Esercizio 1

Si consideri il seguente insieme di istanze di training

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

$$\text{Gini} = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

$I(\text{children}) = \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$ → impurità collettiva

$$\Delta_{gain} = I(parent) - I(children)$$

(a) Calcolare l'indice GINI per l'insieme complessivo di tutte le istanze

Gini(all instances) = $1 - 0.5^2 - 0.5^2 = 0.5$

Le istanze sono equamente divise tra le due classi C0 e C1, pertanto la frequenza relativa di ciascuna classe è 0.5.

(b) Calcolare l'indice GINI per l'attributo Customer ID

L'attributo "customer ID" è diverso per ciascuna istanza di training; pertanto, per un qualsiasi valore *x* id "customer ID" l'indice Gini ha un valore pari a 0 → il nodo è puro perché contiene una sola istanza che appartiene ad una sola classe.

Gini(customer ID = 1) = $1 - (1/1)^2 - (0/1)^2 = 0$

→ per customer ID = 1 ho una sola istanza che ha classe 0 e nessuna istanza che ha classe 1. Lo stesso vale per tutti gli altri possibili valori di "customer ID".

Gini(customer ID) = 0

→ sommando 0 per tutti i valori di "customer ID" si ottiene complessivamente 0

(c) Calcolare l'indice GINI per l'attributo Gender

Gini(Gender = M) = $1 - (6/10)^2 - (4/10)^2 = 0.48$

→ Ci sono 10 istanze della classe C0, di cui 6 hanno "Gender = M", quindi la frequenza relativa $p_i(t)$ per la classe C0 è 6/10
→ Ci sono 10 istanze della classe C1, di cui 4 hanno "Gender = M", quindi la frequenza relativa $p_i(t)$ per la classe C1 è 6/10

Gini(Gender = F) = $1 - (4/10)^2 - (6/10)^2 = 0.48$

Gini(Gender) = (10/20) * Gini(Gender = M) + (10/20) * Gini(Gender = F)
= (10/20) * 0.48 + (10/20) * 0.48 = **0.48**

→ Per calcolare l'indice Gini complessivo per l'attributo Gender, si esegue la media pesata degli indici Gini relativi ai vari valori di Gender. Il peso è dato dal numero di istanze che hanno quel particolare valore di attributo, rispetto al totale. Nel caso specifico ci sono 10 istanze su 20 con Gender = M, e 10 istanze su 20 con Gender = F.

(d) Calcolare l'indice GINI per l'attributo Car Type

Gini(Car Type = 'Family') = $1 - (1/4)^2 - (3/4)^2 = 0.375$

→ ci sono 4 istanze con "car type = family", di cui 1 con classe = C0 e 3 con classe C1

Gini(Car Type = 'Sports') = $1 - (8/8)^2 - (0/8)^2 = 0$

Gini(Car Type = 'Luxury') = $1 - (1/8)^2 - (7/8)^2 = 0.2188$

Gini(Car Type) = (4/20) * Gini(Car Type = 'Family') +
(8/20) * Gini(Car Type = 'Sports') +
(8/20) * Gini(Car Type = 'Luxury')
= (4/20) * 0.375 + (8/20) * 0 + (8/20) * 0.2188 = **0.1625**

(e) Calcolare l'indice Gini per l'attributo Shirt Size

Gini(Shirt size = 'S') = $1 - (3/5)^2 - (2/5)^2 = 0.48$

Gini(Shirt size = 'M') = $1 - (3/7)^2 - (4/7)^2 = 0.4898$

Gini(Shirt size = 'L') = $1 - (2/4)^2 - (2/4)^2 = 0.5$

Gini(Shirt size = 'XL') = $1 - (2/4)^2 - (2/4)^2 = 0.5$

Gini(Shirt size) = (5/20) * Gini(Shirt size = 'S') + (7/20) * Gini(Shirt size = 'M')
+ (4/20) * Gini(Shirt size = 'L') + (4/20) * Gini(Shirt size = 'XL')
= (5/20) * 0.48 + (7/20) * 0.4898 + (4/20) * 0.5 + (4/20) * 0.5 = **0.4919**

(f) Quale attributo tra Gender, Car Type e Shirt Size è migliore come attributo di split di un albero decisionale?

Il migliore attributo tra i tre è "Car type" perché presenta l'indice Gini più basso. Gli altri due hanno un valore Gini molto vicino a 0.50 che rappresenta il massimo grado di impurità.

Esercizio 2

Si consideri il seguente insieme di istanze di training

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Entropy = $-\sum_{i=0}^{c-1} p_i(t) log_2\, p_i(t)$

I(children) = $\sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$ → impurità collettiva

$\Delta_{gain} = I(parent) - I(children)$

(a) Calcolare l'entropia per l'insieme complessivo di tutte le istanze

E(all instances) = (4/9)$\log_2$(4/9) + (5/9)$\log_2$(5/9) = 0.9911

→ L'entropia è molto vicina a 1 (max grado di impurità), infatti le istanze sono quasi equamente divise tra le due classi.

(b) Qual è il guadagno di informazione (information gain) dell'attributo a_1 e dell'attributo a_2?

Consideriamo l'attributo a_1

a_1	+	-
T	3	1
F	1	4

E(a_1 = T) = - (3/4) $\log_2$(3/4) - (1/4) $\log_2$(1/4) = 0.811
E(a_1 = F) = - (1/5) $\log_2$(1/5) - (4/5) $\log_2$(4/5) = 0.7219
E(a_1) = (4/9) * E(a_1 = T) + (5/9) * E(a_1 = F) = (4/9) * 0.811 + (5/9) * 0.7219 = 0.7615

$\Delta_{gain}(a_1)$ = 0.9911 – 0.7615 = **0.2296**

Consideriamo l'attributo a_2

a_2	+	-
T	2	3
F	2	2

$E(a_2 = T) = - (2/5) \log_2(2/5) - (3/5) \log_2(3/5) = 0.9709$
$E(a_2 = F) = - (2/4) \log_2(2/4) - (2/4) \log_2(2/4) = 1$
$E(a_2) = (5/9) * E(a_2 = T) + (4/9) * E(a_2 = F) = (5/9) * 0.9709 + (4/9) * 1 = 0.9839$
$\Delta_{gain}(a_2)$ = 0.9911 – 0.9839 = **0.0072**

L'attributo a_1 risulta un candidato migliore rispetto ad a_2 per la definizione della condizione di split.

(c) Calcolare l'information gain per l'attributo a_3 considerando tutte le possibili posizioni di split.

Split pos.	2.0		3.5		4.5		5.5		6.5		7.5	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
+	1	3	1	3	2	2	2	2	3	1	4	0
-	0	5	1	4	1	4	3	2	3	2	4	1
Entropy	0.8484 (*)		0.9885		0.9183		0.9839		0.9728		0.889	
Δgain	0.1427 (**)		0.0026		0.0728		0.0072		0.0183		0.1022	

(*) $E(a_3 <= 2.0) = -1 \log_2(1) – 0 \log_2(0) = 0$

$E(a_3 > 2.0) = -(3/8) \log_2(3/8) – (5/8) \log_2(5/8) = 0.9544$

$E(a_3) = (1/9) * 0 + (8/9) * 0.9544 = 0.8484$

(**) $\Delta gain(a_3$ split 2.0) = 0.9911 – 0.8484 = 0.1427

Il miglior valore di split è 2.0 perché presenta il valore di information gain più alto.

(d) Qual è il miglior attributo tra a_1, a_2 e a_3 per la condizione di split?

Il maggior guadagno di informazione si ottiene considerando l'attributo a_1, poiché in questo caso l'information gain è di 0.2296, mentre con a_2 si ottiene un information gain di 0.0072 e con a_3 0.1427.

Esercizio 3

Si consideri il seguente insieme di istanze di training

X	Y	Z	No. of Class C1 Examples	No. of Class C2 Examples
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

Classification error rate → error = $1 - \max_i(p_i(t))$

I(children) = $\sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$ → impurità collettiva

$$\Delta_{gain} = I(parent) - I(children)$$

(a) Si costruisca un albero decisionale **di altezza 2** usando l'approccio greedy visto a lezione. Si utilizzi la metrica "classification error rate" per la scelta dell'attributo di split. Qual è l'errore complessivo di classificazione ottenuto con questo albero?

Passo 0) Si crea un unico nodo a cui sono associate tutte le 200 istanze. Il nodo non è puro perché ci sono 100 istanze con classe C1 e 100 istanze con classe C2. Pertanto, si procede ricorsivamente, individuando la migliore condizione di split.

Passo 1) Determinare il miglior attributo di split per il livello 1: è necessario calcolare l'error rate per ciascuno degli attributi X, Y e Z.

Attributo X

X	C1	C2
0	60	60
1	40	40

Err(X = 0) = 1 – max(60/120, 60/120) = 1 – 0.5 = 0.5
Err(X = 1) = 1 – max(40/80, 40/80) = 1 – 0.5 = 0.5

Err(X) = 120/200 * 0.5 + 80/200 * 0.5 = 0.3 + 0.2 = **0.5**

Attributo Y

Y	C1	C2
0	40	60
1	60	40

Err(Y = 0) = 1 – max(40/100, 60/100) = 1 – 0.6 = 0.4
Err(Y = 1) = 1 – max(60/100, 40/100) = 1 – 0.6 = 0.4

Err(Y) = 100/200 * 0.4 + 100/200 * 0.4 = 0.2 + 0.2 = **0.4**

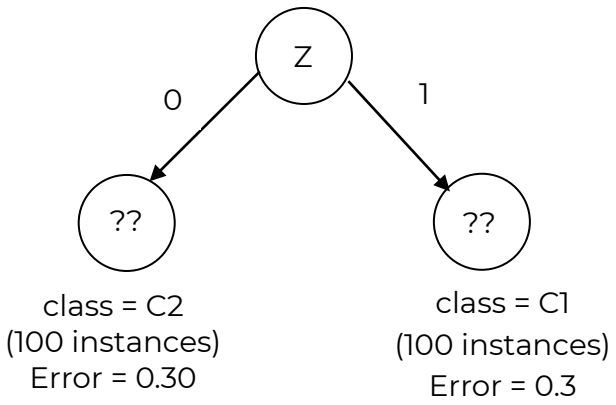
Attributo Z

Z	C1	C2
0	30	70
1	70	30

Err(Z = 0) = 1 – max(30/100, 70/100) = 1 – 0.7 = 0.3
Err(Z = 1) = 1 – max(70/100, 30/100) = 1 – 0.7 = 0.3

Err(Z) = 100/200 * 0.3 + 100/200 * 0.3 = 0.15 + 0.15 = **0.3**

Z presenta il valore di error rate più basso, quindi viene scelto come attributo per eseguire lo split al livello 1.



Entrambi i nodi sono impuri, quindi si procede con il secondo (e ultimo) livello richiesto.

Passo 1) Per entrambi nodi figli si determina qual è il migliore attributo di split

Nodo Z = 0

Attributo X

X	C1	C2
0	15	45
1	15	25

Err(X = 0) = 1 – max(15/60, 45/60) = 1 – 0.75 = 0.25
Err(X = 1) = 1 – max(15/40, 25/40) = 1 – 0.625 = 0.375

Err(X) = 60/100 * 0.25 + 40/100 * 0.375 = 0.15 + 0.15 = **0.3**

Attributo Y

Y	C1	C2
0	15	45
1	15	25

Stessa condizione di X

Err(Y) = 60/100 * 0.25 + 40/100 * 0.375 = 0.15 + 0.15 = **0.3**

Nodo Z = 1

Attributo X

X	C1	C2
0	45	15
1	25	15

Err(X = 0) = 1 – max(45/60, 15/60) = 1 – 0.75 = 0.25
Err(X = 1) = 1 – max(25/40, 15/40) = 1 – 0.625 = 0.375

Err(X) = 60/100 * 0.25 + 40/100 * 0.375 = 0.15 + 0.15 = **0.3**

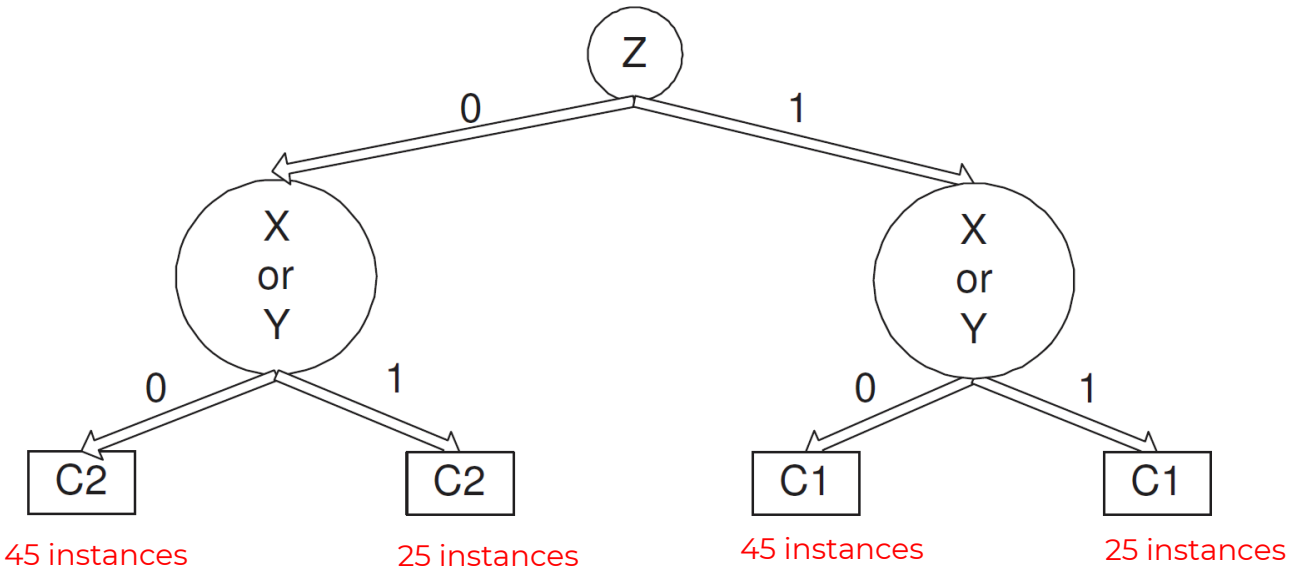
Attributo Y

Y	C1	C2
0	25	15
1	45	15

Stessa condizione di X

Err(Y) = 60/100 * 0.25 + 40/100 * 0.375 = 0.15 + 0.15 = **0.3**

I due attributi X e Y presentano gli stessi valori di classification error rate, pertanto si può scegliere indistintamente l'uno o l'altro come attributo per eseguire lo split al secondo livello in entrambi i nodi figli).



Errore rate = (# of wrong preditions) / (total # of predictions)

= (15 + 15 + 15 + 15)/200 = 0.3

Ognuno dei nodi foglia sbaglia a classificare 15 istanze sul totale di 200.

(b) Si ripeta la costruzione dell'albero assumendo però di usare al primo livello X come attributo di split X, e quindi determinare tramite la metrica di classification error rate il successivo attributo. Come varia l'errore complessivo?

Dopo aver eseguito lo split usando il nodo X, la condizione successiva può considerare l'attributo Y oppure Z.

Nodo X = 0

Attributo Y

Y	C1	C2
0	5	55
1	55	5

Err(Y = 0) = 1 – max(5/60, 55/60) = 5/60
Err(Y = 1) = 1 – max(55/60, 5/60) = 5/60

Err(Y) = 60/120 * 5/60 + 60/120 * 5/60 = 5/60

Attributo Z

Z	C1	C2
0	15	45
1	45	15

Err(Z = 0) = 1 – max(15/60, 45/60) = 15/60
Err(Z = 1) = 1 – max(45/60, 15/60) = 15/60

Err(Z) = 60/120 * 15/60 + 60/120 * 15/60 = 15/60

L'error rate di Y è minore di quello di Z, quindi Y viene scelto come attributo per lo split di secondo livello per il Nodo X = 0.

Nodo X = 1

Attributo Y

Y	C1	C2
0	35	5
1	5	35

Err(Y = 0) = 1 – max(35/40, 5/40) = 5/40
Err(Y = 1) = 1 – max(5/40, 35/40) = 5/40

Err(Y) = 40/80 * 5/40 + 40/80 * 5/40 = 5/40

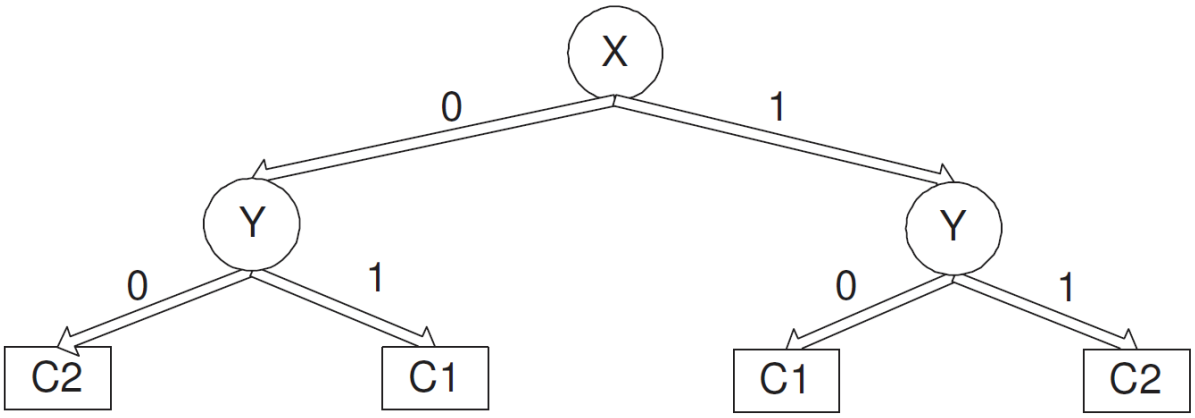
Attributo Z

Z	C1	C2
0	15	25
1	25	15

Err(Z = 0) = 1 – max(15/40, 25/40) = 15/40
Err(Z = 1) = 1 – max(25/40, 15/40) = 15/40

Err(Y) = 40/80 * 15/40 + 40/80 * 15/40 = 15/40

L'error rate di Y è minore di quello di Z, quindi Y viene scelto come attributo per lo split di secondo livello per il Nodo X = 1.



Errore rate = (# of wrong preditions) / (total # of predictions)
= (5 + 5 + 5 + 5)/200 = 0.1

Ognuno dei nodi foglia sbaglia a classificare 5 istanze sul totale di 200.

Esercizio 4

Si consideri il seguente insieme di istanze di training

A	B	C	Number of Instances	
			+	−
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

Classification error rate → error = $1 - \max_i(p_i(t))$

I(children) = $\sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$ → impurità collettiva

$$\Delta_{gain} = I(parent) - I(children)$$

(a) Costruire un albero decisionale di 2 livelli considerando come metrica di valutazione degli attributi di split il classification error rate.

Errore complessivo prima di eseguire qualsiasi split

E_{orig} = 1 – max (50/100, 50/100) = 50/100

Livello 1

Split sull'attributo A

A	+	-
T	25	0
F	25	50

Err(A = T) = 1 – max(25/25, 0/25) = 0
Err(A = F) = 1 – max(25/75, 50/75) = 25/75

Err(A) = 25/100 * 0 + 75/100 * 25/75 = 25/100

Δ_{gain}**(A)** = Eorig – Err(A) = 50/100 – 25/100 = **25/100**

Split sull'attributo B

B	+	-
T	30	20
F	20	30

Err(B = T) = 1 – max(30/50, 20/50) = 20/50
Err(B = F) = 1 – max(20/50, 30/50) = 20/50
Err(B) = 50/100 * 20/50 + 50/100 * 20/50 = 40/100

Δ_{gain}**(B)** = Eorig – Err(B) = 50/100 – 45/100 = **10/100**

Split sull'attributo C

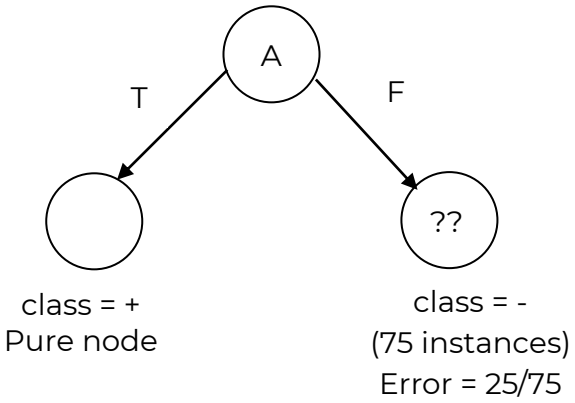
C	+	-
T	25	25
F	25	25

Err(C = T) = 1 – max(25/50, 25/50) = 25/50
Err(C = F) = 1 – max(25/50, 25/50) = 25/50
Err(C) = 50/100 * 25/50 + 50/100 * 25/50 = 50/100

Δ_{gain}**(C)** = Eorig – Err(B) = 50/100 – 50/100 = **0**

L'attributo A viene scelto per lo split di primo livello perché presenta il valore più grande di information gain.

Livello 2



Il nodo A=T è puro quindi non serve procedere ulteriormente, al contrario si procede suddividendo il nodo A=F che presenta le seguenti distribuzioni delle istanze:

B	C	Class label	
		+	−
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

Eorig(A=F) = 25/75

Split sull'attributo B

B	+	-
T	25	20
F	0	30

Err(B = T) = 1 – max(25/45, 20/25) = 20/45
Err(B = F) = 1 – max(0/30, 30/30) = 0

Err(B) = 45/75 * 20/45 + 30/75 * 0 = 20/75

Δ_{gain}**(B)** = Eorig – Err(B) = 25/75 – 20/75 = **5/75**

Split sull'attributo C

C	+	-
T	0	25
F	25	25

Err(C = T) = 1 – max(0/25, 25/25) = 0
Err(C = F) = 1 – max(25/50, 25/50) = 25/50

Err(C) = 25/75 * 0 + 50/75 * 25/50 = 25/75

Δ_{gain}**(C)** = Eorig – Err(C) = 25/75 – 25/75 = **0**

L'attributo B è scelto per lo split di secondo livello.

→ complessivamente rimangono 20 istanze con non sono classificate correttamente → **error rate = 20/100**