Segmentazione binaria

Statistica per le decisioni d'impresa

Segmentazione binaria

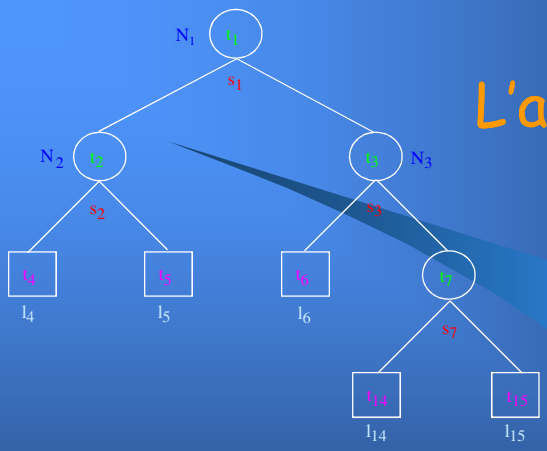
- L'obiettivo

  Classificare un collettivo di individui in gruppi (*segmentazione*), omogenei al loro interno e quanto più possibile differenziati, mediante una successione di divisioni dicotomiche (*binaria*)

- I dati
  - **p** variabili esplicative
  - 1 variabile dipendente *(qualitativa o quantitativa)*

  osservate su N individui *(campione di base)*

L'albero binario



nodi intermedi: t1, t2, t3, t7;

nodi terminali: t4, t5, t6, t14, t15;

split: s1, s2, s3, s7;

etichette: l_4, l_5, l_6, l_{14} l_{15}

Notazioni e definizioni

- **Y: variabile dipendente**
  - Qualitativa $C = (j = 1, \ldots, J)$ classi di risposta
  - Quantitativa spazio R numeri reali
- **$X = (X_1, \ldots, X_p)$: variabili esplicative o predittori**

  $\mathbf{x}_n = [x_{n1}, \ldots, x_{np}]^T$: misurazioni sull'ennesimo individuo
- **Regola di classificazione/previsione:**
  - **Y Qualitativa**

    Partizione di **X** in J gruppi $\{A_1, \ldots A_J\}$, tale che per ogni $\mathbf{x} \in A_j$

    la classe prevista è j: $A_j = \{\mathbf{x}; d(\mathbf{x}) = j\}$
  - **Y Quantitativa**

    Funzione $d(\mathbf{x})$ definita su **X** tale che per ogni **x**: $d(\mathbf{x}) = y$

Fasi di una procedura di segmentazione

1. Un insieme di domande binarie

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

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

2. Un criterio di split

per selezionare la migliore divisione di un nodo

AID **CHAID** **CART** **C4.5** **TWO-STAGE**

... fasi

3. Una regola di arresto

definire una regola per dichiarare un nodo come terminale o intermedio

numerosità dei nodi **test statistici** **pruning**

4. Una regola di assegnazione

per assegnare una classe di risposta o un valore a ciascun nodo terminale

5. Qualità della regola di decisione

stimare il tasso di errata classificazione/previsione

Classification And Regression Trees

- **Classification Trees**

Classificatore ad albero di N individui appartenenti a J gruppi attraverso una partizione dello spazio dei predittori in gruppi internamente omogenei ed esternamente eterogenei
(variabile di risposta qualitativa)

- **Regression Trees**

Regressione ad albero di una variabile dipendente numerica attraverso una partizione dello spazio dei predittori in gruppi internamente omogenei ed esternamente eterogenei
(variabile di risposta quantitativa)

CART: il criterio di split

Obiettivo

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

- ***Classificazione***

Generare nodi figli più omogenei ossia con una proporzione minima di individui appartenenti a classi di risposta differenti

- ***Regressione***

Generare nodi figli con varianza della variabile dipendente minore della varianza nel nodo genitore

Misura dell'impurità di un nodo

Indice di eterogeneità del Gini
Indice di entropia

Varianza o Devianza

L'eterogeneità in un nodo

Indice di eterogeneità del Gini

$$i(t) = 1 - \sum_j p^2(j|t)$$

proporzione di casi della classe j presenti nel nodo t

$$i_{MAX}(t) = \frac{J-1}{J}$$

L'impurità di un nodo è **massima** quando le classi sono equidistribuite al suo interno

t

Classe 1	100
Classe 2	100
Classe 3	100
	300

$$i(t) = 1 - \left[\left(\frac{100}{300}\right)^2 + \left(\frac{100}{300}\right)^2 + \left(\frac{100}{300}\right)^2\right] = 1 - \frac{1}{3} = \frac{3-1}{3} = \frac{2}{3}$$

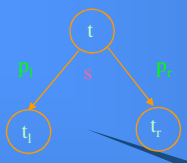
$$i_{MIN}(t) = 0$$

L'impurità di un nodo è **minima** quando esso contiene individui appartenenti solo ad una classe

t

Classe 1	300
Classe 2	0
Classe 3	0
	300

$$i(t) = 1 - \left[\left(\frac{300}{300}\right)^2 + \left(\frac{0}{300}\right)^2 + \left(\frac{0}{300}\right)^2\right] = 1 - 1 = 0$$



Il migliore split

- Impurità di un nodo

$$i(t) = 1 - \sum_j p^2(j \mid t)$$

- Decremento di impurità

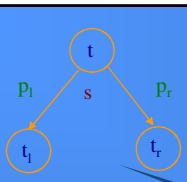
$$\Delta i(s,t) = i(t) - p_l i(t_l) - p_r i(t_r)$$

$$\Delta i(s,t) = i(t)\tau_{Y|s}(t)$$

Indice di predizione di Goodman e Kruskal

- Migliore divisione

$$\Delta i(s^*, t) = \max!$$



La riduzione in devianza

- Impurità di un nodo

$$R(t) = \frac{1}{N}\sum_{x_n \in t}(y_n - \bar{y}_t)^2$$

- Decremento di impurità

$$\Delta R(t,s) = R(t) - R(t_L) - R(t_R)$$

$$N\,\Delta R(t,s) = TSS_Y(t)\eta^2_{Y|s}(t)$$

Rapporto di correlazione del Pearson

- Migliore divisione

$$\Delta R(s^*, t) = \max!$$

La qualità della regola

- **Classificazione ad albero**
- **Regressione ad albero**

Stima del

Tasso di errata classificazione

Tasso di errata previsione

E.Q.M.

Stima del tasso di errata classificazione

- Stima basata sul campione di apprendimento *ottimistica!*

$$R = \frac{\#\ \text{individui mal classificati}}{\#\ \text{individui nel campione}}$$

$$\hat{R}(d) = \frac{1}{N}\sum_n I(d(x_n) \neq j_n)$$

R viene calcolato sullo stesso insieme di dati utilizzato per costruire la regola di classificazione

- Stima basata sul campione test *per grandi campioni!*

$S_{base} \cup S_{test}$ $S_{base} \cap S_{test} = \varnothing$

Occorre definire la proporzione N_{ts}/N

$$R_{test} = \frac{\#\ \text{individui mal classificati nel camp. test}}{N_{test}}$$

Costruzione della regola di classificazione con $N-N_{ts}$ individui

$$\hat{R}_{test}(d) = \frac{1}{N_{test}}\sum_{n \in S_{test}} I(d(x_n) \neq j_n)$$

• **Stima basata sulla cross-validation** *per piccoli campioni!*

- ✓ Il campione S è ripartito in V>2 sottocampioni $S_1, S_2, \ldots, S_v, \ldots, S_V$ di dimensione il più possibile prossima fra di loro
- ✓ La regola di classificazione $d^{(v)}(\mathbf{x})$ viene costruita sul campione $S - S_v$, per v = 1, 2, … V
- ✓ Poiché ***per ogni v*** nessun elemento di S_v è compreso in $S - S_v$,

è possibile effettuare una stima basata sul campione test di $R(d^{(v)})$:

$$\hat{R}_{test}\left(d^{(v)}\right) = \frac{1}{N_v}\sum_{n\in S_v} I\left(d^{(v)}(x_n) \neq j_n\right)$$

ottenendo la stima finale →

$$\hat{R}_{cv}(d) = \frac{1}{V}\sum_{v=1}^{V}\hat{R}_{test}\left(d^{(v)}\right)$$

Se V è sufficientemente elevato, ogni classificatore $d^{(v)}(\mathbf{x})$ viene costruito utilizzando un campione di dimensione $N(1 - 1/V) \approx N$.

Stima del tasso di errata previsione

• **Stima basata sul campione di apprendimento**

$$\hat{R}(d) = \sum_{h\in H_T} R(h) = \frac{1}{N}\sum_{h\in H_T} TSS_Y(h)$$

Devianza interna al nodo terminale h per $h \in H_T$ dove T è l'albero finale associato alla regola d

• **Stima basata sul campione test**

$$\hat{R}_{test}(d) = \sum_{h\in H_T} R_{test}(h)$$

• **Stima basata sulla cross-validation**

$$\hat{R}_{cv}(d) = \frac{1}{V}\sum_{v=1}^{V}\hat{R}_{test}\left(d^{(v)}\right)$$

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***

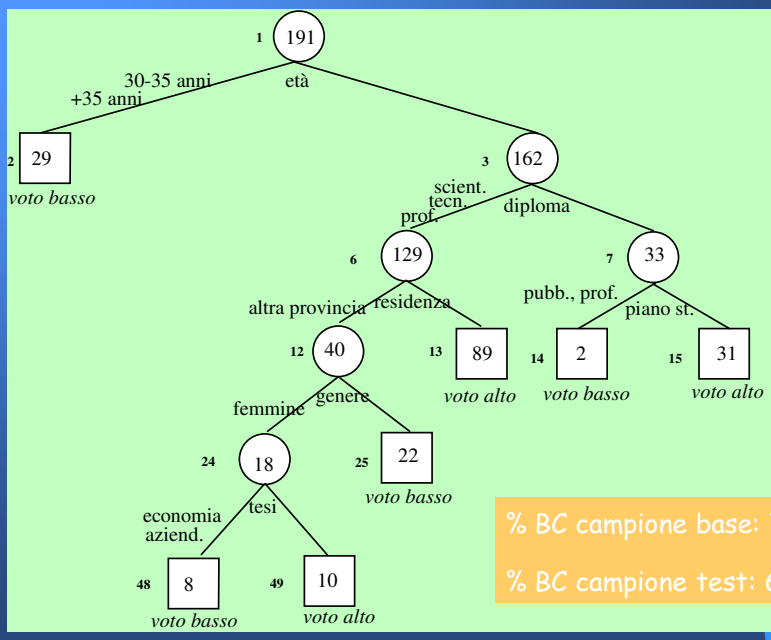
1. 4 anni (*ANN1*) 2. 5-6 anni (*ANN2*) 3. >6 anni (*ANN3*)

- ***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*)

Segmentazione

17



% BC campione base: 74.35%

% BC campione test: 69.47%

Segmentazione

Roberta Siciliano

Dall'albero esplorativo all'albero delle decisioni

Le procedure ad albero seguono una strategia ***divide et impera***

Svantaggi

Complessità:

la struttura risultante da una procedura ricorsiva tende ad essere di dimensioni elevate

difficoltà interpretative

Overfitting:

molte delle branche riflettono caratteristiche particolari dei dati impiegati piuttosto che relazioni sottostanti realmente esistenti tra la variabile di risposta ed i predittori

scarsa accuratezza

L'albero delle decisioni in CART

1. Creazione dell'albero massimo

   ⇨ $\max_{s \in S} \Delta imp(s,t) < \beta$

2. Selezione dei sottoalberi

   ⇨ $T_{\max} \supseteq T_1 \supset T_2 \ldots \supset \{t_1\}$

3. Scelta dell'albero finale

   ⇨ $\tilde{R}(T^*) = \min_k \tilde{R}(T_k)$

Pruning selettivo in CART

Funzione di costo-complessità

per il nodo t

$$R_\alpha(t) = R(t) + \alpha$$

Tassi di errore sul campione di apprendimento

per la branca che si diparte da t

$$R_\alpha(T_t) = R(T_t) + \alpha|\tilde{T}_t|$$

Numero di nodi terminali della branca

$$R_\alpha(t) \geq R_\alpha(T_t)$$

penalità come ...

$$\alpha_t = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

Aumento del tasso di errore per nodo terminale

Segmentazione

La selezione dell'albero finale

- **I fase**

  Viene generata una sequenza nidificata di sottoalberi potando ad ogni passo il sottoalbero che si diparte dal nodo che presenta il più piccolo valore di α **(weakest link)**

- **II fase**

  L' albero finale delle decisioni viene scelto

  - Selezionando nella sequenza l'albero che presenta il minimo tasso di errore sul ***campione test*** **(0-SE rule)**
  - Scegliendo l'albero il cui tasso di errore sul campione test è compatibile con il minimo, compatibilità valutata in funzione dell'errore standard **(1-SE rule)**

Segmentazione

22

Aspetti innovativi in CART

- Funzione di impurità
- Pruning
- Cross-validation
- Trattamento congiunto di predittori qualitativi e quantitativi
- Surrogate splits
- Competitors

Riflessioni sulla segmentazione

- **Vantaggi**
  - Facile leggibilità della regola di classificazione/predizione
  - Selezione automatica delle variabili maggiormente discriminanti
  - Utilizzo di criteri di selezione non parametrici
- **Svantaggi**
  - Possibilità di cadere in regole di classificazione semplicistiche
  - Difficoltà computazionali legate al calcolo, in ciascun nodo, dei valori del criterio di split per tutte le possibili domande binarie