Decision Trees Exercises

Master degree in Computer Science
Dept. Demacs – Unical
Prof. P. Rullo
rullo@unical.it
AY 2021-2022

Exercise 1

- Determine the best splitting attribute, based on IG, among Humidity and Wind

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Exercise 1 - solution

- S = [9+,5-], |S|=14, p_+ = 9/14, p_-=5/15;
- Entropy of S (before splitting)
  - E$(S) = -p_+ log_2 p_+ - p_- log_2 p_- =$ 0.94

- A = wind, Values(A) = {weak, strong}
  - S_w = [6+,2-], |S_w|=8, p+=3/4, p-=1/4
  - E(S_w) = 0.811
  - S_s = [3+,3-], p+=1/2, p-=1/2, |S_s|=6
  - E(S_s) = 1
  - $IG(S,A) = E(S) - \frac{|Sw|}{|S|}E(Sw) - \frac{|Ss|}{|S|}E(Ss)$
  - $IG(S,A) =$ 0.048

A = humidity, Values(A) = {high, normal}

- S_h = [3+,4-], |S_h|=7, p+=3/7, p-=4/7
- E(S_{high}) = 0.985
- S_n = [6+,1-], |S_n|=7, p+=6/7, p-=1/7
- E(S_n) = 0.592
- $IG(S,A) = E(S) - \frac{|Sh|}{|S|}E(S_h) - \frac{|Sn|}{|S|}E(S_n)$
- $IG(S,A) =$ 0.15

- Humidity is the best attribute

Exercise 2

- Build a DT over the dataset S shown below

Id	A1	A2	A3	A4	Class
1	A	D	Si	F	+
2	C	D	Si	M	+
3	A	E	No	F	+
4	C	E	Si	M	+
5	C	E	No	M	-
6	C	E	No	F	-

S = [4+, 2-], |S|=6, p_+ = 2/3, p_-=1/3;
Entropy of S (before splitting)

$$E(S) = -p_+ log_2 p_+ - p_- log_2 p_- = 0.92$$

Exercise 2 - solution

SELECT THE ROOT

- Att = A1, Values(A1) = {A, C}
  - S_A = [2+,0-], S_C = [2+, 2-]
  - $E(S_A) = 0$; $E(S_C) = 1$
  - IG(S,A1) = 0.251
- Att = A2, Values(A2) = {D, E}
  - S_D = [2+,0-], S_E = [2+,2-]
  - $E(S_D) = 0$; $E(S_E) = 1$
  - IG(S,A2) = 0.251
- Att = A3, Values(A3) = {Si, No}
  - S_{Si} = [3+,0-], S_{No} = [1+,2-]
  - $E(S_{Si}) = 0$; $E(S_{No}) = 0.918$
  - IG(S,A3) = 0.459

- Att = A4, Values(A4) = {F, M}
  - S_F = [2+,1-], S_M = [2+, 1-]
  - $E(S_F) = 0.918$; $E(S_M) = 0.918$
  - IG(S,A4) = 0
- The best attribute is A3, which is then the root of the tree

Exercise 2 - solution

BUILD SUBTREES OF THE ROOT

- Root = A3, Values(A3) = {Si, No}
  - S_{Si} = [3+,0-], S_{No} = [1+,2-]
- Left Subtree LS(A3=Si): the elements of $S_{A3=Si}$ all belong to the positive class
  - the root of LS(A3=Si) is a leaf node with label +
- Right Subtree RS(A3=No) is to be created over the dataset $S_{A3=No}$

Dataset $S_{A3=Si}$

Id	A1	A2	A3	A4	Class
1	A	D	Si	F	+
2	C	D	Si	M	+
4	C	E	Si	M	+

Dataset $S_{A3=No}$

Id	A1	A2	A3	A4	Class
3	A	E	No	F	+
5	C	E	No	M	-
6	C	E	No	F	-

Exercise 2 - solution

BUILD SUBTREE RS(A3=No)

- Entropy of T (before splitting)
  - $T = S_{A3=No} = [1+,2-]$, $E(T) = 0.92$
- Att = A1, Values(A1) = {A, C}
  - $T_A = [1+,0-]$, $T_C = [0+, 2-]$
  - $E(T_A) = 0$; $E(T_C) = 0$
  - IG(T,A1) = 0.92
- Att = A4, Values(A4) = {F, M}
  - $T_F = [1+,1-]$, $T_M = [0+,1-]$
  - $E(T_F) = 1$; $E(T_{No}) = 0$
  - IG(T,A4) = 0.258

Dataset $T = S_{A3=No}$

Id	A1	A2	A3	A4	Class
3	A	E	No	F	+
5	C	E	No	M	-
6	C	E	No	F	-

A1 is the best splitting attribute which becomes the root of the subtree RS(A_3=No)

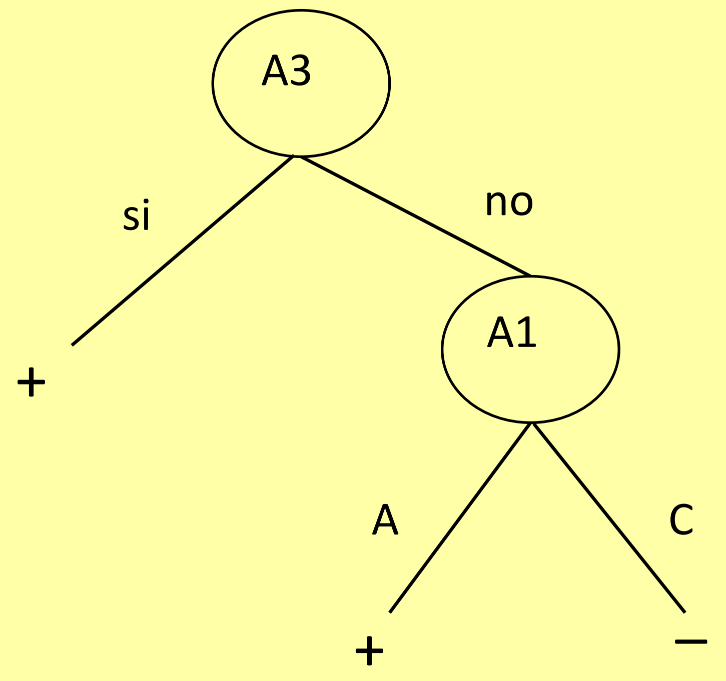
Exercise 2 - solution

BUILD SUBTREES OF A1

A1 splits the dataset T into two homogeneous subsets

- T_A = [1+,0-]
- T_C = [0+, 2-]

which then become leaf nodes



Exercise 3

- What are the best splits of S according to IG and the Gini Index?

Data set S

Instance	a1	a2	Class
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-
7	F	F	-
8	T	F	+
9	F	T	-

- $IG(A,S) = E(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} E(S_v)$
- $G(S) = 1 - \sum p(c)^2$
- $GI(A,S) = G(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} G(S_v)$

Exercise 3 - solution

Instance	A1	A2	Class
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-
7	F	F	-
8	T	F	+
9	F	T	-

INFORMATION GAIN

- S = [4+,5-], E(S) = 0.99
- $S_{A1=T}$ = [3+,1-], $S_{A1=F}$ = [1+,4-]
- E($S_{A1=T}$) = 0,81; E($S_{A1=F}$) = 0.72
- IG(S,A1) = 0.99-4/9*0.81-5/9*0.72 = 0.23
- $S_{A2=T}$ = [2+,3-]; $S_{A2=F}$ = [2+,2-]
- E($S_{A1=T}$) = 0,97; E($S_{A2=F}$) = 1
- IG(S,A2) = 0.99-5/9*0.97-4/9*1 = 0.01

Based on IG, the attribute A1 has the greatest discriminating power

Exercise 3 - solution

Instance	A1	A2	Class
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-
7	F	F	-
8	T	F	+
9	F	T	-

GINI INDEX

- S = [4+,5-], G(S) = 0.49
- $S_{A1=T}$ = [3+,1-], $S_{A1=F}$ = [1+,4-]
- $G(S_{A1=T}) = 1 - (3/4)^2 - (1/4)^2 = 0,38$;
- $G(S_{A1=F}) = 1 - (1/5)^2 – (4/5)^2 = 0.32$
- GI(S,A1) = 0.49-4/9*0.38-5/9*0.32 = 0.15

- $S_{A2=T}$ = [2+,3-]; $S_{A2=F}$ = [2+,2-]
- $G(S_{A2=T}) = 1-(2/5)^2 – (3/5)^2 = 0,48$;
- $G(S_{A2=F}) = 1- 0,25 – 0,25 = 0.50$
- GI(S,A2) = 0.49-5/9*0.48-4/9*0.50 = 0.001

Based on Gini Index, the attribute A1 has the greatest discriminating power