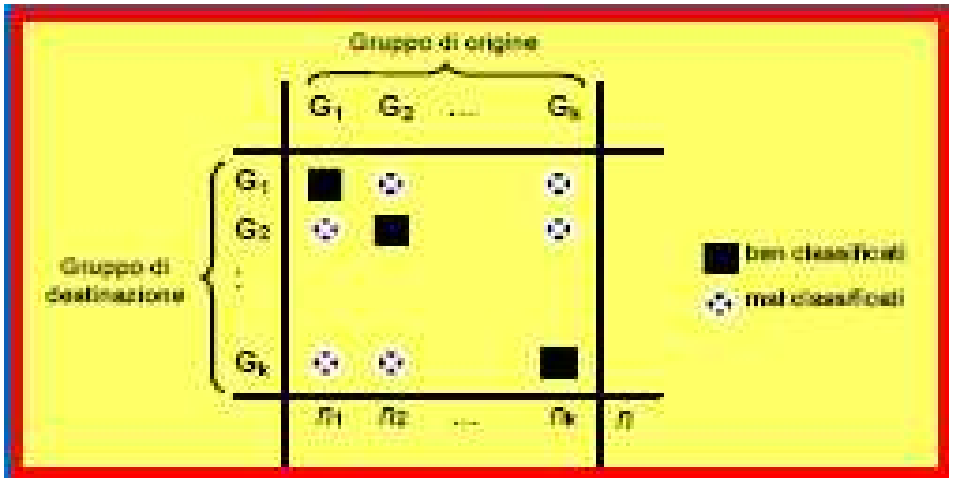
Fase 4: una regola di assegnazione

ad ogni nodo terminale viene assegnata una delle J classi della variabile di risposta nominale o un valore della variabile di risposta continua

- Se la foglia comprende casi appartenenti ad una sola classe, la classe assegnata al nodo è quella corrispondente alle unità che ne fanno parte (**regola dell'unanimità**)
- Se la foglia comprende unità di classi diverse ed una delle classi ha frequenza più alta, la classe assegnata al nodo è quella corrispondente alla frequenza più alta (**regola della maggioranza**)
- Se la foglia comprende unità di classi diverse con la stessa frequenza, si ha una situazione di indecisione che viene risolta, in genere, **assegnando casualmente** la classe al nodo

Fase 5: La valutazione della qualità

Stimare il rischio di errore di classificazione o di previsione associato



A parità di semplicità della rappresentazione ad albero viene selezionata la regola che consente di collocare correttamente la percentuale più elevata di unità statistiche

Ambiti di applicazione
(da L.Fabbris, 1997)

- **Ricerca di interazioni fra variabili predittive**

l'interazione è l'effetto che una combinazione di modalità delle variabili predittive ha sulla variabile dipendente

l'interazione può essere di sinergia o di antagonismo

e può essere scoperta soltanto analizzando le distribuzioni condizionate

ATTENZIONE: quello che interessa non sono le correlazioni fra le variabili predittive, ma come le modalità identificative di subcampioni incidano sul comportamento della variabile dipendente

Identificazione di gruppi devianti

Si tratta di una particolare forma di ricerca di interazioni

***deviante** è un gruppo di unità che possiede valori estremi della variabile dipendente e che si cerca di caratterizzare con combinazioni delle modalità delle variabili predittive*

La segmentazione è stata positivamente utilizzata per identificare: guidatori a rischio, allievi con problemi di lettura; categorie a rischio di comportamenti devianti o di malattie rare

Identificazione di dati anomali

Anche in questo caso si tratta di una ricerca di interazioni:

Alcune foglie di un albero (di solito di dimensioni estremamente piccole) possono contenere valori che si collocano sulle code delle distribuzioni congiunte

- **Generazione di ipotesi di ricerca e di modelli interpretativi (in senso causale) del comportamento della variabile dipendente**

in questo senso la segmentazione è uno strumento tipico di strategie di DM: è uno strumento per la ricerca di pattern e di modelli interpretativi

Naturalmente lo si può concepire come una fase, seguita da una analisi successiva di carattere confermativo

- **Produzione di regole di previsione o di classificazione**

si tratta di una regola di predizione della variabile di risposta, creata sulla base delle osservazioni, al fine di inserire in un gruppo un nuovo soggetto per cui siano note le variabile eaplicative

*ATTENZIONE: il rischio dell'**overfitting** può rendere poco generalizzabile i risultati*

- **Sintesi delle informazioni contenute nei dati**

un ulteriore e positivo risultato fornito dalla s.b. è la sua capacità di ***identificare*** *delle situazioni di ridondanza di informazioni, relative alla variabile dipendente e contenuta in più variabili esplicative, ed è in grado di* ***eliminarla*** *nella costruzione dell'albero*

- **Ricerca di relazioni non lineari tra variabili quantitative e non monotone tra variabili ordinali**

si noti che se un insieme di unità statistiche viene suddiviso in almeno 3 sottoinsiemi (quindi non segmenazione binaria) sulla base di un predittore su scala quatitativa, si può intuire la non linearità della relazione che le lega. Analogamete, per variabili predittive ordinali, se l'andamento è prima crescente e poi decrescente (o viceversa), si ha indicazioni di una rlazione non monotone

Imputazione di dati mancanti

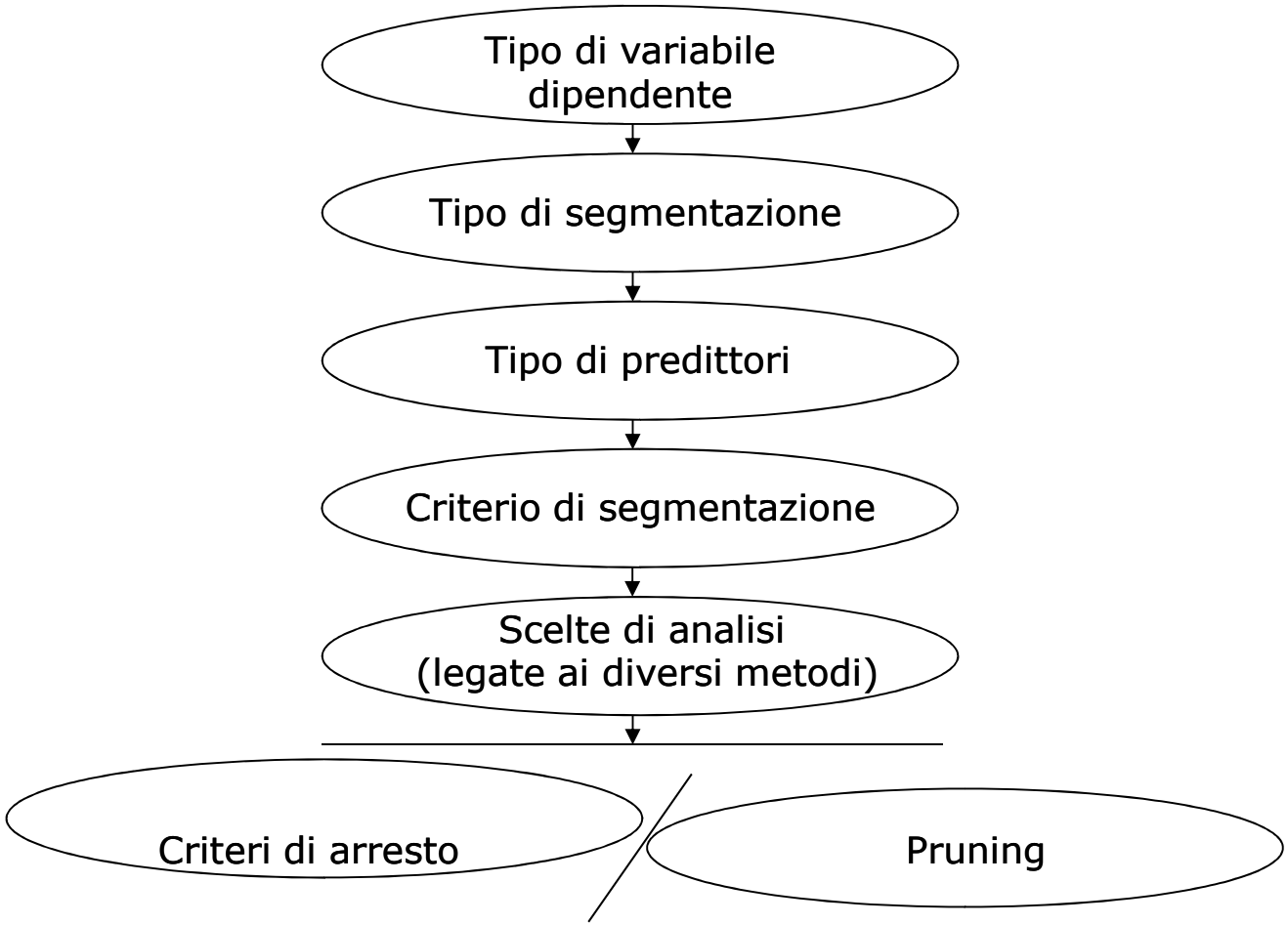
La segmentazione consente di identificare gruppi di unità omogenee rispetto alla variabile di risposta e di sostituire così il dato mancante con uno appartenente alla stessa classe. Si tratta quindi di una variante combinata del metodo della regressione e del metodo dell'analisi dei gruppi per forzare un dato valido al posto di un dato mancante (o incoerente)

Scelte (da L.Fabbris, 1997)

- **Prerequisiti**
  - ***Campione di grandi dimensioni***, *sia nei confronti dell'errore campionario, ma anche extra-campionario*
  - *Attenzione, nel caso di variabili auantitative o dicotomiche, a variabili dipendenti fortemente asimmetriche, poiché si identificano rapidamente i gruppi di coda*
  - *Predittori con un numero di modalità contenuto e con distribuzioni quasi uniformi sono da preferire per vari motivi:*
    - *predittori con distribuzione fortemente asimmetriche finiscono con l'essere esclusi nella costruzione dell'albero, almeno per le modalità rare*
    - *predittori con molte modalità sono in genere preferiti dai criteri di split, al di là del loro potere esplicativo*

Sono suggerite 3 o 4 classi, come struttura ottimale dei predittori che si può ottenere con opportune categorizzazioni, o aggregazioni di modalità

Scelte


Tipo di variabile dipendente
Tipo di segmentazione
Tipo di predittori
Criterio di segmentazione
Scelte di analisi (legate ai diversi metodi)
Criteri di arresto
Pruning

La ricerca del miglior SPLIT

Obiettivo

Costruzione di un albero che sia il più compatto possibile

Ostacoli

Trovare l'albero minimale è computazionalmente oneroso

Strumenti

Algoritmo ricorsivo, con strategia divide-et-impera, si tratta di una strategia euristica greedy e non garantisce, quindi, l'ottimalità: può convergere ad un minimo locale

Decisione principale: l'attributo da scegliere

Desiderata: attributi che splittano gli esempi in insiemi che sono relativamente "puri" e che devono portare più rapidamente possibile ad un nodo foglia

Richiami

- **algoritmo ricorsivo**: risulta particolarmente utile per eseguire dei compiti ripetitivi su un insieme di variabili. La sua esecuzione su un insieme di dati comporta la semplificazione o suddivisione dell'insieme di dati e la sua applicazione agli insiemi di dati semplificati. L'algoritmo termina al verificarsi di una condizione particolare
- **strategia *divide-et-impera***:
  - *divide* trasformare un problema di dimensione **n** in b problemi indipendenti di dimensione n/b
  - *impera* risolvere i b problemi elementari
  - *Combinare* i b risultati
- **euristica *greedy:*** un algoritmo di tipo *greedy* è un algoritmo che cerca di ottenere una soluzione ottima da un punto di vista globale attraverso la scelta della soluzione più golosa ad ogni passo locale. Questa tecnica, quando è applicabile, non sempre arriva ad una soluzione ottima, ma trova soluzioni ottimali per determinati problemi in tempi contenuti

I principali algoritmi di segmentazione

Metodi/ Algoritmo	SEGMENTAZIONE	VARIABILE DIPENDENTE	PREDITTORI
AID	Binaria	Quantitativa	Qualitativi
CHAID	Multipla	Qualitativa	Qualitativi
CART	Binaria	Qualitativa e quantitativa	Qualitativi e Quantitativi
C4.5	Binaria	Qualitativa e quantitativa	Qualitativi e Quantitativi

Algoritmo AID *(Morgan&Sonquist - 1963 - Automatic Interaction Detector)*

- Variabile dipendente QUANTITATIVA
- Variabili esplicative QUALITATIVE

SEGMENTAZIONE BINARIA

Criterio di suddivisione:

scomposizione della devianza della variabile dipendente nella quota entro e tra i gruppi che si vengono a determinare nei diversi livelli dell'albero, quindi sul coefficiente di correlazione di Pearson

Ad ogni passo della procedura di segmentazione, viene selezionata

i) una var. X e

ii) una certa partizione dello spazio su cui X è definita, che massimizza la DEV(Y) tra i gruppi (formati sulla base della X stessa, *r* è il numero di gruppi

$$\text{MAX Dev}_{\text{tra}}(Y) = \left[\sum_r (\bar{y}_r - \bar{y})^2 n_r \right]$$

AID: *Una singola variabile dipendente quantitativa*

$$\eta^2 = \frac{BSS}{TSS} = \frac{N_1(\bar{Y}_1 - \bar{Y})^2 + N_2(\bar{Y}_2 - \bar{Y})^2}{\sum_{i=1}^{2}\sum_{j=1}^{i}(Y_{ij} - \bar{Y})^2}$$

dove $\bar{Y}$ è il valore medio della variabile dipendente in un nodo padre e $\bar{Y}_i$ sono le medie dei sottogruppi formati dalla suddivisione del nodo padre e N_i le rispettive numerosità

Massimizzare significa massimizzare il potere esplicativo dello split

AID: *Più variabili dipendenti quantitative*

Una generalizzazione al caso di più variabili dipendenti è stata proposta, utilizzando come criterio di split la naturale generalizzazione di , basato sulla scomposizione della traccia della matrice di varianza e covarianza:

tr(**T**) = tr(B) + tr(**W**)

$$M^2 = 1 - \left(tr(\mathbf{W}) / tr(\mathbf{T})\right) = tr(\mathbf{B}) / tr(\mathbf{T})$$

dove **T** è la matrice di varianza e covarianza delle variabili dipendenti nel nodo padre, B quella dei 2 gruppi formati dallo split e W quella interna ai gruppi

M^2 può essere interpretata come una misura globale del potere esplicativo dello split

AID: *Regole di arresto*

Le *stopping rules* degli algoritmi AID e varianti sono in genere legate a due elementi:

1. Una soglia legata al potere esplicativo dello *split* : un sottogruppo deve contenere al meno una proporzione minima della varianza totale, per essere candidato interessante per un successivo *split*
2. Il numero finale dei sottogruppi: l'algoritmo si ferma quando l'albero ha raggiunto un numero pre-fissato di foglie. La conseguenza può essere che alcuni rami possono ottenere una segmentazione più fine di altri

CHAID *(Kass -1980, Chi-square Automatic Interaction Detection)*

- Variabile dipendente QUALITATIVA
- Variabili esplicative QUALITATIVE

SEGMENTAZIONE MULTIPLA

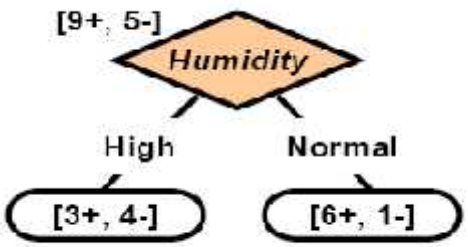
Criterio di suddivisione dei nodi è basato su un **test** per la verifica dell'**ipotesi di indipendenza** statistica tra la variabile dipendente e la variabile esplicativa

Regole di stop

definendo la dimensione massima dell'albero, il numero massimo di livelli, oppure il numero minimo di elementi in un nodo

CHAID – Idee di base

1. **Per ogni attributo di ogni predittore e per ogni combinazione di attributi, si costruisce la tabella di contingenza corrispondente, rispetto alle modalità della variabile dipendente**
2. **Si calcola il — e il corrispondente p-value**
3. **Si seleziona l'attributo X_j che ha il più piccolo p-value (p*min*) e lo si confronta con un valore di arresto —**
   - Se p*min* < —, allora si utilizza X_j come attributo di split
   - Se p*min* > , allora si è individuata una foglia



Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

Humidity/PlayTennis	Yes	No	Totale
High	3	4	7
Normal	6	1	7
Totale	9	5	14

χ^2=2.5, d.f.=1 $p_{Humidity}$ = 0.0588

Wind/PlayTennis	Yes	No	Totale
Light	6	2	8
Strong	3	3	6
Totale	9	5	14

χ^2=0.9, d.f.=1 p_{Wind} = 0.2590

Outlook/PlayTennis	Yes	No	Totale
Sunny	2	3	5
Rain	4	0	4
Overcast	3	2	5
Totale	9	5	14

χ^2=3.5467, d.f.=2 $p_{Outlook}$ = 0.0849

Temperature/PlayTennis	Yes	No	Totale
Hot	2	2	4
Mild	4	2	6
Cool	3	1	4
Totale	9	5	14

χ^2=0.5704, d.f.=2 $p_{Outlook}$ = 0.3759

Aggregazioni di modalità

Se un attributo ha più di 2 valori, si possono raggruppare i valori, al fine di mettere insieme valori omogenei rispetto alla variabile dipendente

Come:

Se un attributo X ha più di 2 valori, trova la coppia di valori meno significativa (con p-value più alto) rispetto alla variabile dipendente

Outlook/PlayTennis	Yes	No	Totale
Sunny	2	3	5
Rain	4	0	4
Overcast	3	2	5
Totale	9	5	14

χ_2=3.5467, d.f.=2 $p_{Outlook}$ = 0.0849

Outlook/PlayTennis	Yes	No	Totale
Sunny	2	3	5
Overcast	3	2	5
Totale	95	5	10

χ^2=4, d.f.=1 $p_{Outlook}$ =0.5164

Outlook/PlayTennis	Yes	No	Totale
Sunny	2	3	5
Rain	4	0	4
Totale	6	3	9

χ^2=3.6, d.f.=1 $p_{Outlook}$ =0.0348

Outlook/PlayTennis	Yes	No	Totale
Rain	4	0	4
Overcast	3	2	5
Totale	7	2	9

χ^2=2.0571, d.f.=1 $p_{Outlook}$ = 0.0997

Outlook/PlayTennis	Yes	No	Totale
Sunny,Rain	6	3	9
Overcast	3	2	5
Totale	9	5	14

χ^2=0.0622, d.f.=1 $p_{Outlook}$ = 1.5503

Outlook/PlayTennis	Yes	No	Totale
Sunny	2	3	5
Rain,Overcast	7	2	9
Totale	9	5	14

χ^2=1.9980, d.f.=1 $p_{Outlook}$ =0.1039

Outlook/PlayTennis	Yes	No	Totale
Sunny,Overcast	5	5	10
Rain	4	0	4
Totale	9	5	14

χ^2=3.1111, d.f.=1 $p_{Outlook}$ =0.0477

Algoritmo CHAID

Un ESEMPIO sull'efficacia del direct marketing (da Brasini,Tassinari&Tassinari)

Un'azienda contatta i propri clienti con una lettera proponendo alle famiglie di un centro urbano di sottoscrivere un abbonamento annuale ad una rivista

Costituisce così una base di dati di 81.040 famiglie classificate rispetto alla sottoscrizione dell'abbonamento (1=SI; 2=NO)

Questa è la variabile **criterio**. Le variabili **esplicative** sono: *sesso* del capofamiglia (M/F); *età* del capofamiglia (7 classi); *dimensione familiare* (da 1 a 5 =5 o più); *occupazione* del capofamiglia (I=impiegato; O=operaio; A=altro)

E' solo una piccola quota che ha sottoscritto, ma l'aziende vuole conoscere il *profilo* di queste famiglie, applicando la tecnica CHAID di *segmentazione multipla*

Le variabili esplicative e la variabile dipendente: % di successi dell'azione di DirectMKTG

1: famiglia di una persona 1,09%

2: famiglia di 2-3 persone con capofamiglia impiegato 2,39%

3: famiglia di 2-3 persone con capofamiglia **non** impiegato 1,42%

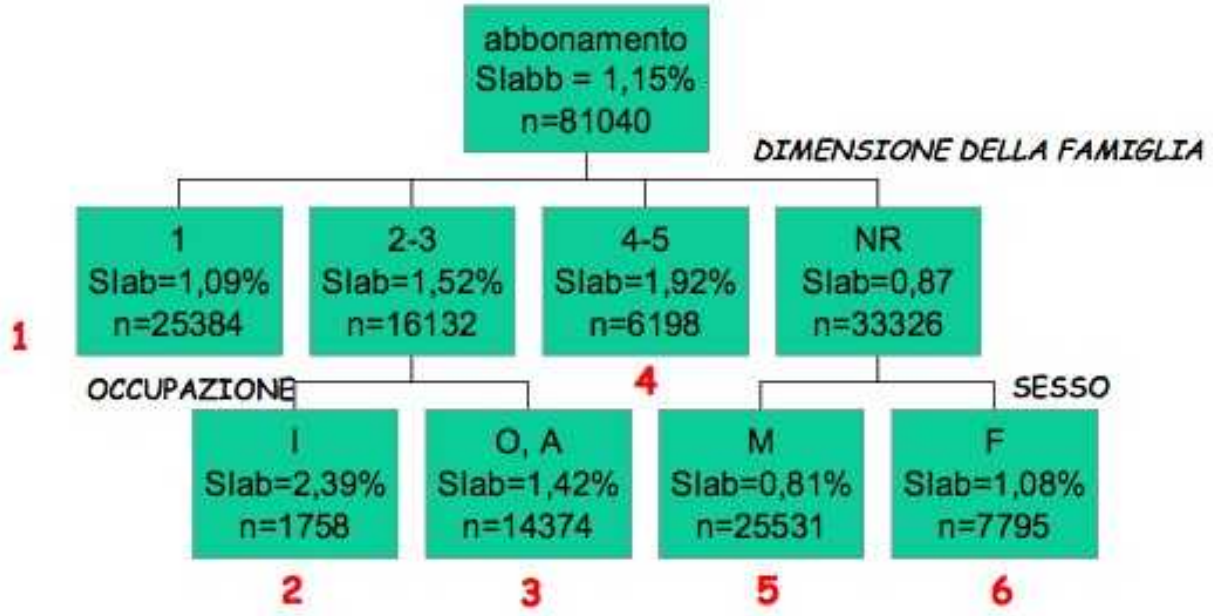
4: famiglia di 4 o 5 persone o più 1,92%

5: famiglia con numero di componenti sconosciuto e capofamiglia maschio 0,81%

6: famiglia con numero di componenti sconosciuto e capofamiglia femmina 1,08%

L'efficacia di un'azione di Direct Marketing

Albero di segmentazione con l'algoriitmo CHAID



CART *(Breiman, Friedman, Olshen – 1984, Classification And Regression Tree)*

- Variabile dipendente QUALITATIVA/QUANTITATIVA
- Variabili esplicative QUALITATIVE/QUANTITATIVE

Obiettivi:

Classification Trees

Individuare il miglior classificatore di N individui appartenenti a J gruppi che siano internamente omogenei ed esternamente eterogenei (*variabile di risposta in classi*)

Regression Trees:

Predire una variabile dipendente (continua) attraverso M variabili esplicative
(*variabile di risposta continua*)

CART: il criterio di split

Obiettivo:

Generare nodi figli che siano più "puri" del nodo genitore

- **Classificazione ad albero**

Generare nodi figli **omogenei**

*perché con una **proporzione minima di individui** di classi **differenti** della variabile di risposta*

- **Regressione ad albero**

Generare nodi figli con **varianza** della variabile di risposta **minore** del nodo genitore

Misura dell'impurità di un nodo

- **Classificazione ad albero**
  - Indice di eterogeneità del Gini
  - Indice di entropia
- **Regressione ad albero**
  - Media ponderata delle varianze nei nodi figli

L'indice di eterogeneità di Gini

L'indice di Gini I è una misura della eterogeneità (omogeneità) di una distribuzione statistica a partire dai valori delle frequenze relative associate alle *k* modalità di una generica variabile X

Se i dati sono distribuiti in modo eterogeneo su tutte le *k* modalità di X (cioè, se le modalità hanno numerosità simili), l'indice di Gini è elevato

Il suo massimo (tutte le numerosità sono uguali) è pari a *k*/(*k*-1)

Viceversa, in caso di distribuzione di frequenza omogenea (tutti gli individui appartengono ad una sola classe) l'indice sarà pari a 0.

$$I = 1 - \sum_{i=1}^{k} f_i^2$$

Esempi di uso dell'indice di Gini

Nella scelta dello split, una volta che un nodo t è stato suddiviso in *k* nodi figlio, si calcola il cosiddetto I_{split} :

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

dove n_i è il numero di elementi nel nodo figlio i

e n è il numero di elementi nel nodo t

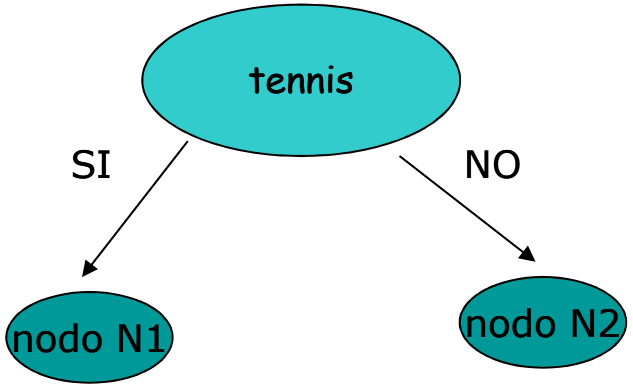
n_i/n è il peso dei vari indici di I(*i*)

Si sceglie lo split corrispondente all' I_{split} più piccolo

Per scegliere quindi il miglior *split* occorre enumerare tutte le possibili partizioni

Il caso di attributi binari

Si cerca la partizione **binaria** che produca nodi **più grandi** e **più puri** possibile:



I caso: SI = 6

NO = 6

N1	N2
6	0
0	6

I = 0,00

5	1
1	5

I = 0,271

4	2
3	3

I = 0,486

3	3
3	3

I = 0,5

Il caso di variabili nominali

1. Per ciascuna classe della variabile dipendente si calcolano le frequenze di ciascuna modalità

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

2. Si cerca la migliore partizione binaria delle categorie delle variabili esplicative

	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Il caso di variabili continue

Esistono diverse possibili strategie la più efficiente da un punto di vista computazionale è quella che, per ciascuna delle variabili quantitative:

1. ordina i valori
2. scandisce linearmente questi valori, aggiornando ogni volta la matrice dei conteggi necessario per calcolare I

   considera che, a ogni passo, un singolo elemento (appartenente ad una certa classe) passa da una partizione all'altra (• +/- 1 in una particolare riga)

3. si sceglie lo split con l'I minore

Mente	No		No		No		SI		SI		SI		No		No		No		No	
	Reddito tassabile																			
	60		70		75		85		90		95		100		120		125		220	
	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
SI	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
NO	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
I	0.420		0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400		0.420	

L'entropia di Shannon

La misura di entropia H, sviluppata nell'ambito della teoria dell'informazione misura la quantità di incertezza, o di informazione contenuta in un segnale

$$H = -\sum f_i \lg(f_i)$$

essa assume valore minimo zero quando tutte le unità del collettivo presentano la stessa modalità e valore massimo log k nel caso di equiripartizione delle unità tra le diverse modalità:

$$H_{\max} = -\sum_{i=1}^{k} \left[\frac{1}{k} \lg\left(\frac{1}{k} \right) \right]$$

CART: Qualità della regola di classificazione

- Stima del Tasso di Errata Classificazione:

  R = numero individui mal classificati / numero di individui nel campione

Nota bene – R si calcola sullo stesso campione su cui si è calcolato l'albero di classificazione

- Metodo del campione test

  $N = N_{base} + N_{test}$

Il campione si partiziona in un campione base (o *training set*), di dimensione N_{base} su cui si costruisce l'albero ed un campione test (o *validation test*), di dimensione N_{test}, su cui si controlla la bontà dei risultati

R = numero individui mal classificati nel campione test / N_{test}

CART: Il pruning

CART ha un comportamento peculiare nel confronto della regola di arresto

In realtà l'algoritmo è composta da 2 fasi principali:

1. la generazione dell'albero completo
2. il *pruning* dell'albero

Un albero completo, infatti, è spesso costituito da un numero molto elevato di rami e di foglie

Conviene sfoltire (POTARE) la ridondanza dell'albero, eliminando i rami meno interssanti

CART: Generazione dell'albero

Lo schema più classico della procedura CART parte da una suddivisione del set di dati in un *training-set* e in un *valiation-set*

Lavorando sul *training-set*, si comincia una suddivisione divisione binaria in classi. Il criterio di *split* si basa sul concetto generale di **diversità** (chiamata anche **gain ratio** o **information gain**)

Questa grandezza è una funzione della frazione di record classificati in ciascuna classe. Indicando con p_c la frazione di record classificati nel nodo c, possiamo definire una funzione **diversità** che fornisce una misura di quanto i record sono equamente distribuiti nelle classi

Nei due casi limite i record sono rispettivamente concentrati solo in una classe (*omogeneità*) e divisi in parti uguali fra le due classi (*eterogeneità*). Abbiamo visto 2 fra le più comuni misure (Gini, Entropia)

CART: L'algoritmo di generazione dell'albero

L'albero viene generato nodo per nodo, individuando il mglior split

1. **Sort**. Per ogni campo si ordinano tutti i possibili valori in base a quanto permettono di diminuire la funzione di diversità prescelta
2. **Max**. Fra i migliori *split*, si prende quello quello che minimizza la diversità
3. Si procede **iterando** dal punto 1 fino a quando non si riesce più a diminuire la funzione di diversità.

CART: Pruning

L'algoritmo descritto precedentemente genera **tutto** l'albero sulla base di un *training set* . Per applicare efficacemente l'albero ottenuto ad altri dataset è necessario eliminare alcuni rami, diminuendo l'errore complessivo dell'albero

Si costruisce l'albero massimo Tmax con H nodi terminali utilizzando il *training set*

1. Si costruisce l'insieme C_h di tutti i sotto alberi di Tmax con h < H nodi terminali
2. Si sceglie il migliore sottoalbero T_h per ciascuna tipologia di sottoalberi, sulla base di un criterio che può essere il tasso di corretta classificazione:

   $$R(T_h) = \min \{R(T) \; T \in C_h\}$$

4. Sul Validation set, si sceglie il sottoalbero T* con tasso di errore minimo

   $$R(T^*) \min\{R^{ts}(T_h); 0 \le h \le H\ 1\}$$

CART: Vantaggi e Svantaggi

- PRO:
  - Facile leggibilità della regola di classificazione
  - Selezione automatica delle variabili discriminanti
  - utilizzo di criteri di selezione non parametrici

- CONTRO
  - Possibilità di cadere in regole di classificazioni banali
  - Difficoltà computazionali legate al calcolo, in ciascun nodo dei valori di split per tutte le possibili dicotomizzazioni

Come funziona CART (dal manuale SPSS) Notazioni

Y variabile dipendente (ordinale, nominale o continua)
Se Y è nominale ha J classi ($C = 1, \ldots, J$}
$\mathbf{X}_m$, $m = 1, \ldots, M$ è l'insieme dei predittori. $\mathbf{X}_m$ può essere ordinale, nominale, o continua

$\hbar = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$ è l'intero campione di apprendimento

$\hbar(t)$ è il campione di apprendimento nel nodo t

$p(j,t)\ ,\ j = 1, \ldots, J$ è la probabilità di un caso nella classe j nel nodo t

$p(t)$ è la probabilità di un caso nel nodo t

$p(j \mid t)$, $j = 1, \ldots, J$ è la probabilità di un caso nella classe j dato che cade nel nodo t

$C(i \mid j)$ è il costo di cattiva classificazione di un elemento nella classe i nella classe j , con, ovviamente:

$C(j \mid j) = 0$

La costruzione dell'albero

L'idea di base è scegliere la suddivisione fra tutte le possibili ad ogni nodo, che produca il nodo figlio più "puro"

In questo algoritmo si considerano tutti i possibili split, ma solo: ogni *split* dipende dal valore di un solo predittore

Si parte dalla radice, iterando i seguenti passi:

1. *si individua il miglior split per ogni predittore: per ogni predittore continuo oppure ordinale, si ordinano i valori dal più piccolo al più grande. Si esaminano tutti i possibili punti candidati per essere uno split, per trovare il migliore, ossia quello che massimizza il criterio di split*

   Per ogni predittore nominale, si esaminano tutti i possibili sottoinsiemi di modalità, per trovare il miglior split
2. *Fra i migliori splitindividuati nella fase 1., si sceglie quello che massimizza il criterio di split*
3. *si diveide il nodo per cui si è identificato il miglior split nella fase 2, SE la REGOLA di ARRESTO non è soddisfatta*

Regole di arresto

- Se un nodo diventa puro; ossia se tutti i casi nel nodo hanno un valore identico per la variabile dipendente
- Se nel nodo i predittori hanno tutti lo stesso valore
- Se l'albero ha raggiunto la dimensione massima prefissata
- Se la dimensione di un nodo è inferiore alla dimensione minima prefissata
- Se lo *split* produce un nodo figlio più piccolo di quanto prefissato

Indagine sugli sbocchi occupazionali dei laureati in Economia e Commercio di Napoli (1997)

- **Voto di laurea**

1. Basso (VOT1) 2. Alto (VOT2)

- **Genere**

1. maschio (MASC) 2. Femmina (FEMM)

- **Residenza**

1. Napoli (RENA) 2. provincia di Napoli (REPR) 3. altre province (REAP)

- **Età attuale**

1. minore di 25 anni (ETA1) 2. tra 26 e 30 anni (ETA2)
3. tra 31 e 35 (ETA3) 4. oltre 30 anni (ETA4)

- **Diploma**

1. maturità classica (DICL) 2. maturità scientifica (DISC)
3. diploma tecnico (DITN) 4. Magistrale (DIMA) 5. altri diplomi (DIPR)

- **Piano di studi**

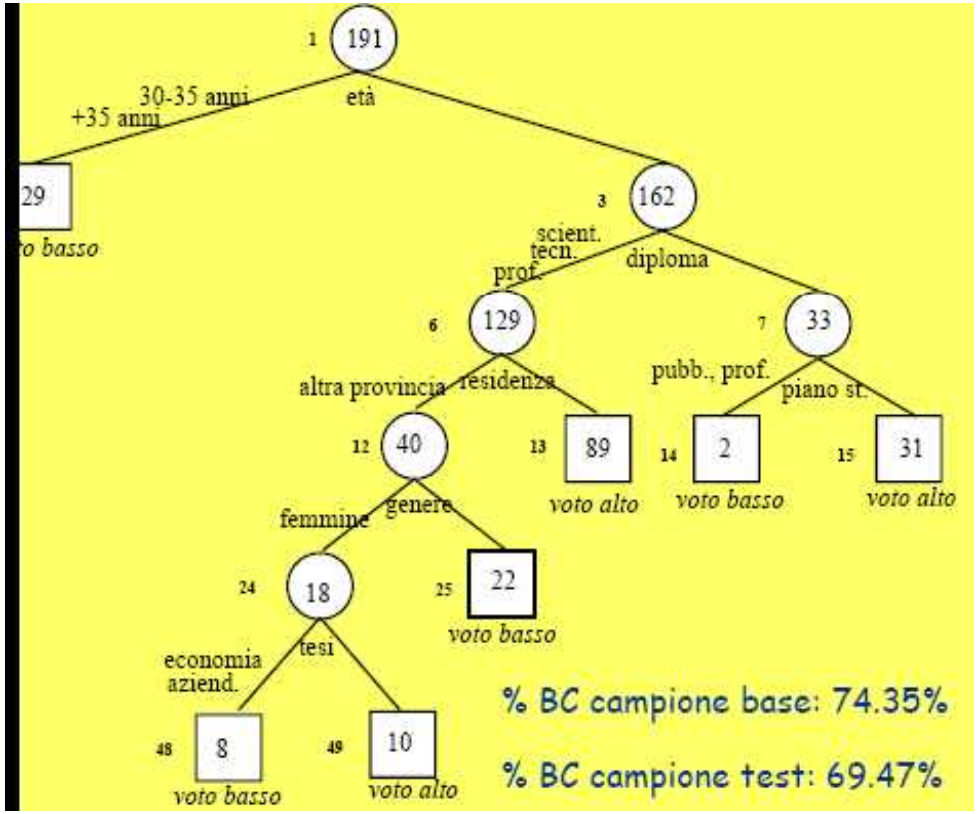
1. individuale (PUIN) 2. Aziendale (PIAZ) 3. Generale (PIGE)
4. Quantitativo (PIMA) 5. Pubblico (PIPU) 6. Professionale (PIPR)

- **Anni impiegati per la laurea**

1. 4 anni (ANN1) 2. 5-6 anni (ANN2)3. >6 anni (ANN3)

- **Tesi di laurea**

1. T-economiche (TEEC) 2. T-giuridiche (TEGI) 3. T-quantitative (TEQU)4. T-storiche soc. e geog. (TESS) 5. T-aziendali (TEAZ)


1
191
età
30-35 anni
+35 anni
29
o basso
3
162
scient.
tecn.
prof.
diploma
6
129
7
33
altra provincia
residenza
pubb., prof.
piano st.
12
40
13
89
14
2
15
31
voto alto
voto basso
voto alto
genere
femmine
24
18
25
22
voto basso
economia
aziend.
tesi
48
8
49
10
voto basso
voto alto
% BC campione base: 74.35%
% BC campione test: 69.47%

Dati medici

VARIABILI:

- peso del bambino alla nascita

1. peso maggiore o uguale a 2500 gr *2. peso inferiore a 2500 gr*

- età della mamma
- peso della mamma prima della gravidanza
- razza della mamma

1. bianca *2. nera* *3. altro*

- vizio del fumo

1. sì *2. no*

- storia di precedenti aborti

1. nessun aborto precedente *2. almeno un aborto*

- problemi di ipertensione

1. sì *2. no*

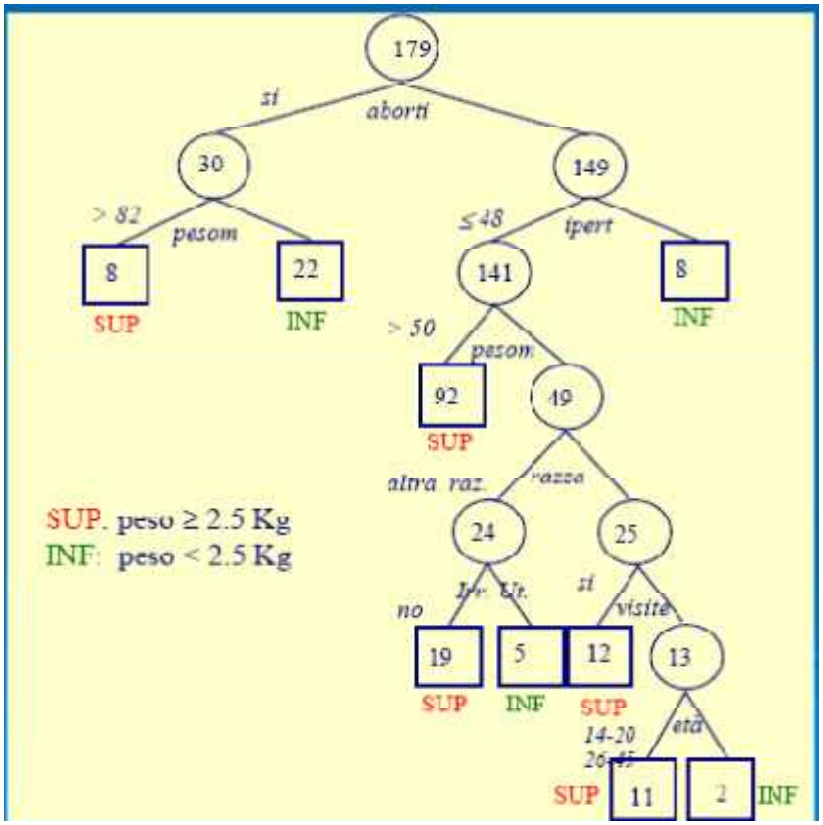
- problemi di irritabilità uterina

1. sì *2. no*

- visite ginecologiche durante il primo trimestre

1. nessuna visita *2. almeno una visita*

INDIVIDUI: 189 donne in stato di gravidanza


179
si
aborti
30
149
> 82
pesom
≤ 48
ipert
8
SUP
22
INF
141
8
INF
> 50
pesom
92
SUP
49
altra raz.
razza
24
25
no
si
visite
19
SUP
5
INF
12
SUP
13
età
SUP
11
2
INF
SUP. peso ≥ 2.5 Kg
INF: peso < 2.5 Kg

Allocazione degli individui nelle classi

Campione base

	Classe di origine	
Classe di assegnazione	SUP	INF
SUP	112 *91.1*	30 *53.6*
INF	11 *8.9*	26 *46.4*
TOTALE	123 *100.0*	56 *100.00*

Campione test

	Classe di origine	
Classe di assegnazione	SUP	INF
SUP	5 *71.4*	3 *100.0*
INF	2 *28.6*	0 *.0*
TOTALE	7 *100.0*	3 *100.0*

Qualità della clasificazione

Campione di base

C	EFF	TOT	%
SUP	112	123	91.06
INF	26	56	46.43
Tot	138	179	77.09

- **C**: classe di assegnazione;
- **EFF**: numero di individui ben classificati;
- **TOT**: numero di individui originariamente collocati in ogni classe;
- **%**: percentuale di ben classificati (EFF/TOT);

Campione di convalida

C	EFF	TOT	%
SUP	5	7	71.43
INF	0	3	0.0
Tot.	5	10	50.00

L'algoritmo C4.5 1

L'algoritmo C4.5 è una evoluzione dell'algoritmo ID3 (Iterative Dichotomiser 3) e si basa su un'euristica collegata al Rasoio di Occam: la scelta è quella che privilegia l'albero più piccolo (anche se non produce necessariamente l'albero più piccolo)

Il criterio è quello dell'Entropia di Shannon, calcolato su un insieme di dati di apprendimento.

L'algortmo funziona così:

I dati di apprendimento sono un insieme S = $s_1,s_2,\ldots$ di campioni già classificati. Ogni campione $s_i = x_1,x_2,\ldots$ È un vettore in cui $x_1,x_2,\ldots$ Sono gli attributi o le caratteristiche che descrivono il campione

I dati di apprendimento sono aumentati, aggiungendo un vettore $C = c_1,c_2,\ldots$ dove $c_1,c_2,\ldots$ rappresentano le classi di appartenenza di ciascun elemento.

L'algoritmo C4.5 2

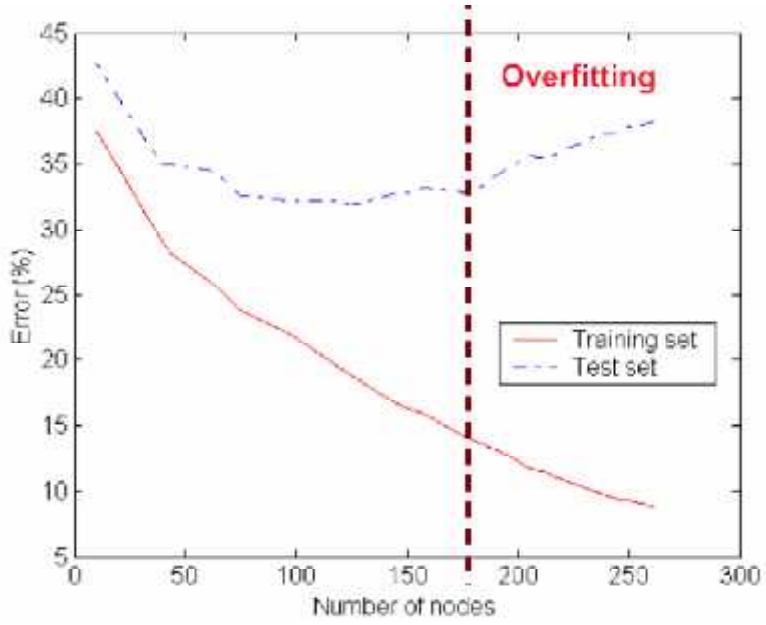
C4.5 si basa sull'idea che ciascun attributo può essere utilizzato per prendere una decisione sulla suddivisione dei dati in sottoinsiemi di dimensione minore. C4.5 utilizza il cosiddetto information gain (la differenza fra le entropie) normalizzato, per scegliere il miglior attributo per lo split: l'attributo con l' information gain normalizzato più alto è quello utilizzato per lo split. L'algoritmo procede sui sottoinsiemi identificati. Quando tutti i campioni considerati appartengono alla stessa classe, allora si è identificata una foglia. Può anche succedere che nessuna delle caratteristiche produca un information gain, in questo caso C4.5 crea un nodo al livello superiore. Lo stesso accade se una classe dovesse risultare vuota

Qualche confronto

1. CHAID e C4.5 sono alberi di segmentazione multipla, AID e CART binaria
2. Mentre gli altri metodi utilizzano un solo data set, CART ha un *training set* e un *validation set*
3. CHAID e AID utilizzano criteri statistici per la regola di stop, mentre CART e C4.5 effettuano il pruning
4. CHAID produce più che un albero una serie di cespugli, ossia spesso conduce a molte foglie, provenienti da un ramo, facilmente rappresentabile in una tabella di contingenza, Questo fa sì che CHAID sia un metodo molto utilizzato nelle ricerche di mercato, per la sua capacità descrittiva, basata su un test statistico
5. CART ha un utilizzo predittivo più evidente

Overfitting

Il rischio che i metodi di segmentazione cercano di evitare, attraverso la semplificazione dell'albero è l'overfitting, ossia l'eccessivo adattamento dell'albero alle specificità del data set osservato. Questo si manifesta, infatti, come un numero eccessivo di rami, alcuni dei quali possono riflettere anomalie dovuti a rumori o valori anomali



Overfitting: Early stopping rule

Pre-Pruning:

l'algoritmo si arresta prima di produrre un albero completo

Le tipiche condizioni di arresto:

- **Stop se tutte le istanze appartengono alla stessa classe**
- **Stop se tutti i valori degli attributi hanno lo stesso valor**
- **Stop se il numero di elementi in un nodo è minore di uno soglia fornita in input:**
  - Evita la creazione di piccole partizioni
  - Difficile trovare il threshold
- **Stop se espandendo il nodo corrente non miglioriamo la misura di impurità (es.., Gini o Information Gain).**

Post-Pruning:

l'algoritmo costruisce l'albero completo

Successivamente:

- **Elimina i nodi in modo *bottom-up***
- **Usa un insieme di dati differenti dal training data**

per decidere qual è il *best pruned tree*

- **Se l'errore sui dati di test migliora dopo la sostituzione del sotto-albero con un nodo foglia, lo sostituisce in maniera permanente l'etichetta di classe da assegnare al nuovo nodo foglia è determinato dalla classe maggioritaria nel sotto albero**

Estrarre regole di classificazione dagli alberi

Rappresentare la conoscenza nella forma di regole IF … THEN è un modo facile per comunicare i risultati:

- Seguire un percorso dell'albero, lungo i rami e fino alle foglie è estremamente semplice:
  - **Ciascuna coppia attributo-valore lungo un cammino forma una *congiunzione***
  - **Il nodo foglia restituisce la predizione della classe per la regola estratta**

Esempi

IF *outlook* = "sunny" AND *Humidity* = "normal" THEN *play_tennis* = "yes"

IF *outlook* = "sunny" AND *Humidity* = "high" THEN *play_tennis* = "no"

Alcuni riferimenti bibliografici

- BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J., (1984), *Classification and regression trees*. Belmont C.A. Wadsworth.
- BOUROCHE J.M., TENENHAUS M. (1970), "Quelques méthodes de segmentation", Revue Francaise d'Informatique de Recherche Opérationelle, vol. V-2.
- CELEUX, LECHEVALLIER Y. (1982), "Methodes de segmentation non parametriques" Revue de Statistique Appliquees, n.4, vol.XXX, pagg. 39-53.
- FABBRIS L. (1989), "Analisi di segmentazione binaria mediante AID" in Fabbris, Analisi esplorativa dei dati multidimensionali, pagg. 239-275.
- FIELDING A. (1975), "Binary segmentation: the Automatic Interaction Detector and Related Techniques for Exploring Data Structure" in O'Muircheartaigh C.A., Payne C., The analysis of survey data, Vol.1, pagg. 221-257.
- GUEGUEN A., NAKACHE J.P., (1988), "Methode de discrimination basee sur la construction d'un arbre de decision binaire", Revue de Statistique Appliquees n.1, vol.XXXVI, pagg. 19-38.
- MORGAN J.N, SONQUIST J.A., (1963), "Problems in the analysis of survey data and proposal", Journal of American Statistical Association, vol. 58, pagg. 415-434
- QUINLAN J.R., (1986), "Induction of Decision Trees", Machine Learning, n.1, pagg. 81-106.