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 Rilevante

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).

d) si classifichi l'istanza:

?	Si

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)=$5/16*(-4/5*\log_2 4/5-1/5*\log_2 1/5)+5/16(-3/5*\log_2 3/5-2/5*\log_2$ 2/5) +$6/16(-1/6*\log_2 1/6-5/6*\log_2 5/6)=$ // tre entropie calcolate sui tre sottoinsiemi con diverso valore dell'attr. Distanza
=0.312*0.722+0.312*0.971+0.375 *0.650=0.772

gain(Distanza) =16/17*(1-0.772) =0.215

ESERCIZIO ALBERI DECISIONALI

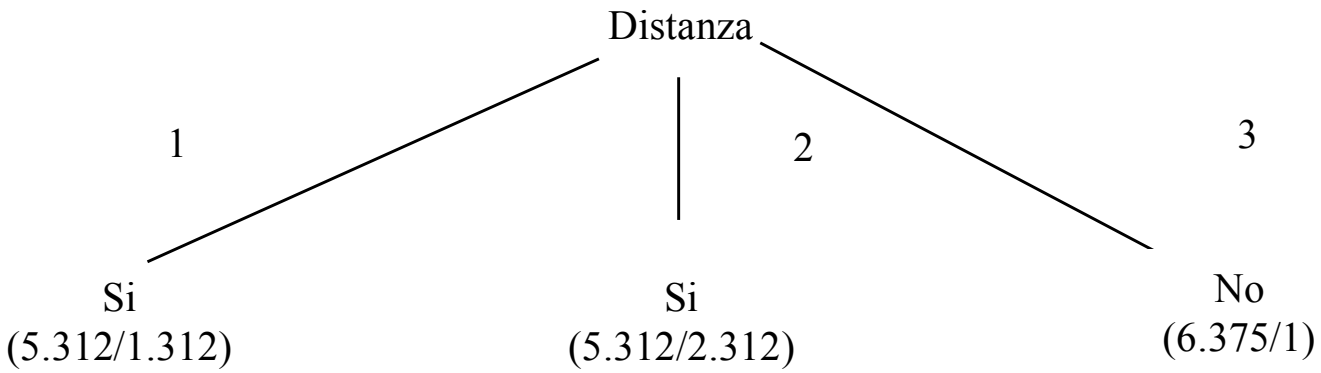
Per calcolare il guadagno dell'attributo Uguali non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno Uguali noto (insieme F):

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

gain(Uguali) =15/17*(0.997-0.993)=0.004

ESERCIZIO ALBERI DECISIONALI

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



ESERCIZIO ALBERI DECISIONALI

d) l'istanza viene divisa in tre parti, di peso rispettivamente

5.312/17=0.312,
5.312/17=0.312
6.375/17=0.375.

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
Si con probabilità 1-84.3%=15.7%.

Quindi in totale la classificazione dell'istanza è
Si: 0.312*75.3%+0.312*56.5%+0.375*15.7%=47.0%
No: 0.312*24.7%+0.312*43.5%+0.375*84.3% =52.9%

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).

d) si classifichi l'istanza:

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$ 1/5)+5/16(-3/5*$\log_2$ 3/5-2/5*$\log_2$ 2/5) +6/16(-1/6*$\log_2$ 1/6-5/6*$\log_2$ 5/6)= // tre entropie calcolate sui tre sottoinsiemi con diverso valore dell'attr. Distanza

=0.312*0.722+0.312*0.971+0.375 *0.650=0.772

gain(Distanza) =16/17*(1-0.772) =0.215

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) =

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

gain(Helical) =1-0.950=0.050

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

info(F)=-7/14*log2 7/14– 7/14*log2 7/14=1

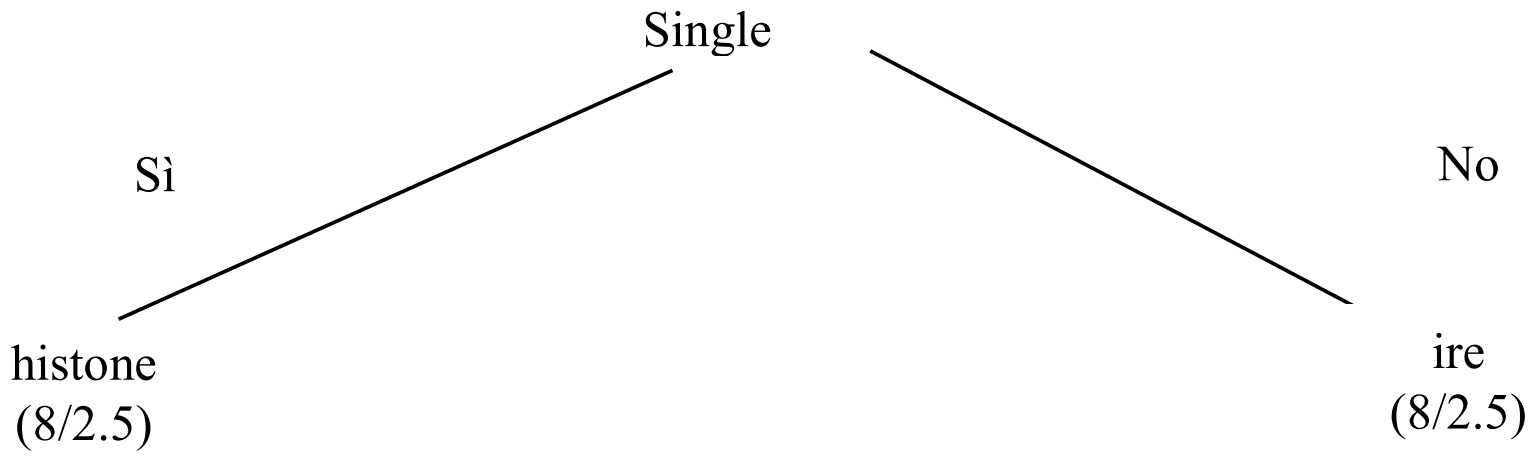
infoSingle(F)=7/14*(-5/7*log25/7-2/7*log2 2/7)+7/14(-2/7*log2 2/7-5/7*log2 5/7)=

=0.5*0.863+0.5*0.863=0.863

gain(Single) =14/16*(1-0.863)=0.120

ESERCIZIO ALBERI DECISIONALI

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



ESERCIZIO ALBERI DECISIONALI

d) l'istanza viene divisa in due parti, di peso rispettivamente 8/16=0.5 e 8/16=0.5.
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%
ire: 0.5*31.3%+0.5*68.7%=50%