Alberi decisionali

- Si consideri il dataset

Istanza	a1	a2	classe
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-

- Si costruisca l'albero seguendo l'algoritmo di c4.5 nell'ipotesi che il minimo numero di esempi in almeno due sottoinsiemi sia pari a 2

Risposta

- Nodo radice: info(T)=1
- Test su A1: $info_{A1}(T)=3/6*(-2/3*\log_2 2/3-1/3*\log_2 1/3)+3/6*(-1/3*\log_2 1/3 -2/3*\log_2 2/3)= 0.5*0.92+ 0.5*0.92=0.92$

  gain(A1)=1-0.92=0.08
- Test su A2: infoA2(T)=$4/6*(-2/4*\log_2 2/4 -2/4*\log_2 2/4)+2/6(-1/2*\log_2 1/2 -1/2*\log_2 1/2)=0.66*1+0.33*1=1$

  gain(A2)=(1-1)=0
- Viene preferito A1

Risposta

- $T_{A1=T}$={

1	T	T	+
2	T	T	+
3	T	F	-

}

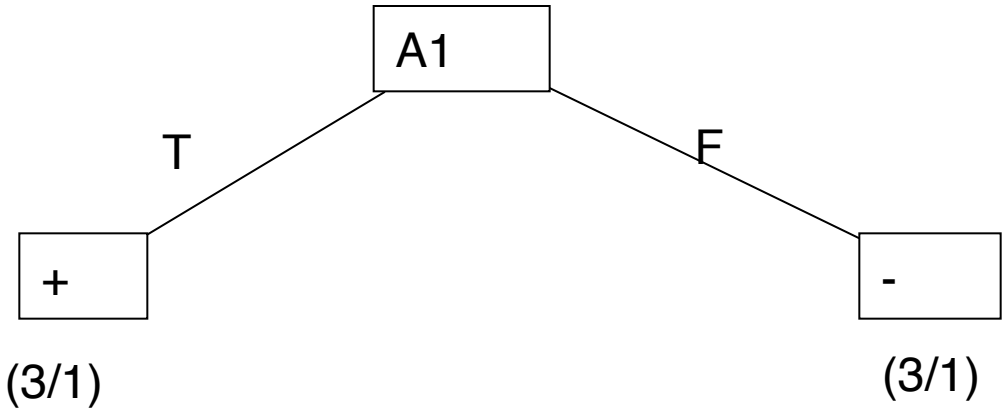
$T_{A1=F}$={

4	F	F	+
5	F	T	-
6	F	T	-

}

- c4.5 si ferma in quanto $T_{A1=T}$ e $T_{A1=F}$ non possono essere suddivisi in modo che almeno due sottoinsiemi abbiamo almeno due elementi

Risposta



Alberi decisionali

- Il maestro Yoda e' preoccupato dal numero di apprendisti Jedi che hanno deciso di darsi al Lato Oscuro, quindi ha deciso di apprendere un albero di decisione su alcuni dati storici per identificare i casi problematici in futuro.
- La tabella T descrive 12 iniziati specificando se sono passati al Lato Oscuro sulla base dell'eta' in cui il loro apprendistato Jedi e' cominciato, se hanno completato il loro apprendistato, la loro disposizione generale e la loro specie.

Tabella T

Eta' di inizio dell'apprendistato	Apprendistato completato	Disposizione	Specie	Lato Oscuro
5	1	Felice	Umana	0
9	1	Felice	Gungan	0
6	0	Felice	Wookie	0
6	1	Triste	Mon Calamari	0
7	0	Triste	Umana	0
8	1	Arrabbiata	Umana	0
5	1	Arrabbiata	Ewok	0
9	0	Felice	Ewok	1
8	0	Triste	Umana	1
8	0	Triste	Umana	1
6	0	Arrabbiata	Wookie	1
7	0	Arrabbiata	Mon Calamari	1

Domande

1. Qual'e' l'entropia della tabella T rispetto all'attributo Lato Oscuro?

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right)$$

2. Dire ad occhio (senza calcolare la funzione euristica) quale attributo verrebbe scelto come radice dell'albero dall'algoritmo di apprendimento di alberi di decisione?
3. Qual'e' il guadagno di informazione dell'attributo scelto nella risposta precedente?

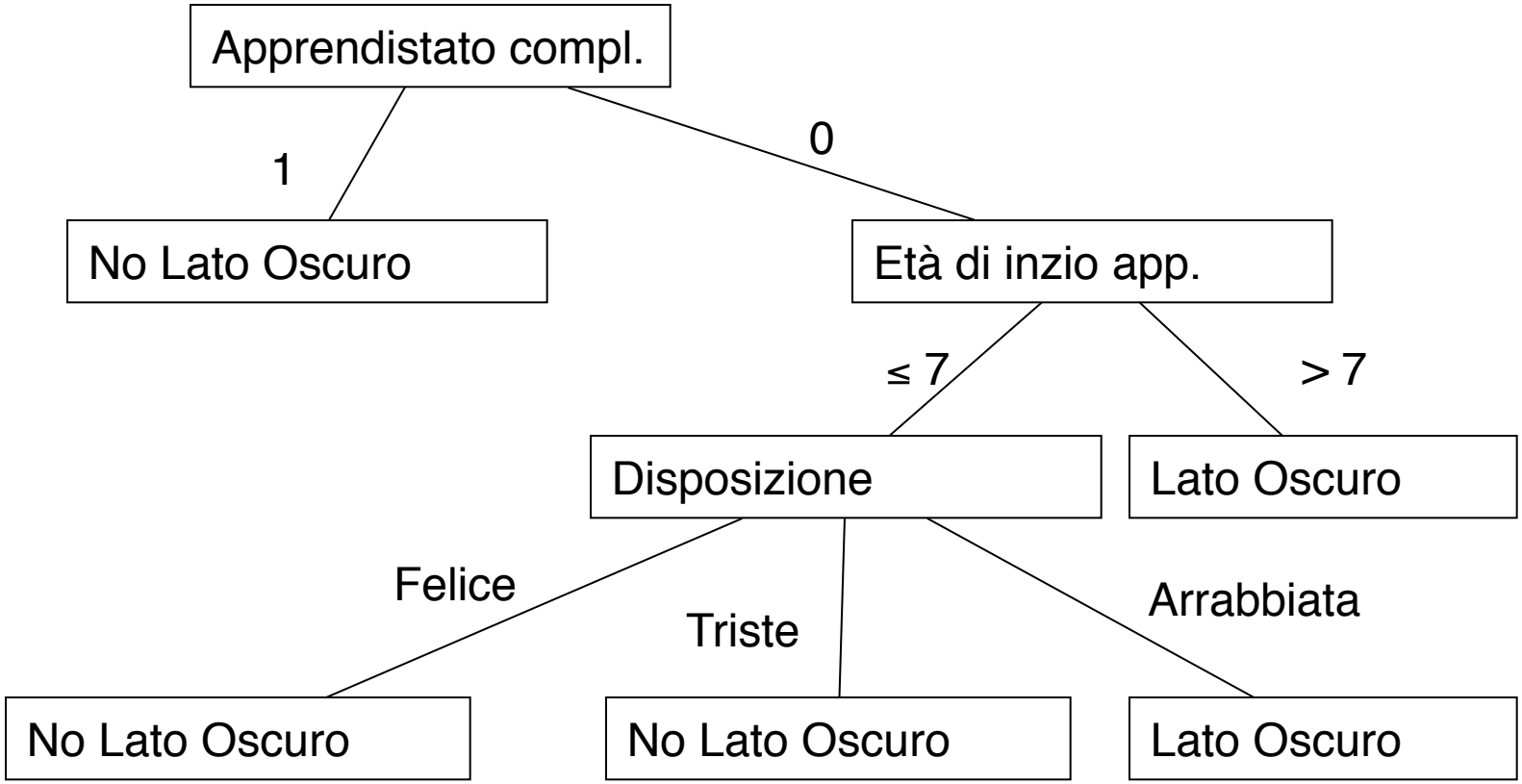
Risposte

1. info(T)=-5/12$\log_2$(5/12)-7/12$\log_2$(5/12)=0.980
2. Apprendistato completato
3. gain(App)=info(T)-(5/12*(-0/5$\log_2$(0/5)-5/5$\log_2$(5/5))
   +7/12(-5/7$\log_2$(5/7)-2/7$\log_2$(2/7)))=0.476

Domanda

4. Disegnare l'albero decisionale che sarebbe appreso da questi dati (suggerimento: l'albero ha al massimo 3 divisioni. Si costruisca l'albero guardando solo la composizione degli insiemi, senza calcolare il guadagno di informazione)

Risposta


Apprendistato compl.
1
0
No Lato Oscuro
Età di inzio app.
≤ 7
> 7
Disposizione
Lato Oscuro
Felice
Triste
Arrabbiata
No Lato Oscuro
No Lato Oscuro
Lato Oscuro

Esercizio

- Si consideri la seguente tabella

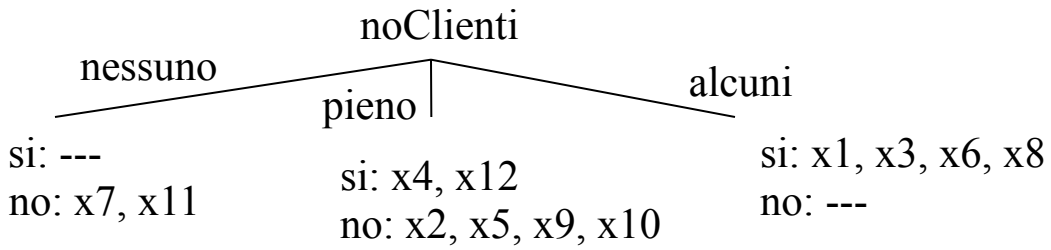
Es.	attributi										Dec
	alt.	bar	V/S	fame	noC	prez	piov	pren	tipo	att	
x1	si	no	no	si	alc	£££	no	si	F	0-10	Si
x2	si	no	no	si	pieno	£	no	no	Thai	30-60	No
x3	no	si	no	no	alc	£	no	no	hamb	0-10	Si
x4	si	no	si	si	pieno	£	no	no	Thai	10-30	Si
x5	si	no	si	no	pieno	£££	no	si	F	>60	No
x6	no	si	no	si	alc	££	si	si	I	0-10	Si
x7	no	si	no	no	ness	£	si	no	hamb	0-10	No
x8	no	no	no	si	alc	££	si	si	Thai	0-10	Si
x9	no	si	si	no	pieno	£	si	no	hamb	>60	No
x10	si	si	si	si	pieno	£££	no	si	I	10-30	No
x11	no	no	no	no	ness	£	no	no	Thai	0-10	No
x12	si	si	si	si	pieno	£	no	no	hamb	30-60	Si

Esercizio

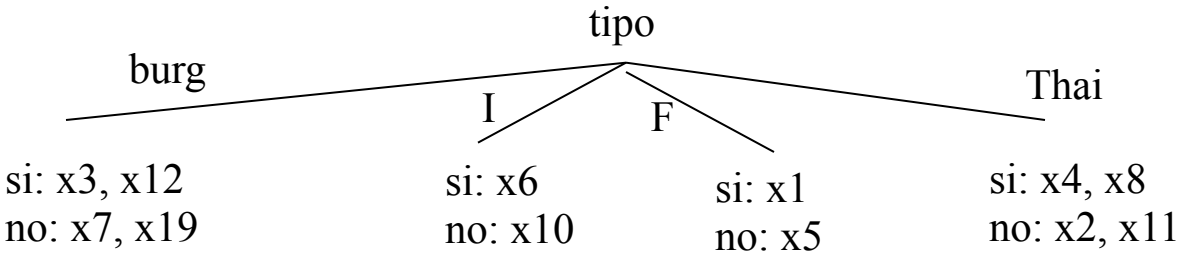
- Si costruisca un albero di decisione a partire dalla tabella.
  - Al primo passo si scelga usando il criterio del guadagno quale tra gli attributi no Clienti e tipo e' piu' conveniente usare.
  - Dopo questa scelta si prosegua selezionando l'attributo fame, poi l'attributo tra noClienti e tipo non ancora usato per la generazione dell'albero e poi l'attributo ven/sab.
  - NOTA: si prosegua fino a trovare foglie omogenee senza usare il criterio di terminazione di C4.5

Soluzione

– attributo **noClienti**: per due valori discrimina completamente (“nessuno” e “alcuni”)



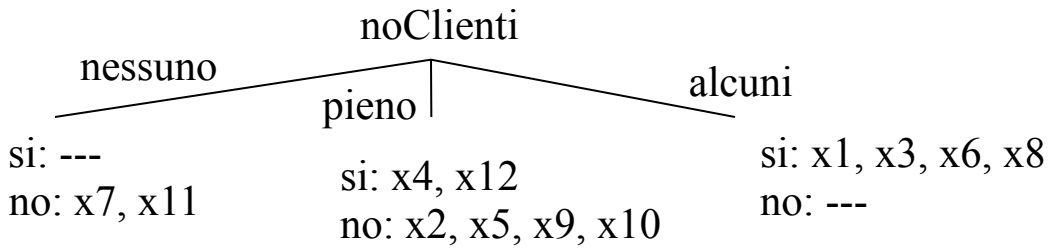
– attributo **tipo**: discrimina male per tutti i valori



– Tra i due noClienti è la scelta migliore
– in generale tra tutti è quello con entropia più bassa

Soluzione

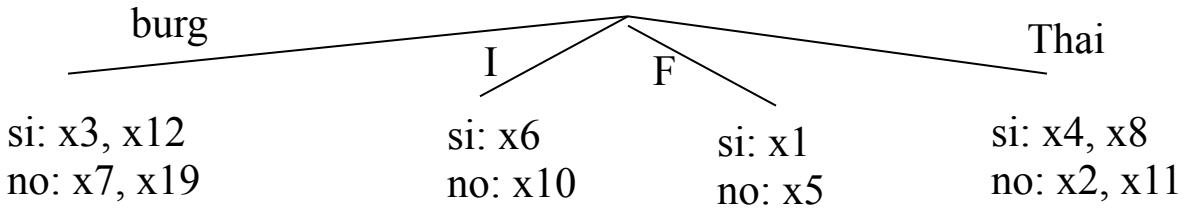
- Esempio
  - attributo **noClienti**: per due valori discrimina completamente (“nessuno” e “alcuni”)



  - info(T) = - 1/2 $\log_2$ (1/2) - 1/2 $\log_2$ (1/2) = 1
  - info($T_{clienti}$) = 2/12 info(T_{alcuni}) + 4/12 info(T_{ness}) + 6/12 info(T_{pieno})
    - info(T_{alcuni}) = 0
    - info(T_{ness}) = 0
    - info(T_{pieno}) = -2/6 $\log_2$ (2/6) - 4/6 $\log_2$ (4/6)
  - gain($T_{clienti}$) = 1- info($T_{clienti}$) = 0.541

Soluzione

- Esempio
  - attributo **tipo**: discrimina male per tutti i valori

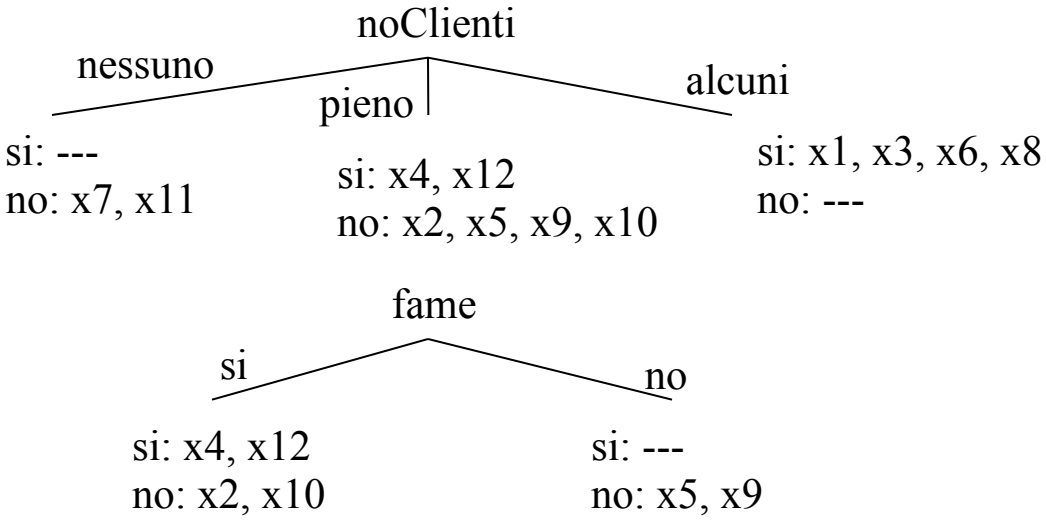


- info(T) = - 1/2 $\log_2$ (1/2) - 1/2 $\log_2$ (1/2) = 1
- info(T_{tipo}) = 2/12 info(T_I) + 2/12 info(T_F) + 4/12 info(T_{burg}) + 4/12 info(T_{Thai})
  - info(T_I) = 1
  - info(T_F) = 1
  - info(T_{burg}) = 1
  - info(T_{Thai}) = 1
- gain(T_{tipo}) = 1- 1 =0

Soluzione

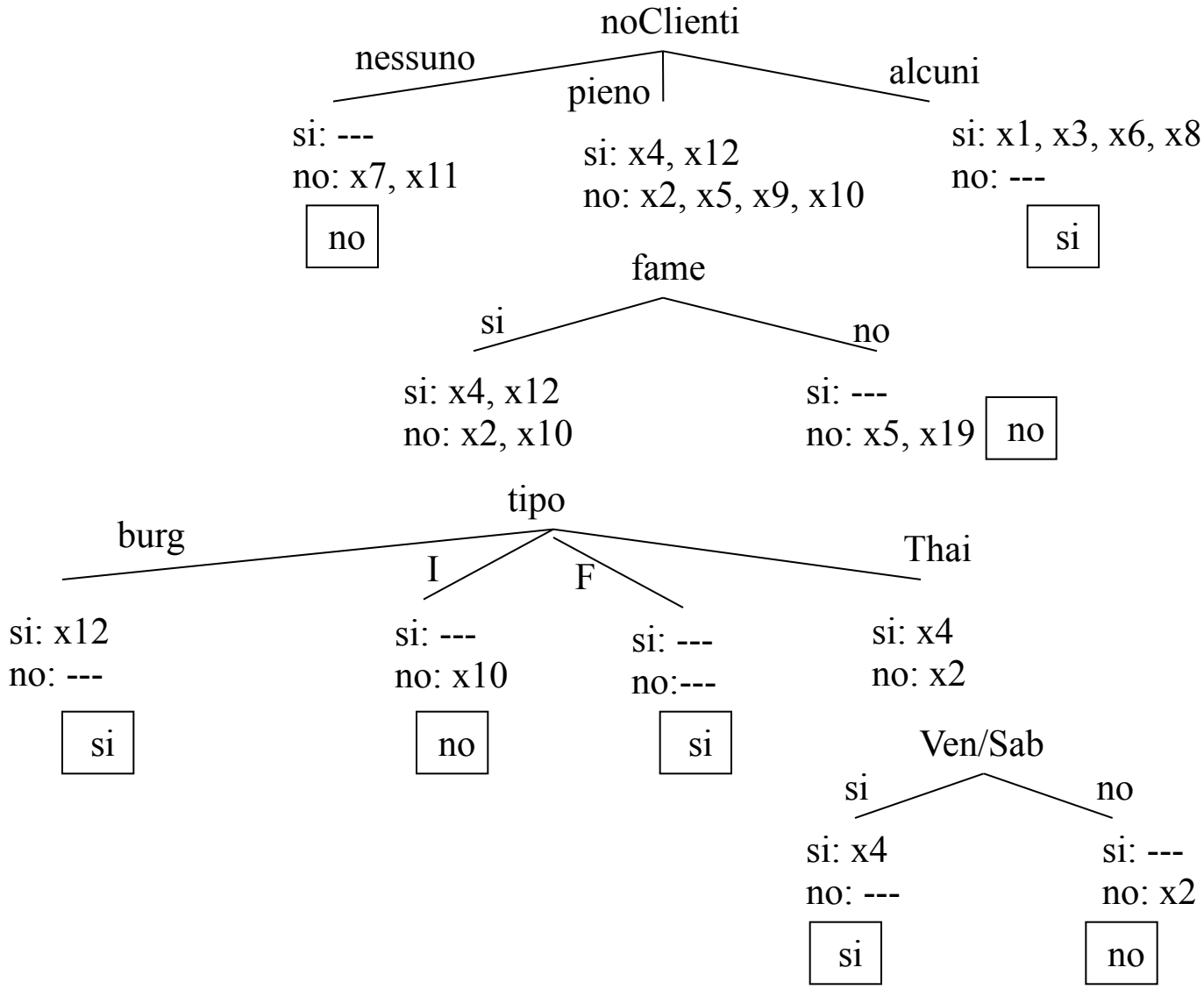
– L'algoritmo procede ricorsivamente considerando il valore "pieno" di noClienti e considerando gli esempi per quel valore
  - si analizzano gli altri attributi
  - si seleziona quello che discrimina meglio

  nel caso "fame": per uno dei due valori si ha classificazione completa



– Analogamente si procede ricorsivamente su ramo "si"

Albero risultante



•ESERCIZIO ALBERI DECISIONALI

- Si consideri il seguente Training set

Distanza	Uguali	Stessa entità
1	Si	Si
2	No	Si
3	No	No
1	?	Si
2	Si	No
2	No	Si
3	Si	No
2	No	No
1	Si	Si
1	No	No
2	No	Si
?	Si	No
3	Si	Si
3	No	No
3	Si	No
1	Si	Si
3	?	No

•ESERCIZIO ALBERI DECISIONALI

a) Si calcoli l'entropia del training set rispetto all'attributo Stessa_entita

- b) Si calcoli il rapporto di guadagno dei due attributi rispetto a questi esempi di training

- c) si costruisca un albero decisionale ad un solo livello per il training set dato, indicando le etichette delle foglie (numero di esempi finiti nella foglia/numero di esempi finiti nella foglia non appartenenti alla classe della foglia)

•ESERCIZIO ALBERI DECISIONALI

a) 8 #pos, 9 #neg, 17 #tot

- info(S)=-8/17*$\log_2$ 8/17-9/17*$\log_2$ 9/17=0.998
-
- b) Per calcolare il guadagno dell'attributo Distanza non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno Distanza noto (insieme F):

- info(F)=-8/16*$\log_2$ 8/16– 8/16*$\log_2$ 8/16=1
- $info_{Distanza}$(F)=5/16*(-4/5*$\log_2$4/5-1/5*$\log_2$

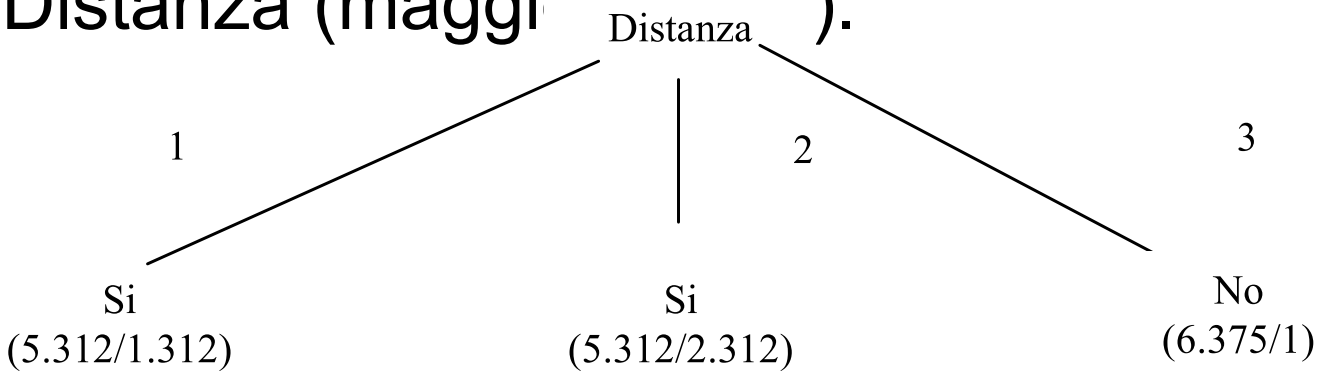
•ESERCIZIO ALBERI DECISIONALI

- Per calcolare il guadagno dell'attributo Uguali non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno Uguali noto (insieme F):

- info(F)=-7/15*log2 7/15– 8/15*log2 8/15=0.997
- infoUguali(F)= 8/15(-4/8*log2 4/8-4/8*log2 4/8) +7/15*(-3/7*log2 3/7-4/7*log2 4/7) =
- =0.533 *1+0.467*0.985=0.993

•ESERCIZIO ALBERI DECISIONALI

- c) L'attributo scelto per la radice dell'albero è Distanza (maggi).



•5.312/17=0.312

•6.375/17=0.375.

ESERCIZIO ALBERI DECISIONALI

•La prima parte viene mandata lungo il ramo 1 e viene classificata come

•Si con probabilità 4/5.312=75.3% e come

•No con probabilità 1-75.3%=24.7%.

•La seconda parte viene mandata lungo il ramo 2 e viene classificata come Si con probabilità 3/5.312 =56.5% e come

•No con probabilità 1-56.5%=43.5%.

•La terza parte viene mandata lungo il ramo 3 e viene classificata come

•No con probabilità 5.375/6.375 =84.3% e come

•ESERCIZIO ALBERI DECISIONALI

- Si consideri il seguente Training set

Helical	**Single**	**Classe**
Uno	Sì	histone
Tre	No	ire
Due	No	histone
Uno	No	ire
Tre	No	ire
Due	?	histone
Uno	Sì	ire
Tre	Sì	histone
Due	Sì	histone
Tre	No	ire
Uno	Sì	histone
Tre	No	ire
Due	Sì	ire
Uno	No	histone
Due	?	ire
Tre	Sì	histone

•ESERCIZIO ALBERI DECISIONALI

a) Si calcoli l'entropia del training set rispetto all'attributo Classe

- b) Si calcoli il rapporto di guadagno dei due attributi rispetto a questi esempi di training

- c) si costruisca un albero decisionale ad un solo livello per il training set dato, indicando le etichette delle foglie (numero di esempi finiti nella foglia/numero di esempi finiti nella foglia non appartenenti alla classe della foglia)

Tre ?

•ESERCIZIO ALBERI DECISIONALI

a)8 #pos, 9 #neg, 17 #tot

- info(S)=-8/17*$\log_2$ 8/17-9/17*$\log_2$ 9/17=0.998
-
- b) Per calcolare il guadagno dell'attributo Distanza non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno Distanza noto (insieme F):

- info(F)=-8/16*$\log_2$ 8/16– 8/16*$\log_2$ 8/16=1
- $info_{Distanza}$(F)=5/16*(-4/5*$\log_2$4/5-1/5*$\log_2$

•ESERCIZIO ALBERI DECISIONALI

a) info(S)=-8/16*$\log_2$ 8/16-8/16*$\log_2$ 8/16=1

- b) infoHelical(S)=5/16*(-3/5*$\log_2$3/5-2/5*$\log_2$ 2/5)+5/16*(-3/5*$\log_2$3/5-2/5*$\log_2$ 2/5) +6/16*(-2/6*$\log_2$2/6-4/6*$\log_2$ 4/6) =

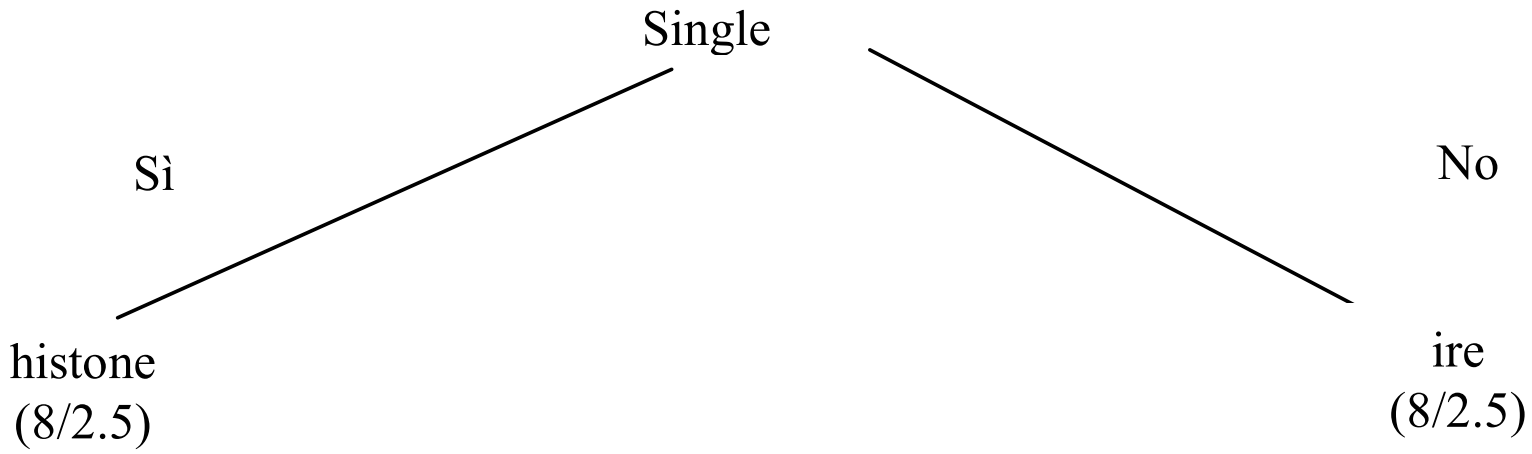
a) =0.312*0.971+0.312*0.971+0.375*0.918=0.950

b) gain(Helical) =1-0.950=0.050

c) Per calcolare il guadagno dell'attributo

•ESERCIZIO ALBERI DECISIONALI

- c) L'attributo scelto per la radice dell'albero è Single (maggiore gain).



rispettivamente

•8/16=0.5 e 8/16=0.5.

ESERCIZIO ALBERI DECISIONALI

•La prima parte viene mandata lungo il ramo Sì e viene classificata come histone con probabilità 5.5/8=68.7% e come

•ire con probabilità 1-68.7%=31.3%.

•La seconda parte viene mandata lungo il ramo No e viene classificata come

•ire con probabilità 5.5/8 =68.7% e

•histone con probabilità 1-68.7%=31.3%.

•Quindi in totale la classificazione dell'istanza è

•histone: 0.5*68.7%+0.5*31.3%=50%