



UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA

Laurea triennale in Informatica



Viviana Pentangelo

✉ tutoratofia@gmail.com

Fondamenti di Intelligenza Artificiale

Help Teaching - Esercitazione 3



FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Costruire l'albero di decisione sulla base del dataset sottostante. Siamo interessati a predire la colonna **Frutta**.

Colore	Altezza (mm)	Larghezza (mm)	Frutta
Verde	57	62	Mela
Giallo	40	180	Banana
Verde	69	72	Mela
Arancione	35	30	Arancia
Arancione	45	35	Arancia
Arancione	50	45	Arancia
Giallo	40	170	Banana
Giallo	30	140	Banana
Giallo	60	62	Mela
Giallo	52	58	Mela

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$


Quante volte nel dataset D compare la classe c

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

Colore	Altezza (mm)	Larghezza (mm)	Frutta
Verde	57	62	Mela
Giallo	40	180	Banana
Verde	69	72	Mela
Arancione	35	30	Arancia
Arancione	45	35	Arancia
Arancione	50	45	Arancia
Giallo	40	170	Banana
Giallo	30	140	Banana
Giallo	60	62	Mela
Giallo	52	58	Mela

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

Colore	Altezza (mm)	Larghezza (mm)	Frutta	
Verde	57	62	Mela	Mela = 4/10
Giallo	40	180	Banana	
Verde	69	72	Mela	
Arancione	35	30	Arancia	
Arancione	45	35	Arancia	
Arancione	50	45	Arancia	
Giallo	40	170	Banana	
Giallo	30	140	Banana	
Giallo	60	62	Mela	
Giallo	52	58	Mela	

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

Colore	Altezza (mm)	Larghezza (mm)	Frutta	
Verde	57	62	Mela	Mela = 4/10
Giallo	40	180	Banana	Banana = 3/10
Verde	69	72	Mela	
Arancione	35	30	Arancia	
Arancione	45	35	Arancia	
Arancione	50	45	Arancia	
Giallo	40	170	Banana	
Giallo	30	140	Banana	
Giallo	60	62	Mela	
Giallo	52	58	Mela	

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

Colore	Altezza (mm)	Larghezza (mm)	Frutta	
Verde	57	62	Mela	Mela = 4/10
Giallo	40	180	Banana	
Verde	69	72	Mela	Banana = 3/10
Arancione	35	30	Arancia	Arancia = 3/10
Arancione	45	35	Arancia	
Arancione	50	45	Arancia	
Giallo	40	170	Banana	
Giallo	30	140	Banana	
Giallo	60	62	Mela	
Giallo	52	58	Mela	

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

$$H(D) = - \frac{4}{10} \cdot \log_2 \frac{4}{10} - \frac{3}{10} \cdot \log_2 \frac{3}{10} - \frac{3}{10} \cdot \log_2 \frac{3}{10}$$

Mela = 4/10

Banana = 3/10

Arancia = 3/10

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

$$H(D) = - \frac{4}{10} \cdot \log_2 \frac{4}{10} - \frac{3}{10} \cdot \log_2 \frac{3}{10} - \frac{3}{10} \cdot \log_2 \frac{3}{10}$$

Mela = 4/10
Banana = 3/10
Arancia = 3/10

Alberi di decisione - Esercizio 1

Ci serviamo dell'**entropia** come misura per stabilire la **purezza** dei valori di un campione.

$$H(D) = - \sum p(c) \cdot \log_2 p(c)$$

$$H(D) = - \frac{4}{10} \cdot \log_2 \frac{4}{10} - \frac{3}{10} \cdot \log_2 \frac{3}{10} - \frac{3}{10} \cdot \log_2 \frac{3}{10}$$

Mela = 4/10

Banana = 3/10

$$H(D) = 0.53 + 0.52 + 0.52 = 1.57$$

Arancia = 3/10



Entropia iniziale del dataset

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Come primo step siamo interessati a capire quale delle feature può rappresentare la **miglior radice** per l'albero di decisione che vogliamo costruire

Colore	Altezza (mm)	Larghezza (mm)	Frutta
Verde	57	62	Mela
Giallo	40	180	Banana
Verde	69	72	Mela
Arancione	35	30	Arancia
Arancione	45	35	Arancia
Arancione	50	45	Arancia
Giallo	40	170	Banana
Giallo	30	140	Banana
Giallo	60	62	Mela
Giallo	52	58	Mela

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Come lo decidiamo?

Come primo step siamo interessati a capire quale delle feature può rappresentare la **miglior radice** per l'albero di decisione che vogliamo costruire



Colore	Altezza (mm)	Larghezza (mm)	Frutta
Verde	57	62	Mela
Giallo	40	180	Banana
Verde	69	72	Mela
Arancione	35	30	Arancia
Arancione	45	35	Arancia
Arancione	50	45	Arancia
Giallo	40	170	Banana
Giallo	30	140	Banana
Giallo	60	62	Mela
Giallo	52	58	Mela

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo l'**Information Gain** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

Colore	Altezza (mm)	Larghezza (mm)	Frutta
Verde	57	62	Mela
Giallo	40	180	Banana
Verde	69	72	Mela
Arancione	35	30	Arancia
Arancione	45	35	Arancia
Arancione	50	45	Arancia
Giallo	40	170	Banana
Giallo	30	140	Banana
Giallo	60	62	Mela
Giallo	52	58	Mela

Alberi di decisione - Esercizio 1

Calcoliamo l'***Information Gain*** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

L'***Information Gain*** rappresenta la differenza di entropia precedente e successiva ad una divisione in base ad un determinato attributo.

$$Gain(D, A) = H(D) - \sum_{v \in A} \frac{|D_v|}{|D|} \cdot H(D_v)$$

Alberi di decisione - Esercizio 1

Calcoliamo l'***Information Gain*** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

L'***Information Gain*** rappresenta la differenza di entropia precedente e successiva ad una divisione in base ad un determinato attributo.

$$Gain(D, A) = H(D) - \sum_{v \in A} \frac{|D_v|}{|D|} \cdot H(D_v)$$


Entropia del dataset

Alberi di decisione - Esercizio 1

Calcoliamo l'**Information Gain** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

L'**Information Gain** rappresenta la differenza di entropia precedente e successiva ad una divisione in base ad un determinato attributo.

$$Gain(D, A) = H(D) - \sum_{v \in A} \frac{|D_v|}{|D|} \cdot H(D_v)$$



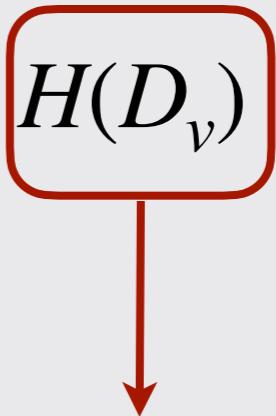
Quante volte un valore
di A compare sul totale

Alberi di decisione - Esercizio 1

Calcoliamo l'**Information Gain** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

L'**Information Gain** rappresenta la differenza di entropia precedente e successiva ad una divisione in base ad un determinato attributo.

$$Gain(D, A) = H(D) - \sum_{v \in A} \frac{|D_v|}{|D|} \cdot H(D_v)$$


Entropia di quel valore di A

Alberi di decisione - Esercizio 1

Calcoliamo l'***Information Gain*** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

L'***Information Gain*** rappresenta la differenza di entropia precedente e successiva ad una divisione in base ad un determinato attributo.

$$Gain(D, A) = H(D) - \sum_{v \in A} \frac{|D_v|}{|D|} \cdot H(D_v)$$

Siamo interessati a scegliere il nodo che avrà Infomation Gain maggiore.

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo l'**Information Gain** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

Colore	Frutta
Verde	Mela
Giallo	Banana
Verde	Mela
Arancione	Arancia
Arancione	Arancia
Arancione	Arancia
Giallo	Banana
Giallo	Banana
Giallo	Mela
Giallo	Mela

Colore può assumere tre valori = {Verde, Giallo, Arancione}

$$H(D_{verde}) =$$

$$H(D_{giallo}) =$$

$$H(D_{arancione}) =$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo l'**Information Gain** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

Colore	Frutta
Verde	Mela
Giallo	Banana
Verde	Mela
Arancione	Arancia
Arancione	Arancia
Arancione	Arancia
Giallo	Banana
Giallo	Banana
Giallo	Mela
Giallo	Mela

Colore può assumere tre valori = {Verde, Giallo, Arancione}

$$H(D_{verde}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{giallo}) =$$

$$H(D_{arancione}) =$$



Ci sono **2** mele verdi

Alberi di decisione - Esercizio 1

Calcoliamo l'***Information Gain*** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

Colore	Frutta
Verde	Mela
Giallo	Banana
Verde	Mela
Arancione	Arancia
Arancione	Arancia
Arancione	Arancia
Giallo	Banana
Giallo	Banana
Giallo	Mela
Giallo	Mela

Colore può assumere tre valori = {Verde, Giallo, Arancione}

$$H(D_{verde}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{giallo}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$H(D_{arancione}) =$$



Ci sono **3** banane e **2** mele gialle

Alberi di decisione - Esercizio 1

Calcoliamo l'**Information Gain** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

Colore	Frutta
Verde	Mela
Giallo	Banana
Verde	Mela
Arancione	Arancia
Arancione	Arancia
Arancione	Arancia
Giallo	Banana
Giallo	Banana
Giallo	Mela
Giallo	Mela

Colore può assumere tre valori = {Verde, Giallo, Arancione}

$$H(D_{verde}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{giallo}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$H(D_{arancione}) = -\frac{3}{3} \log_2 \frac{3}{3} = 0$$



Ci sono **3** arance arancioni

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo l'**Information Gain** per ciascuno degli attributi candidati.

Partiamo da **Colore**.

Colore	Frutta
Verde	Mela
Giallo	Banana
Verde	Mela
Arancione	Arancia
Arancione	Arancia
Arancione	Arancia
Giallo	Banana
Giallo	Banana
Giallo	Mela
Giallo	Mela

$$H(D_{verde}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{giallo}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$H(D_{arancione}) = -\frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$Gain(D, Colore) = 1.57 - \left(\frac{5}{10} \cdot 0.97 + \frac{2}{10} \cdot 0 + \frac{3}{10} \cdot 0 \right) = 1.09$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Probabilmente ora vi starete domandando...

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Probabilmente ora vi starete domandando...

"Ma ora devo fare quei calcoli per ogni singolo valore che c'è qui dentro?"



Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Probabilmente ora vi starete domandando...

"Ma ora devo fare quei calcoli per ogni singolo valore che c'è qui dentro?"



Sì.

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Probabilmente ora vi starete domandando...

"Ma ora devo fare quei calcoli per ogni singolo valore che c'è qui dentro?"



~~Si.~~

No, possiamo scegliere una strategia per categorizzare questi valori numerici!

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Scegliamo delle soglie per dividere in intervalli di valori le diverse altezze.

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Scegliamo delle soglie per dividere in intervalli di valori le diverse altezze.

"Come scegliamo queste soglie?"

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Scegliamo delle soglie per dividere in intervalli di valori le diverse altezze.

"Come scegliamo queste soglie?"

Ai fini dell'esercizio, scegliamo semplicemente dei valori utili guardando la distribuzione dei dati.

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
57	Mela
40	Banana
69	Mela
35	Arancia
45	Arancia
50	Arancia
40	Banana
30	Banana
60	Mela
52	Mela

Scegliamo delle soglie per dividere in intervalli di valori le diverse altezze.

"Come scegliamo queste soglie?"

Ai fini dell'esercizio, scegliamo semplicemente dei valori utili guardando la distribuzione dei dati.

Ad esempio, scegliamo $x \leq 40$, $40 < x \leq 50$, e $x > 50$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
>50	Mela
<=40	Banana
>50	Mela
<=40	Arancia
>40, <=50	Arancia
>40, <=50	Arancia
<=40	Banana
<=40	Banana
>50	Mela
>50	Mela

Scegliamo delle soglie per dividere in intervalli di valori le diverse altezze.

"Come scegliamo queste soglie?"

Ai fini dell'esercizio, scegliamo semplicemente dei valori utili guardando la distribuzione dei dati.

Ad esempio, scegliamo $x \leq 40$, $40 < x \leq 50$, e $x > 50$

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
>50	Mela
<=40	Banana
>50	Mela
<=40	Arancia
>40, <=50	Arancia
>40, <=50	Arancia
<=40	Banana
<=40	Banana
>50	Mela
>50	Mela

Scegliamo delle soglie per dividere in intervalli di valori le diverse altezze.

"Come scegliamo queste soglie?"

Ai fini dell'esercizio, scegliamo semplicemente dei valori utili guardando la distribuzione dei dati.

Ad esempio, scegliamo $x \leq 40$, $40 < x \leq 50$, e $x > 50$

Calcoliamo quindi l'entropia per i tre intervalli individuati.

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
>50	Mela
<=40	Banana
>50	Mela
<=40	Arancia
>40, <=50	Arancia
>40, <=50	Arancia
<=40	Banana
<=40	Banana
>50	Mela
>50	Mela

$$H(D_{\leq 40}) =$$

$$H(D_{>40, \leq 50}) =$$

$$H(D_{>50}) =$$

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
>50	Mela
<=40	Banana
>50	Mela
<=40	Arancia
>40, <=50	Arancia
>40, <=50	Arancia
<=40	Banana
<=40	Banana
>50	Mela
>50	Mela

$$H(D_{\leq 40}) = - \frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$$

$$H(D_{>40, \leq 50}) =$$

$$H(D_{>50}) =$$

Ci sono **3** banane e **1** arancia

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
>50	Mela
<=40	Banana
>50	Mela
<=40	Arancia
>40, <=50	Arancia
>40, <=50	Arancia
<=40	Banana
<=40	Banana
>50	Mela
>50	Mela

$$H(D_{\leq 40}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$$

$$H(D_{>40, \leq 50}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{>50}) =$$

Ci sono **2** arance

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
>50	Mela
<=40	Banana
>50	Mela
<=40	Arancia
>40, <=50	Arancia
>40, <=50	Arancia
<=40	Banana
<=40	Banana
>50	Mela
>50	Mela

$$H(D_{\leq 40}) = - \frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$$

$$H(D_{>40, \leq 50}) = - \frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{>50}) = - \frac{4}{4} \log_2 \frac{4}{4} = 0$$



Ci sono **4** mele

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Altezza**.

Altezza (mm)	Frutta
>50	Mela
<=40	Banana
>50	Mela
<=40	Arancia
>40, <=50	Arancia
>40, <=50	Arancia
<=40	Banana
<=40	Banana
>50	Mela
>50	Mela

$$H(D_{\leq 40}) = - \frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$$

$$H(D_{>40, \leq 50}) = - \frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{>50}) = - \frac{4}{4} \log_2 \frac{4}{4} = 0$$

$$Gain(D, Altezza) = 1.57 - \left(\frac{4}{10} \cdot 0.81 + \frac{2}{10} \cdot 0 + \frac{4}{10} \cdot 0 \right) = 1.24$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
62	Mela
180	Banana
72	Mela
30	Arancia
35	Arancia
45	Arancia
170	Banana
140	Banana
62	Mela
58	Mela

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
62	Mela
180	Banana
72	Mela
30	Arancia
35	Arancia
45	Arancia
170	Banana
140	Banana
62	Mela
58	Mela

Anche qui, scegliamo delle soglie per dividere in intervalli i valori numerici.

Scegliamo, ad esempio, $x \leq 45$, $45 < x \leq 100$, $x > 100$

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
>45, <=100	Mela
>100	Banana
>45, <=100	Mela
<=45	Arancia
<=45	Arancia
<=45	Arancia
>100	Banana
>100	Banana
>45, <=100	Mela
>45, <=100	Mela

Anche qui, scegliamo delle soglie per dividere in intervalli i valori numerici.

Scegliamo, ad esempio, $x \leq 45$, $45 < x \leq 100$, $x > 100$

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
>45, <=100	Mela
>100	Banana
>45, <=100	Mela
<=45	Arancia
<=45	Arancia
<=45	Arancia
>100	Banana
>100	Banana
>45, <=100	Mela
>45, <=100	Mela

Anche qui, scegliamo delle soglie per dividere in intervalli i valori numerici.

Scegliamo, ad esempio, $x \leq 45$, $45 < x \leq 100$, $x > 100$

Calcoliamo l'entropia per ciascuno dei tre intervalli.

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
>45, <=100	Mela
>100	Banana
>45, <=100	Mela
<=45	Arancia
<=45	Arancia
<=45	Arancia
>100	Banana
>100	Banana
>45, <=100	Mela
>45, <=100	Mela

$$H(D_{\leq 45}) =$$

$$H(D_{>45, \leq 100}) =$$

$$H(D_{>100}) =$$

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
>45, <=100	Mela
>100	Banana
>45, <=100	Mela
<=45	Arancia
<=45	Arancia
<=45	Arancia
>100	Banana
>100	Banana
>45, <=100	Mela
>45, <=100	Mela

$$H(D_{\leq 45}) = - \frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{>45, \leq 100}) =$$

$$H(D_{>100}) =$$

Ci sono **3** arance

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
>45, <=100	Mela
>100	Banana
>45, <=100	Mela
<=45	Arancia
<=45	Arancia
<=45	Arancia
>100	Banana
>100	Banana
>45, <=100	Mela
>45, <=100	Mela

$$H(D_{\leq 45}) = - \frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{>45, \leq 100}) = - \frac{4}{4} \log_2 \frac{4}{4} = 0$$

$$H(D_{>100}) =$$

Ci sono **4** mele

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
>45, <=100	Mela
>100	Banana
>45, <=100	Mela
<=45	Arancia
<=45	Arancia
<=45	Arancia
>100	Banana
>100	Banana
>45, <=100	Mela
>45, <=100	Mela

$$H(D_{\leq 45}) = - \frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{>45, \leq 100}) = - \frac{4}{4} \log_2 \frac{4}{4} = 0$$

$$H(D_{>100}) = - \frac{3}{3} \log_2 \frac{3}{3} = 0$$

Ci sono **3** banane

Alberi di decisione - Esercizio 1

Calcoliamo ora l'Information Gain dato da **Larghezza**.

Larghezza (mm)	Frutta
>45, <=100	Mela
>100	Banana
>45, <=100	Mela
<=45	Arancia
<=45	Arancia
<=45	Arancia
>100	Banana
>100	Banana
>45, <=100	Mela
>45, <=100	Mela

$$H(D_{\leq 45}) = - \frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{>45, \leq 100}) = - \frac{4}{4} \log_2 \frac{4}{4} = 0$$

$$H(D_{>100}) = - \frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$Gain(D, Largezza) = 1.57 - \left(\frac{3}{10} \cdot 0 + \frac{4}{10} \cdot 0 + \frac{3}{10} \cdot 0 \right) = 1.57$$

Alberi di decisione - Esercizio 1

A questo punto, possiamo finalmente scegliere il nodo radice dell'albero.

Scegliamo l'attributo A che ha dato $Gain(D,A)$ maggiore.

$$Gain(D, Colore) = 1.57 - \left(\frac{5}{10} \cdot 0.97 + \frac{2}{10} \cdot 0 + \frac{3}{10} \cdot 0 \right) = 1.09$$

$$Gain(D, Altezza) = 1.57 - \left(\frac{4}{10} \cdot 0.81 + \frac{2}{10} \cdot 0 + \frac{4}{10} \cdot 0 \right) = 1.24$$

$$Gain(D, Largezza) = 1.57 - \left(\frac{3}{10} \cdot 0 + \frac{4}{10} \cdot 0 + \frac{3}{10} \cdot 0 \right) = 1.57$$

Alberi di decisione - Esercizio 1

A questo punto, possiamo finalmente scegliere il nodo radice dell'albero.

Scegliamo l'attributo A che ha dato $Gain(D,A)$ maggiore.

$$Gain(D, Colore) = 1.57 - \left(\frac{5}{10} \cdot 0.97 + \frac{2}{10} \cdot 0 + \frac{3}{10} \cdot 0 \right) = 1.09$$

$$Gain(D, Altezza) = 1.57 - \left(\frac{4}{10} \cdot 0.81 + \frac{2}{10} \cdot 0 + \frac{4}{10} \cdot 0 \right) = 1.24$$

$$Gain(D, Largezza) = 1.57 - \left(\frac{3}{10} \cdot 0 + \frac{4}{10} \cdot 0 + \frac{3}{10} \cdot 0 \right) = 1.57$$

Scegliamo quindi l'attributo *Larghezza*.

In questo caso siamo fortunati, poiché considerare questo attributo ci porta già a non avere impurità.

Alberi di decisione - Esercizio 1

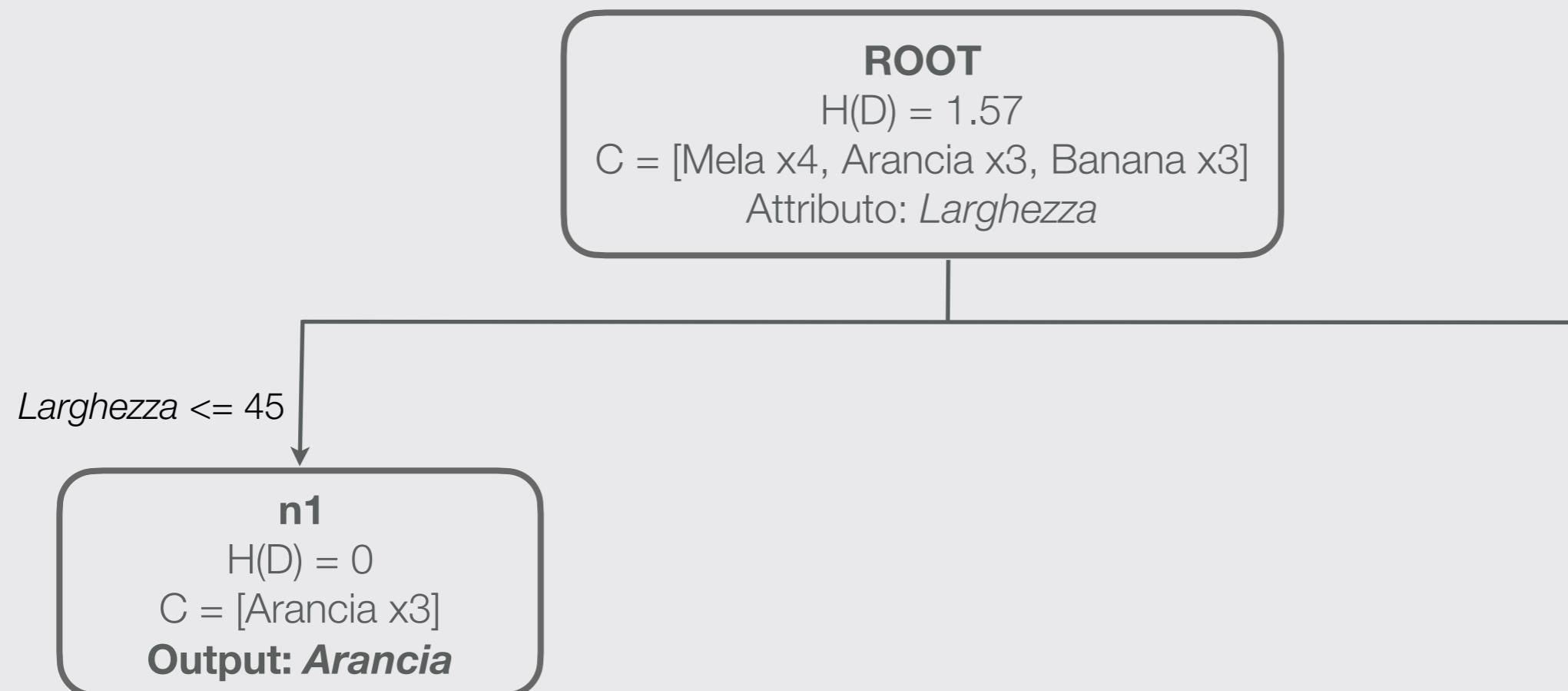
Infatti, utilizzando *Larghezza* come primo attributo di split otterremo:

ROOT
 $H(D) = 1.57$
 $C = [\text{Mela x4}, \text{Arancia x3}, \text{Banana x3}]$

FIA Help Teaching - Esercitazione 3

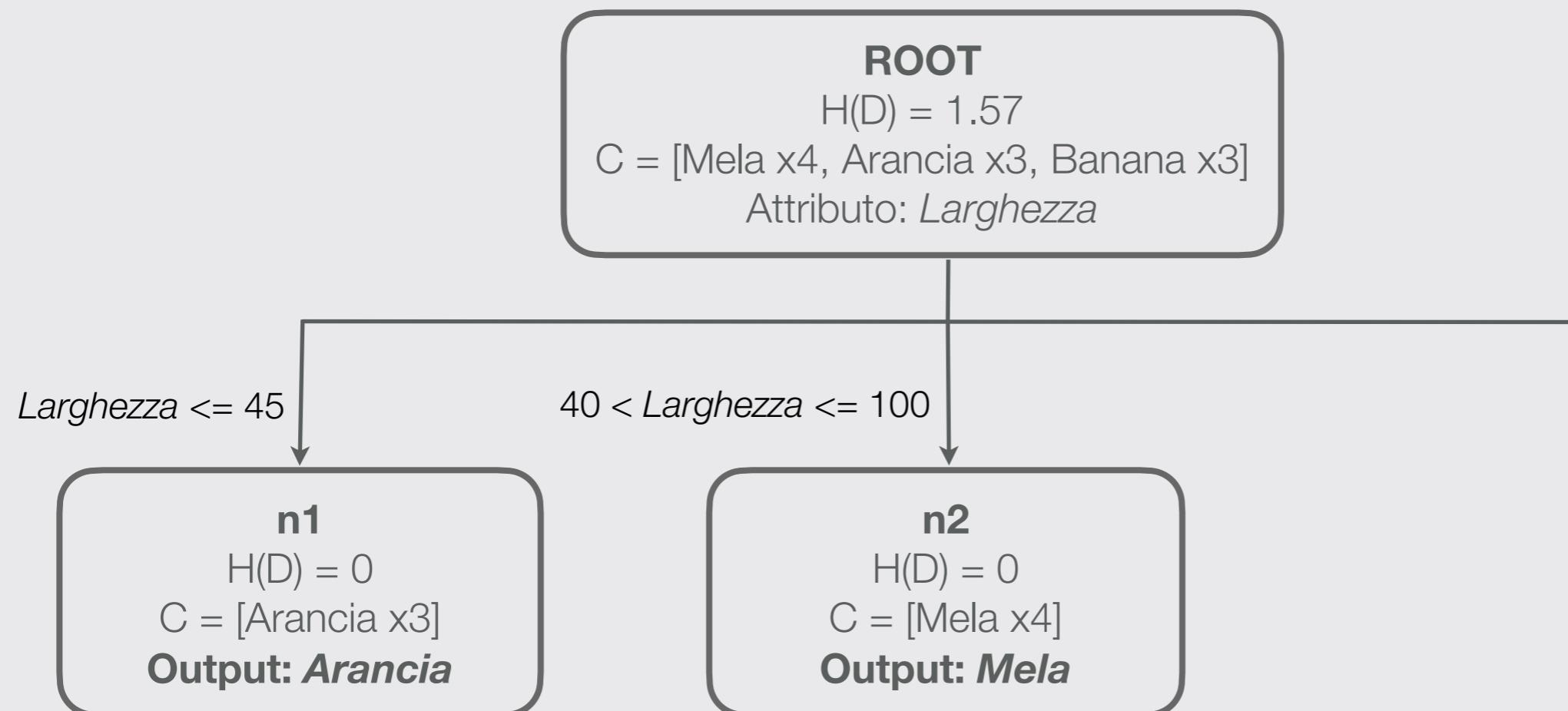
Alberi di decisione - Esercizio 1

Infatti, utilizzando *Larghezza* come primo attributo di split otterremo:



Alberi di decisione - Esercizio 1

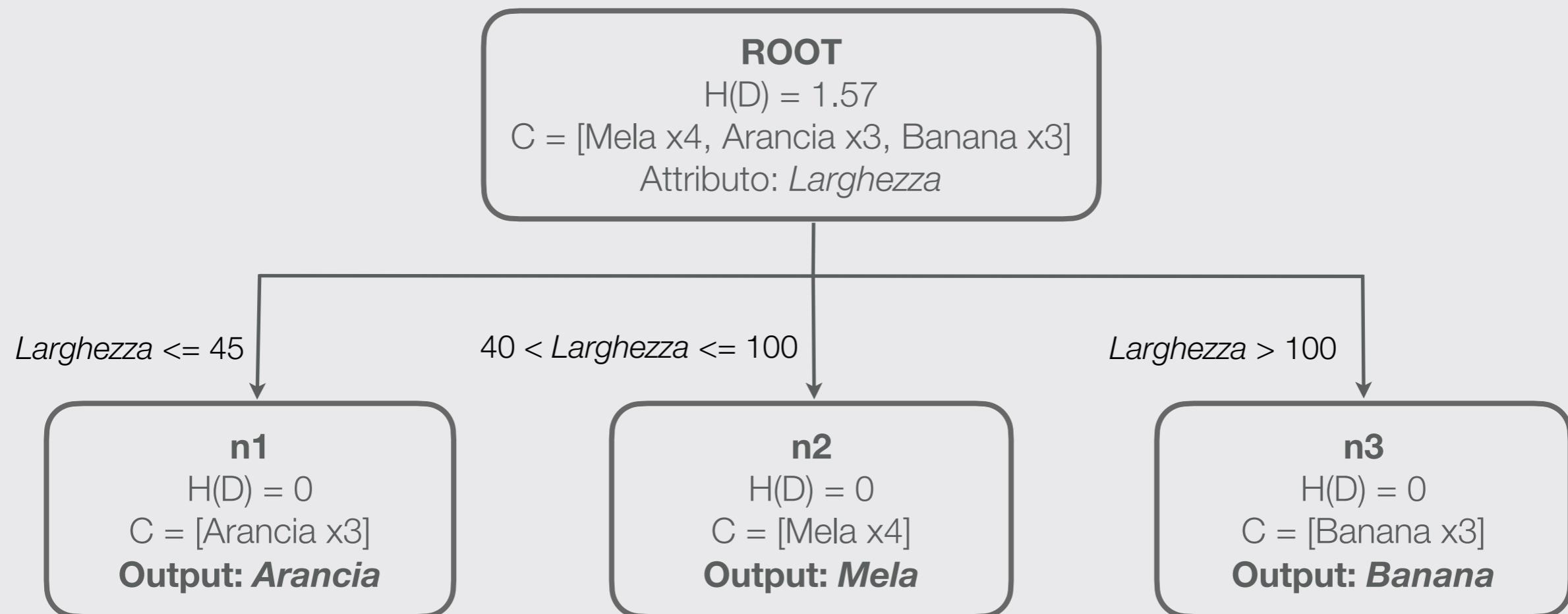
Infatti, utilizzando *Larghezza* come primo attributo di split otterremo:



FIA Help Teaching - Esercitazione 3

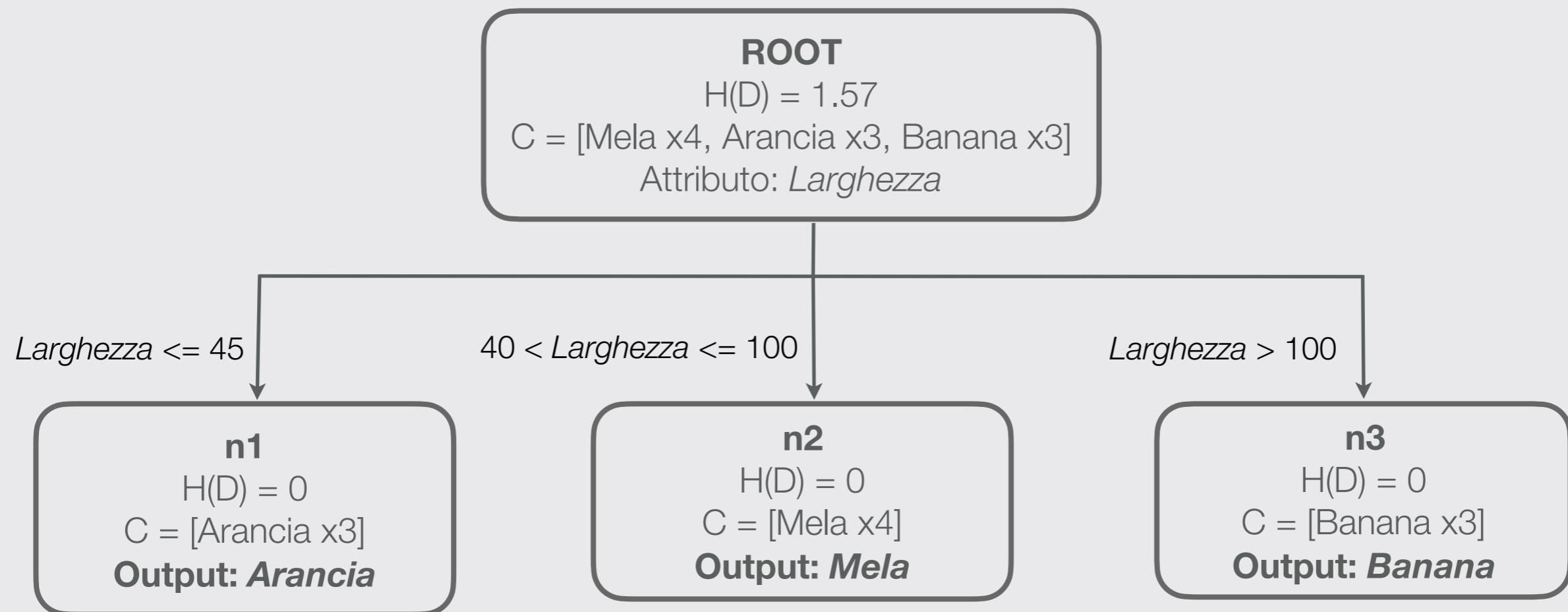
Alberi di decisione - Esercizio 1

Infatti, utilizzando *Larghezza* come primo attributo di split otterremo:



Alberi di decisione - Esercizio 1

Infatti, utilizzando *Larghezza* come primo attributo di split otterremo:

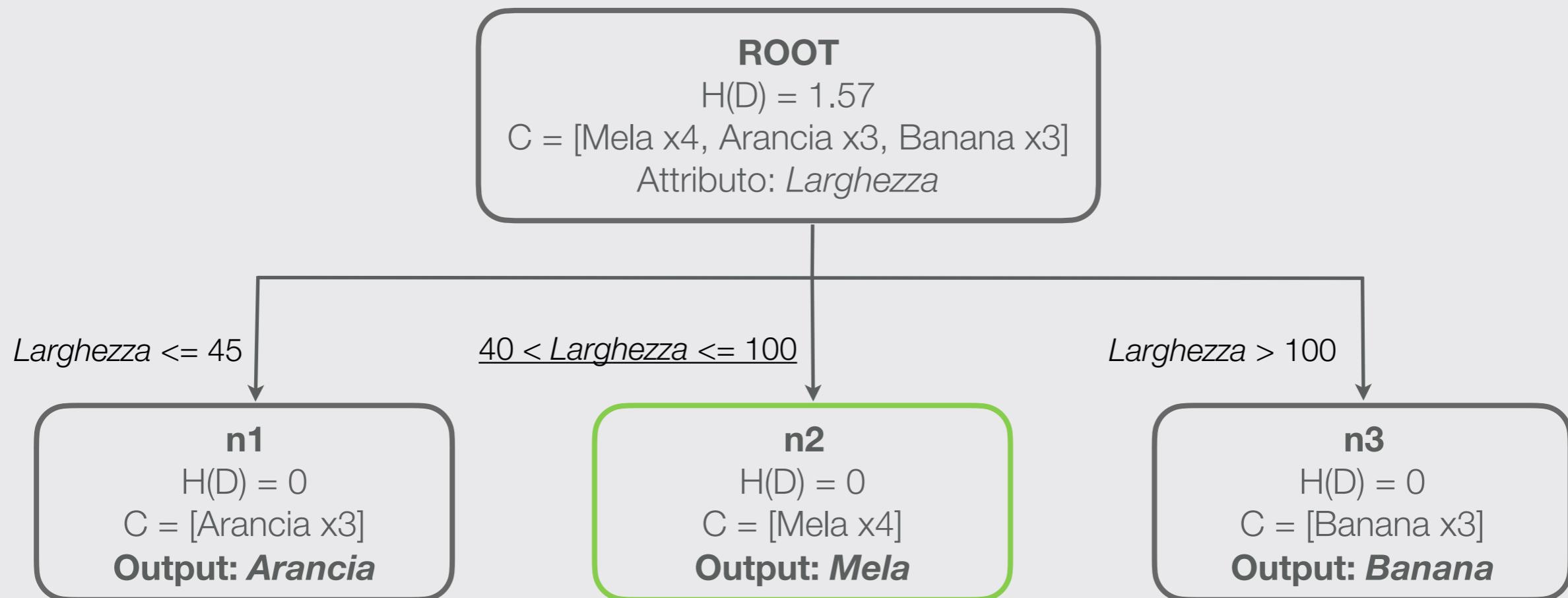


Se volessimo utilizzare quest'albero per predire il seguente frutto, quale sarebbe la risposta?

Giallo	58	63	?
--------	----	----	---

Alberi di decisione - Esercizio 1

Infatti, utilizzando *Larghezza* come primo attributo di split otterremo:



Se volessimo utilizzare quest'albero per predire il seguente frutto, quale sarebbe la risposta?

Giallo	58	63	Mela
--------	----	----	-------------

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Siete un'azienda che vende computer, e siete interessati a capire che tipo di clienti potrebbero potenzialmente acquistare i vostri prodotti. Volete definire un albero di decisione in grado di predire se un cliente comprerà un computer sulla base delle informazioni pregresse che avete dai clienti passati.

Età	Disponibilità economica	Studente	Lavoratore	Compra il computer?
<=30	Alta	No	Sì	Sì
<=30	Media	No	No	No
>30, <=40	Alta	No	No	Sì
>40	Media	No	No	No
>40	Bassa	Sì	Sì	Sì
>40	Bassa	Sì	No	No
>30, <=40	Bassa	Sì	No	No
<=30	Alta	No	No	Sì
<=30	Bassa	Sì	Sì	Sì
>40	Media	Sì	Sì	Sì
>30, <=40	Media	No	No	No

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Calcoliamo l'entropia iniziale del dataset. Abbiamo una classificazione binaria, con Sì = 6/11 e No = 5/11

Età	Disponibilità economica	Studente	Lavoratore	Compra il computer?
<=30	Alta	No	Sì	Sì
<=30	Media	No	No	No
>30, <=40	Alta	No	No	Sì
>40	Media	No	No	No
>40	Bassa	Sì	Sì	Sì
>40	Bassa	Sì	No	No
>30, <=40	Bassa	Sì	No	No
<=30	Alta	No	No	Sì
<=30	Bassa	Sì	Sì	Sì
>40	Media	Sì	Sì	Sì
>30, <=40	Media	No	No	No

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Calcoliamo l'entropia iniziale del dataset. Abbiamo una classificazione binaria, con Sì = 6/11 e No = 5/11

$$H(D) = -\frac{6}{11} \log_2 \frac{6}{11} - \frac{5}{11} \log_2 \frac{5}{11} = 0.99$$

Età	Disponibilità economica	Studente	Lavoratore	Compra il computer?
<=30	Alta	No	Sì	Sì
<=30	Media	No	No	No
>30, <=40	Alta	No	No	Sì
>40	Media	No	No	No
>40	Bassa	Sì	Sì	Sì
>40	Bassa	Sì	No	No
>30, <=40	Bassa	Sì	No	No
<=30	Alta	No	No	Sì
<=30	Bassa	Sì	Sì	Sì
>40	Media	Sì	Sì	Sì
>30, <=40	Media	No	No	No

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Età***.

Età	Compra il computer?
<=30	Sì
<=30	No
>30, <=40	Sì
>40	No
>40	Sì
>40	No
>30, <=40	No
<=30	Sì
<=30	Sì
>40	Sì
>30, <=40	No

$$H(D_{\leq 30}) =$$

$$H(D_{>30, \leq 40}) =$$

$$H(D_{>40}) =$$

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Età***.

Età	Compra il computer?
<=30	Sì
<=30	No
>30, <=40	Sì
>40	No
>40	Sì
>40	No
>30, <=40	No
<=30	Sì
<=30	Sì
>40	Sì
>30, <=40	No

$$H(D_{\leq 30}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$$

$$H(D_{>30, \leq 40}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.91$$

$$H(D_{>40}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Età***.

Età	Compra il computer?
<=30	Sì
<=30	No
>30, <=40	Sì
>40	No
>40	Sì
>40	No
>30, <=40	No
<=30	Sì
<=30	Sì
>40	Sì
>30, <=40	No

$$H(D_{\leq 30}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$$

$$H(D_{>30, \leq 40}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.91$$

$$H(D_{>40}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

$$Gain(D, Eta) = 0.99 - \left(\frac{4}{11} \cdot 0.81 + \frac{3}{11} \cdot 0.91 + \frac{4}{11} \cdot 1 \right) = 0.083$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Disponibilità economica***.

Disponibilità economica	Compro il computer?
Alta	Sì
Media	No
Alta	Sì
Media	No
Bassa	Sì
Bassa	No
Bassa	No
Alta	Sì
Bassa	Sì
Media	Sì
Media	No

$$H(D_{alta}) =$$

$$H(D_{media}) =$$

$$H(D_{bassa}) =$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Disponibilità economica***.

Disponibilità economica	Compro il computer?
Alta	Sì
Media	No
Alta	Sì
Media	No
Bassa	Sì
Bassa	No
Bassa	No
Alta	Sì
Bassa	Sì
Media	Sì
Media	No

$$H(D_{alta}) = -\frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{media}) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.81$$

$$H(D_{bassa}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Disponibilità economica***.

Disponibilità economica	Compra il computer?
Alta	Sì
Media	No
Alta	Sì
Media	No
Bassa	Sì
Bassa	No
Bassa	No
Alta	Sì
Bassa	Sì
Media	Sì
Media	No

$$H(D_{alta}) = -\frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{media}) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.81$$

$$H(D_{bassa}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

$$Gain(D, DispEc) = 0.99 - (\frac{3}{11} \cdot 0 + \frac{4}{11} \cdot 0.81 + \frac{4}{11} \cdot 1) = 0.331$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Studente***.

Studente	Compra il computer?
No	Sì
No	No
No	Sì
No	No
Sì	Sì
Sì	No
Sì	No
No	Sì
Sì	Sì
Sì	Sì
No	No

$$H(D_{si}) =$$

$$H(D_{no}) =$$

Alberi di decisione - Esercizio 2

Calcoliamo l'**Information Gain** per **Studente**.

Studente	Compra il computer?
No	Sì
No	No
No	Sì
No	No
Sì	Sì
Sì	No
Sì	No
No	Sì
Sì	Sì
Sì	Sì
No	No

$$H(D_{si}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$H(D_{no}) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = 1$$

Alberi di decisione - Esercizio 2

Calcoliamo l'**Information Gain** per **Studente**.

Studente	Compra il computer?
No	Sì
No	No
No	Sì
No	No
Sì	Sì
Sì	No
Sì	No
No	Sì
Sì	Sì
Sì	Sì
No	No

$$H(D_{si}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.97$$

$$H(D_{no}) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = 1$$

$$Gain(D, Studente) = 0.99 - \left(\frac{5}{11} \cdot 0.97 + \frac{6}{11} \cdot 1 \right) = 0.003$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Calcoliamo l'***Information Gain*** per ***Lavoratore***.

Lavoratore	Compra il computer?
Sì	Sì
No	No
No	Sì
No	No
Sì	Sì
No	No
No	No
No	Sì
Sì	Sì
Sì	Sì
No	No

$$H(D_{si}) =$$

$$H(D_{no}) =$$

Alberi di decisione - Esercizio 2

Calcoliamo l'**Information Gain** per **Lavoratore**.

Lavoratore	Compra il computer?
Sì	Sì
No	No
No	Sì
No	No
Sì	Sì
No	No
No	No
No	Sì
Sì	Sì
Sì	Sì
No	No

$$H(D_{si}) = -\frac{4}{4} \log_2 \frac{4}{4} = 0$$

$$H(D_{no}) = -\frac{2}{7} \log_2 \frac{2}{7} - \frac{5}{7} \log_2 \frac{5}{7} = 0.86$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Calcoliamo l'**Information Gain** per **Lavoratore**.

Lavoratore	Compra il computer?
Sì	Sì
No	No
No	Sì
No	No
Sì	Sì
No	No
No	No
No	Sì
Sì	Sì
Sì	Sì
No	No

$$H(D_{si}) = -\frac{4}{4} \log_2 \frac{4}{4} = 0$$

$$H(D_{no}) = -\frac{2}{7} \log_2 \frac{2}{7} - \frac{5}{7} \log_2 \frac{5}{7} = 0.86$$

$$Gain(D, Lavoratore) = 0.99 - \left(\frac{4}{11} \cdot 0 + \frac{7}{11} \cdot 0.86 \right) = 0.442$$

Alberi di decisione - Esercizio 2

A questo punto, possiamo finalmente scegliere il nodo radice dell'albero.

Scegliamo l'attributo A che ha dato $Gain(D,A)$ maggiore.

$$Gain(D, Eta) = 0.99 - \left(\frac{4}{11} \cdot 0.81 + \frac{3}{11} \cdot 0.91 + \frac{4}{11} \cdot 1 \right) = 0.083$$

$$Gain(D, DispEc) = 0.99 - \left(\frac{3}{11} \cdot 0 + \frac{4}{11} \cdot 0.81 + \frac{4}{11} \cdot 1 \right) = 0.331$$

$$Gain(D, Studente) = 0.99 - \left(\frac{5}{11} \cdot 0.97 + \frac{6}{11} \cdot 1 \right) = 0.003$$

$$Gain(D, Lavoratore) = 0.99 - \left(\frac{4}{11} \cdot 0 + \frac{7}{11} \cdot 0.86 \right) = 0.442$$

Alberi di decisione - Esercizio 2

A questo punto, possiamo finalmente scegliere il nodo radice dell'albero.

Scegliamo l'attributo A che ha dato $Gain(D,A)$ maggiore.

$$Gain(D, Eta) = 0.99 - \left(\frac{4}{11} \cdot 0.81 + \frac{3}{11} \cdot 0.91 + \frac{4}{11} \cdot 1 \right) = 0.083$$

$$Gain(D, DispEc) = 0.99 - \left(\frac{3}{11} \cdot 0 + \frac{4}{11} \cdot 0.81 + \frac{4}{11} \cdot 1 \right) = 0.331$$

$$Gain(D, Studente) = 0.99 - \left(\frac{5}{11} \cdot 0.97 + \frac{6}{11} \cdot 1 \right) = 0.003$$

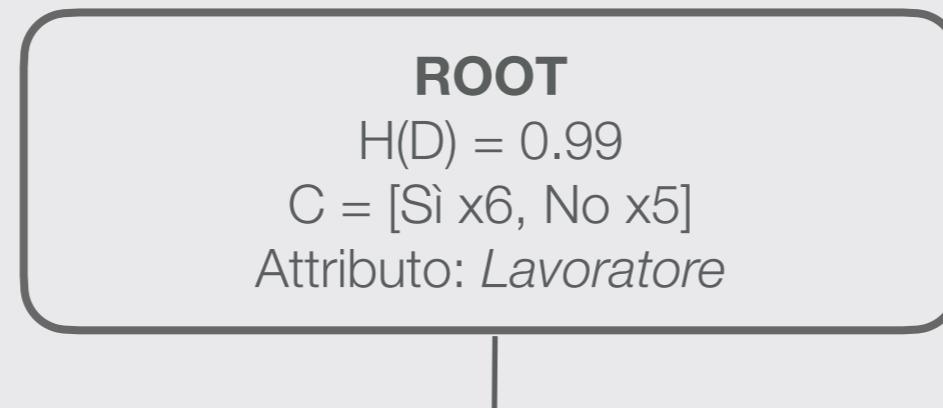
$$Gain(D, Lavoratore) = 0.99 - \left(\frac{4}{11} \cdot 0 + \frac{7}{11} \cdot 0.86 \right) = 0.442$$

Scegliamo quindi l'attributo *Lavoratore*.

In questo caso non siamo fortunati, e dobbiamo continuare.

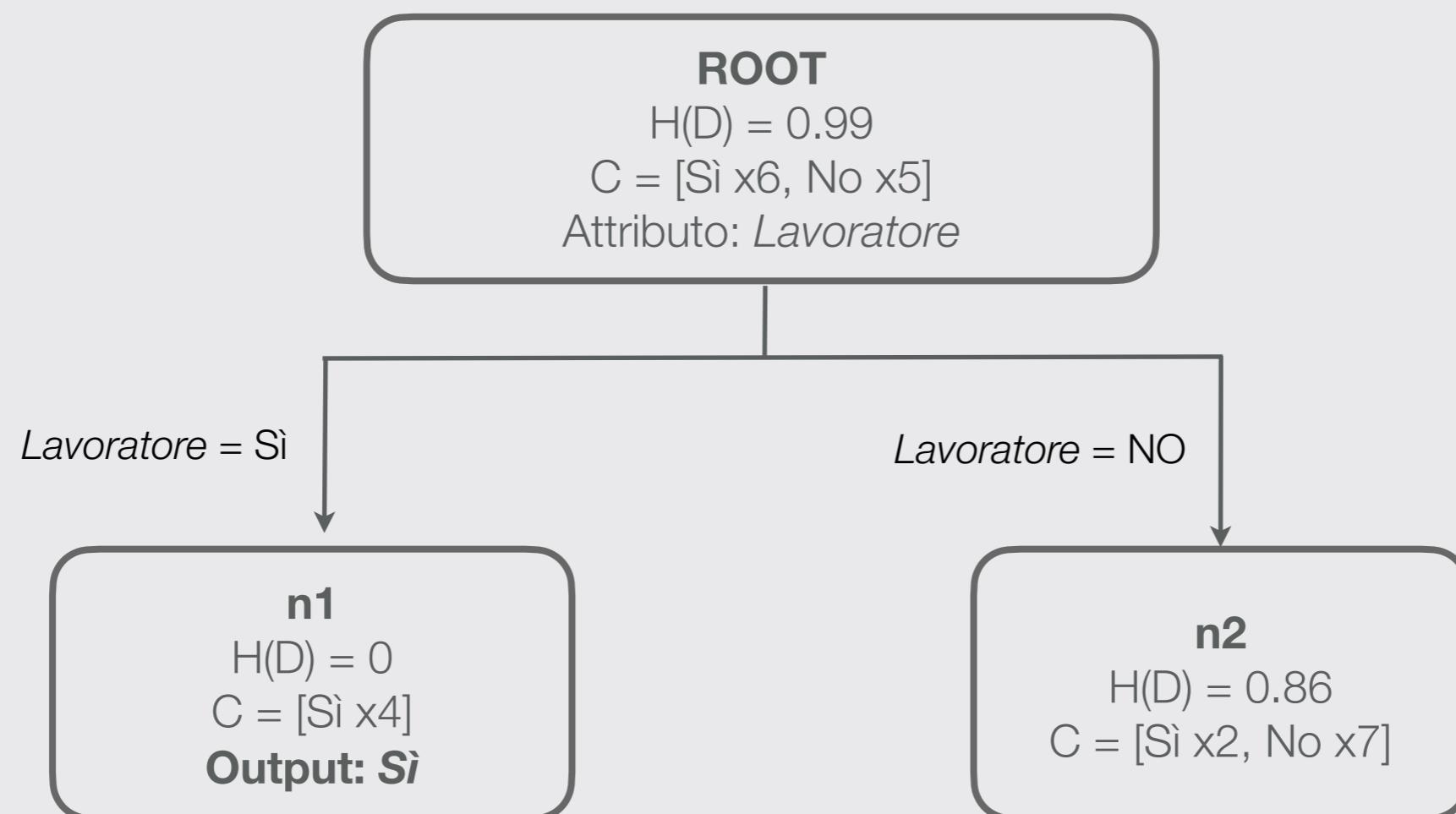
Alberi di decisione - Esercizio 2

Utilizzando *Età* come primo attributo di split otterremo:



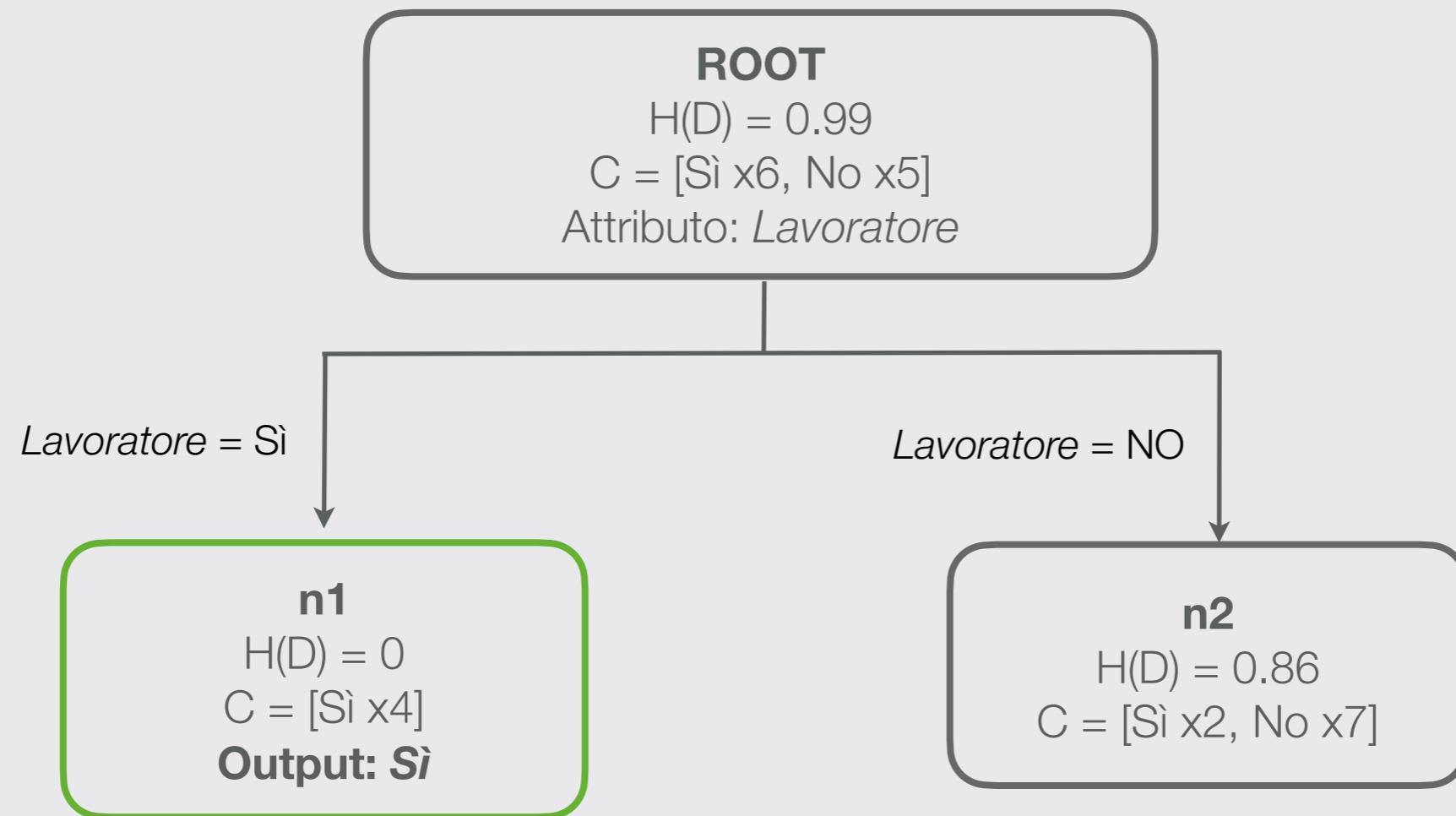
Alberi di decisione - Esercizio 2

Utilizzando *Età* come primo attributo di split otterremo:



Alberi di decisione - Esercizio 2

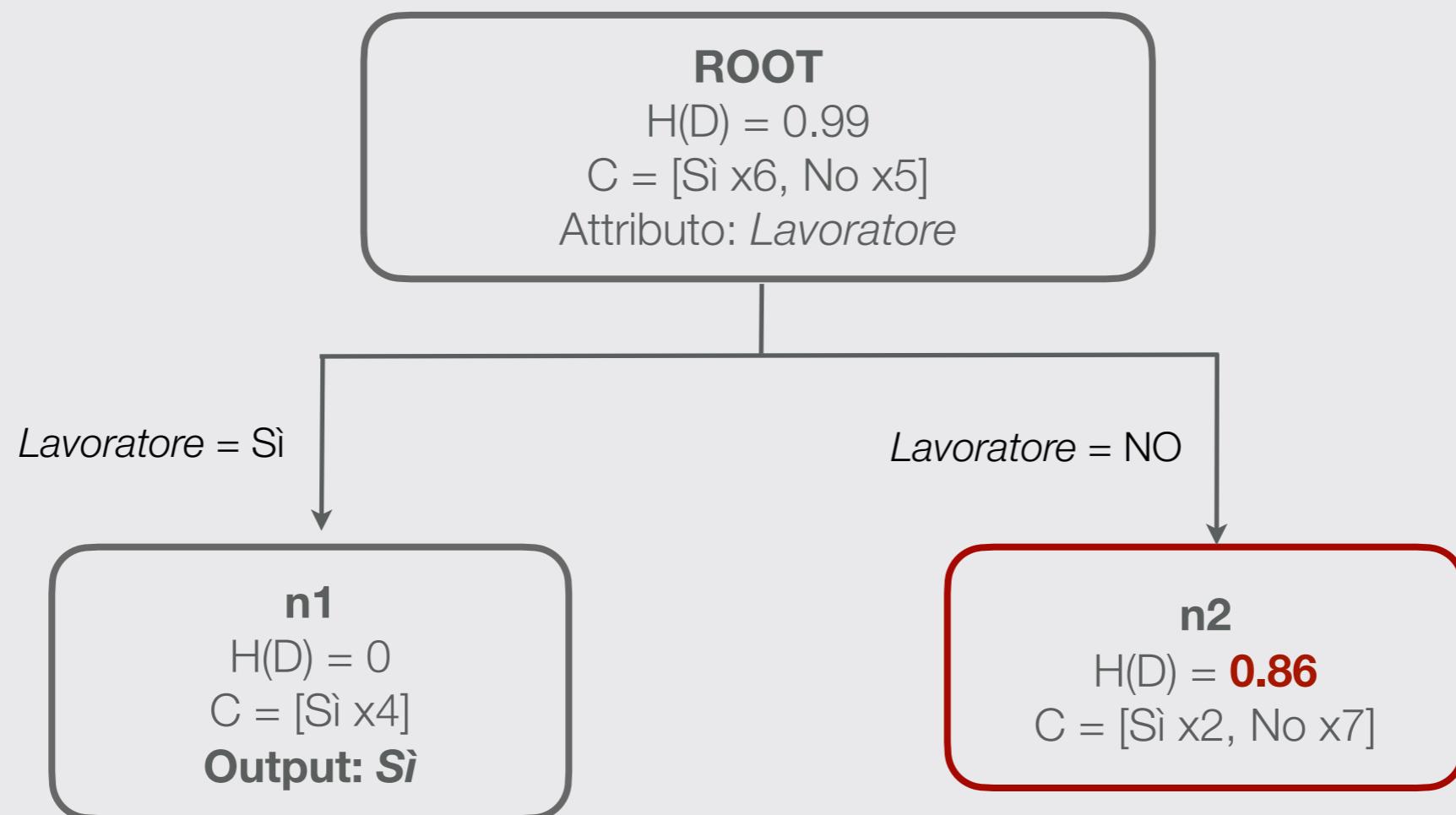
Utilizzando *Età* come primo attributo di split otterremo:



Nel nodo 1, non abbiamo impurità.
Quindi possiamo fermarci e restituire "Sì"
ai campioni che finiscono qui dentro.

Alberi di decisione - Esercizio 2

Utilizzando *Età* come primo attributo di split otterremo:



Nel nodo 2, abbiamo ancora dei campioni misti. Possiamo continuare per creare un altro livello dell'albero e diminuire l'impurità.

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Da questo punto in poi, faremo le nostre considerazioni solo sui dati che sono finiti nel nodo che dobbiamo spartire.

Età	Disponibilità economica	Studente	Lavoratore	Compra il computer?
<=30	Alta	No	Sì	Sì
<=30	Media	No	No	No
>30, <=40	Alta	No	No	Sì
>40	Media	No	No	No
>40	Bassa	Sì	Sì	Sì
>40	Bassa	Sì	No	No
>30, <=40	Bassa	Sì	No	No
<=30	Alta	No	No	Sì
<=30	Bassa	Sì	Sì	Sì
>40	Media	Sì	Sì	Sì
>30, <=40	Media	No	No	No

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Da questo punto in poi, faremo le nostre considerazioni solo sui dati che sono finiti nel nodo che dobbiamo spartire.

Età	Disponibilità economica	Studente	Lavoratore	Compra il computer?
<=30	Alta	No	Sì	Sì
<=30	Media	No	No	No
>30, <=40	Alta	No	No	Sì
>40	Media	No	No	No
>40	Bassa	Sì	Sì	Sì
>40	Bassa	Sì	No	No
>30, <=40	Bassa	Sì	No	No
<=30	Alta	No	No	Sì
<=30	Bassa	Sì	Sì	Sì
>40	Media	Sì	Sì	Sì
>30, <=40	Media	No	No	No

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Da questo punto in poi, faremo le nostre considerazioni solo sui dati che sono finiti nel nodo che dobbiamo spartire.

Età	Disponibilità economica	Studente	Lavoratore	Compra il computer?
<=30	Media	No	No	No
>30, <=40	Alta	No	No	Sì
>40	Media	No	No	No
>40	Bassa	Sì	No	No
>30, <=40	Bassa	Sì	No	No
<=30	Alta	No	No	Sì
>30, <=40	Media	No	No	No

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Da questo punto in poi, faremo le nostre considerazioni solo sui dati che sono finiti nel nodo che dobbiamo splittare.

Consideriamo da ora solo gli altri attributi.

Età	Disponibilità economica	Studente	Lavoratore	Compra il computer?
<=30	Media	No	No	No
>30, <=40	Alta	No	No	Sì
>40	Media	No	No	No
>40	Bassa	Sì	No	No
>30, <=40	Bassa	Sì	No	No
<=30	Alta	No	No	Sì
>30, <=40	Media	No	No	No

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Da questo punto in poi, faremo le nostre considerazioni solo sui dati che sono finiti nel nodo che dobbiamo spartire.

Consideriamo da ora solo gli altri attributi.

Età	Disponibilità economica	Studente	Compra il computer?
<=30	Media	No	No
>30, <=40	Alta	No	Sì
>40	Media	No	No
>40	Bassa	Sì	No
>30, <=40	Bassa	Sì	No
<=30	Alta	No	Sì
>30, <=40	Media	No	No

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Ricominciamo il giro e calcoliamo l'**Information Gain** per **Età**.

Età	Compra il computer?
<=30	No
>30, <=40	Sì
>40	No
>40	No
>30, <=40	No
<=30	Sì
>30, <=40	No

$$H(D_{\leq 30}) =$$

$$H(D_{>30, \leq 40}) =$$

$$H(D_{>40}) =$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Ricominciamo il giro e calcoliamo l'**Information Gain** per **Età**.

Età	Compra il computer?
<=30	No
>30, <=40	Sì
>40	No
>40	No
>30, <=40	No
<=30	Sì
>30, <=40	No

$$H(D_{\leq 30}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$H(D_{>30, \leq 40}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.91$$

$$H(D_{>40}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

Alberi di decisione - Esercizio 2

Ricominciamo il giro e calcoliamo l'***Information Gain*** per ***Età***.

Età	Compra il computer?
<=30	No
>30, <=40	Sì
>40	No
>40	No
>30, <=40	No
<=30	Sì
>30, <=40	No

$$H(D_{\leq 30}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$H(D_{>30, \leq 40}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.91$$

$$H(D_{>40}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$Gain(D, Eta) = 0.86 - (\frac{2}{7} \cdot 1 + \frac{3}{7} \cdot 0.91 + \frac{2}{7} \cdot 0) = 0.184$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Ricominciamo il giro e calcoliamo l'***Information Gain*** per ***Disponibilità Economica***.

Disponibilità economica	Compra il computer?
Media	No
Alta	Sì
Media	No
Bassa	No
Bassa	No
Alta	Sì
Media	No

$$H(D_{alta}) =$$

$$H(D_{media}) =$$

$$H(D_{bassa}) =$$

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Ricominciamo il giro e calcoliamo l'**Information Gain** per **Disponibilità Economica**.

Disponibilità economica	Compra il computer?
Media	No
Alta	Sì
Media	No
Bassa	No
Bassa	No
Alta	Sì
Media	No

$$H(D_{bassa}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{media}) = -\frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{alta}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

Alberi di decisione - Esercizio 2

Ricominciamo il giro e calcoliamo l'***Information Gain*** per ***Disponibilità Economica***.

Disponibilità economica	Compra il computer?
Media	No
Alta	Sì
Media	No
Bassa	No
Bassa	No
Alta	Sì
Media	No

$$H(D_{bassa}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

$$H(D_{media}) = -\frac{3}{3} \log_2 \frac{3}{3} = 0$$

$$H(D_{alta}) = -\frac{2}{2} \log_2 \frac{2}{2} = 0$$

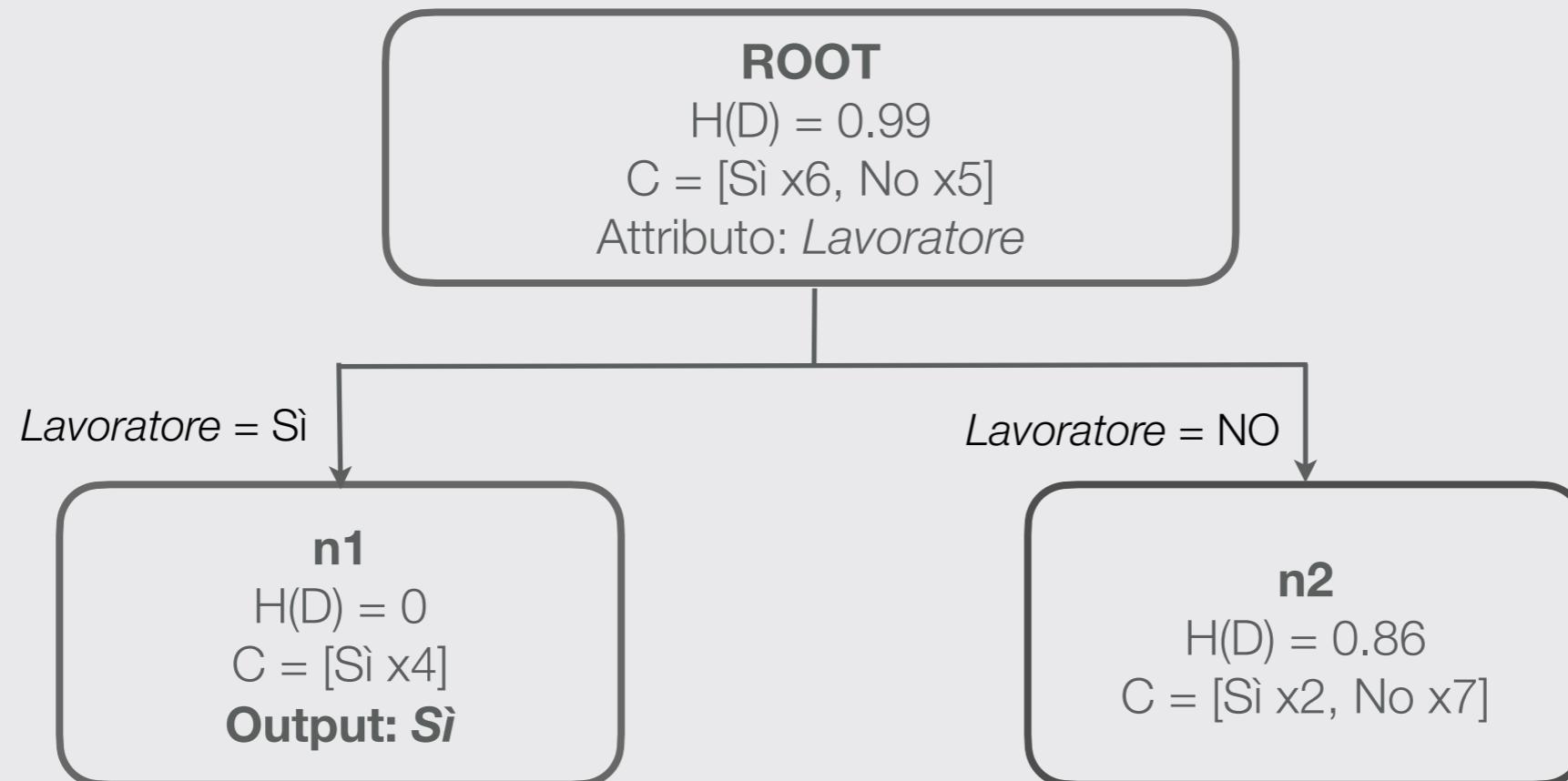
$$Gain(D, DispEc) = 0.86 - (\frac{2}{7} \cdot 0 + \frac{3}{7} \cdot 0 + \frac{2}{7} \cdot 0) = 0.86$$

Possiamo fermarci, poiché abbiamo trovato un attributo di split che azzera l'impurità!

FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

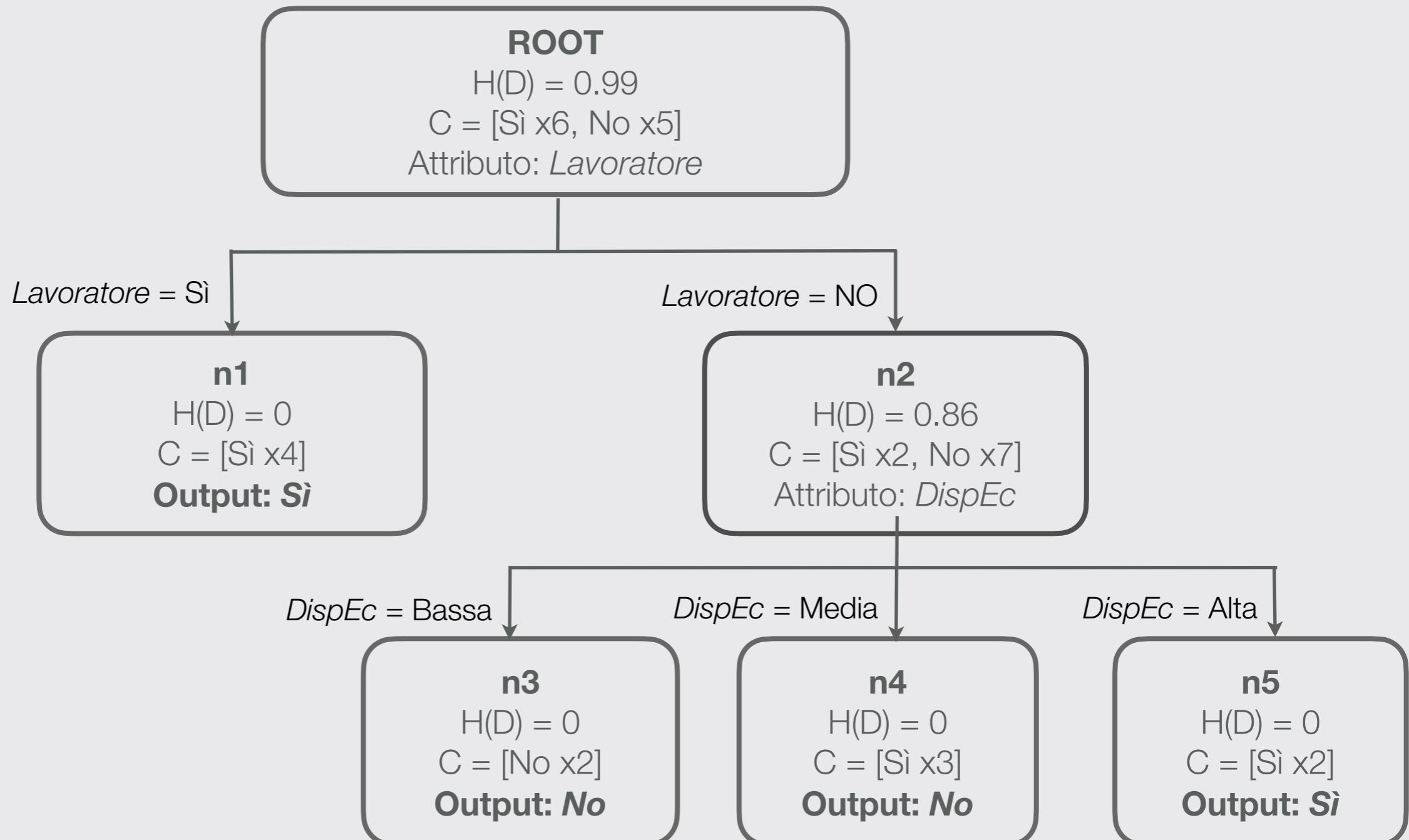
Aggiungiamo le nuove informazioni all'albero:



FIA Help Teaching - Esercitazione 3

Alberi di decisione - Esercizio 2

Aggiungiamo le nuove informazioni all'albero:



Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

c1 = {4,
c2 = {17,
c3 = {23,
c4 = {31,

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1,$$

$$c_2 = \{17,$$

$$c_3 = \{23,$$

$$c_4 = \{31,$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8,$$

$$c_2 = \{17,$$

$$c_3 = \{23,$$

$$c_4 = \{31,$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9,$$

$$c_2 = \{17,$$

$$c_3 = \{23,$$

$$c_4 = \{31,$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\}$$

$$c_2 = \{17,$$

$$c_3 = \{23,$$

$$c_4 = \{31,$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\}$$

$$c_2 = \{17, 13, 14, 19\}$$

$$c_3 = \{23,$$

$$c_4 = \{31,$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\}$$

$$c_2 = \{17, 13, 14, 19, 20\} \leftarrow \text{20 ha la stessa differenza (in valore assoluto) sia da 17 che da 23.}$$

$$c_3 = \{23,$$

$$c_4 = \{31,$$

Scegliamo arbitrariamente dove metterlo.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\}$$

$$c_2 = \{17, 13, 14, 19, 20\}$$

$$c_3 = \{23, 21, 25\}$$

$$c_4 = \{31\}$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\}$$

$$c_2 = \{17, 13, 14, 19, 20\}$$

$$c_3 = \{23, 21, 25, 27\} \quad \leftarrow \text{Stesso ragionamento per } 27$$

$$c_4 = \{31,$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\}$$

$$c_2 = \{17, 13, 14, 19, 20\}$$

$$c_3 = \{23, 21, 25, 27\}$$

$$c_4 = \{31, 48, 128, 333\}$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\}$$

$$c_2 = \{17, 13, 14, 19, 20\}$$

$$c_3 = \{23, 21, 25, 27\}$$

$$c_4 = \{31, 48, 128, 333\}$$

A questo punto dobbiamo ricalcolare i centroidi. Utilizziamo la **media aritmetica**.

N.B. Non è importante se il nuovo centroide che otteniamo non è uno dei punti iniziali. Servirà solo per ridefinire i cluster.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c' = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\}$$

$$c_3 = \{23, 21, 25, 27\}$$

$$c_4 = \{31, 48, 128, 333\}$$

A questo punto dobbiamo ricalcolare i centroidi. Utilizziamo la **media aritmetica**.

N.B. Non è importante se il nuovo centroide che otteniamo non è uno dei punti iniziali. Servirà solo per ridefinire i cluster.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c'_1 = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\} \longrightarrow c'_2 = (17+13+14+19+20)/5 = 16.6$$

$$c_3 = \{23, 21, 25, 27\}$$

$$c_4 = \{31, 48, 128, 333\}$$

A questo punto dobbiamo ricalcolare i centroidi. Utilizziamo la **media aritmetica**.

N.B. Non è importante se il nuovo centroide che otteniamo non è uno dei punti iniziali. Servirà solo per ridefinire i cluster.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c' = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\} \longrightarrow c' = (17+13+14+19+20)/5 = 16.6$$

$$c_3 = \{23, 21, 25, 27\} \longrightarrow c' = (23+21+25+27)/4 = 24$$

$$c_4 = \{31, 48, 128, 333\}$$

A questo punto dobbiamo ricalcolare i centroidi. Utilizziamo la **media aritmetica**.

N.B. Non è importante se il nuovo centroide che otteniamo non è uno dei punti iniziali. Servirà solo per ridefinire i cluster.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c' = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\} \longrightarrow c' = (17+13+14+19+20)/5 = 16.6$$

$$c_3 = \{23, 21, 25, 27\} \longrightarrow c' = (23+21+25+27)/4 = 24$$

$$c_4 = \{31, 48, 128, 333\} \longrightarrow c' = (31+48+128+333)/4 = 135$$

A questo punto dobbiamo ricalcolare i centroidi. Utilizziamo la **media aritmetica**.

N.B. Non è importante se il nuovo centroide che otteniamo non è uno dei punti iniziali. Servirà solo per ridefinire i cluster.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Avendo già i centroidi, il primo step è assegnare ogni punto a ciascuno di essi.

Li assegniamo sulla base della vicinanza. Essendo punti monodimensionali, usiamo semplicemente la **differenza**.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c' = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\} \longrightarrow c' = (17+13+14+19+20)/5 = 16.6$$

$$c_3 = \{23, 21, 25, 27\} \longrightarrow c' = (23+21+25+27)/4 = 24$$

$$c_4 = \{31, 48, 128, 333\} \longrightarrow c' = (31+48+128+333)/4 = 135$$

Una delle medie esce molto più distante dalle altre. Potremmo fare delle considerazioni per fare data cleaning e individuare degli outlier?

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Osserviamo la distribuzione dei dati.



Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Osserviamo la distribuzione dei dati.

Ci sono due valori che si discostano molto dal resto della distribuzione. Questi rappresentano degli **outlier** - valori "anomali" che non rispecchiano le altre osservazioni nei dati - e potrebbero far ottenere dei cluster poco rappresentativi.



Clustering - Esercizio 1

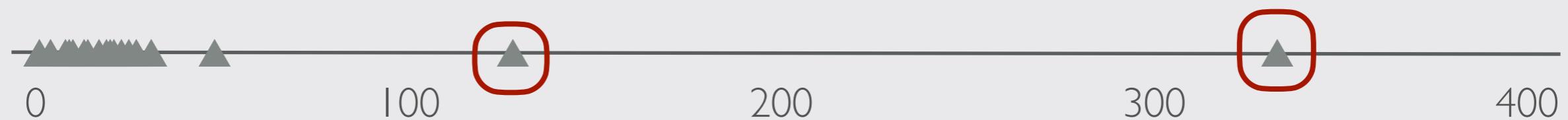
Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, 128, 333}

Osserviamo la distribuzione dei dati.

Ci sono due valori che si discostano molto dal resto della distribuzione. Questi rappresentano degli **outlier** - valori "anomali" che non rispecchiano le altre osservazioni nei dati - e potrebbero far ottenere dei cluster poco rappresentativi.



Sulla base di questa considerazione e della media aritmetica molto diversa dalle altre ottenuta prima, possiamo motivare la loro **rimozione** dall'insieme di dati durante l'esecuzione dell'algoritmo di clustering.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riprendiamo i cluster precedenti, senza considerare gli outlier e ricalcoliamo la media dell'ultimo cluster.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 333~~}

Riprendiamo i cluster precedenti, senza considerare gli outlier e ricalcoliamo la media dell'ultimo cluster.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c' = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\} \longrightarrow c' = (17+13+14+19+20)/5 = 16.6$$

$$c_3 = \{23, 21, 25, 27\} \longrightarrow c' = (23+21+25+27)/4 = 24$$

$$c_4 = \{31, 48, 128, 333\}$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riprendiamo i cluster precedenti, senza considerare gli outlier e ricalcoliamo la media dell'ultimo cluster.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c' = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\} \longrightarrow c' = (17+13+14+19+20)/5 = 16.6$$

$$c_3 = \{23, 21, 25, 27\} \longrightarrow c' = (23+21+25+27)/4 = 24$$

$$c_4 = \{31, 48\}$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riprendiamo i cluster precedenti, senza considerare gli outlier e ricalcoliamo la media dell'ultimo cluster.

$$c_1 = \{4, 1, 8, 9, 10\} \longrightarrow c' = (4+1+8+9+10)/5 = 6.4$$

$$c_2 = \{17, 13, 14, 19, 20\} \longrightarrow c' = (17+13+14+19+20)/5 = 16.6$$

$$c_3 = \{23, 21, 25, 27\} \longrightarrow c' = (23+21+25+27)/4 = 24$$

$$c_4 = \{31, 48\} \longrightarrow c' = (31+48)/2 = 39.5$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riprendiamo i cluster precedenti, senza considerare gli outlier e ricalcoliamo la media dell'ultimo cluster.

$$\begin{array}{ll} c_1 = \{ & c' = 6.4 \\ c_2 = \{ & c' = 16.6 \\ c_3 = \{ & c' = 24 \\ c_4 = \{ & c' = 39.5 \end{array}$$

Riassegniamo ora i dati ai cluster tenendo conto dei nuovi centroidi.

Ci fermeremo quando non cambieranno più i cluster tra un'iterazione e l'altra.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riprendiamo i cluster precedenti, senza considerare gli outlier e ricalcoliamo la media dell'ultimo cluster.

$$c_1 = \{1, 4, 8, 9, 10\}$$

$$c'_1 = 6.4$$

$$c_2 = \{13, 14, 17, 19, 20\}$$

$$c'_2 = 16.6$$

$$c_3 = \{21, 23, 25, 27, 31\}$$

$$c'_3 = 24$$

$$c_4 = \{48\}$$

$$c'_4 = 39.5$$

Riassegniamo ora i dati ai cluster tenendo conto dei nuovi centroidi.

Ci fermeremo quando non cambieranno più i cluster tra un'iterazione e l'altra.

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

I cluster sono cambiati rispetto a prima, per cui ricalcoliamo i centroidi.

$$c_1 = \{1, 4, 8, 9, 10\}$$

$$c_2 = \{13, 14, 17, 19, 20\}$$

$$c_3 = \{21, 23, 25, 27, 31\}$$

$$c_4 = \{48\}$$

$$c'' = 6.4$$

$$c'' = 16.6$$

$$c'' =$$

$$c'' =$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

I cluster sono cambiati rispetto a prima, per cui ricalcoliamo i centroidi.

$$c_1 = \{1, 4, 8, 9, 10\}$$

$$c_2 = \{13, 14, 17, 19, 20\}$$

$$c_3 = \{21, 23, 25, 27, 31\}$$

$$c_4 = \{48\}$$

$$c'' = 6.4$$

$$c'' = 16.6$$

$$c'' = (21+23+25+27+31)/5=25.4$$

$$c'' = 48$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riassegniamo i cluster sulla base dei nuovi centroidi.

c1 =

c2 =

c3 =

c4 =

$c'' = 6.4$

$c'' = 16.6$

$c'' = 25.4$

$c'' = 48$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riassegniamo i cluster sulla base dei nuovi centroidi.

$$c_1 = \{1, 4, 8, 9, 10\}$$

$$c'' = 6.4$$

$$c_2 = \{13, 14, 17, 19, 20\}$$

$$c'' = 16.6$$

$$c_3 = \{21, 23, 25, 27, 31\}$$

$$c'' = 25.4$$

$$c_4 = \{48\}$$

$$c'' = 48$$

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 383~~}

Riassegniamo i cluster sulla base dei nuovi centroidi.

c1 = {1, 4, 8, 9, 10}
c2 = {13, 14, 17, 19, 20}
c3 = {21, 23, 25, 27, 31}
c4 = {48}

$c'' = 6.4$
 $c'' = 16.6$
 $c'' = 25.4$
 $c'' = 48$

I cluster non sono cambiati rispetto all'iterazione precedente, per cui possiamo fermarci!

Clustering - Esercizio 1

Dato il seguente insieme di punti, fornire un clustering di tali punti in 4 partizioni utilizzando l'algoritmo del k-means. Si prenda in considerazione il caso in cui i centroidi iniziali siano 4, 17, 23, 31.

Mostrare tutte le operazioni e le eventuali scelte di date cleaning effettuate. **Commentare poi i risultati ottenuti ed in particolare l'efficacia dell'algoritmo.**

{1, 4, 8, 9, 10, 13, 14, 17, 19, 20, 21, 23, 25, 27, 31, 48, ~~128, 333~~}

Lo farete voi sulla base delle osservazioni fatte durante l'esercizio e dello **studio della teoria!**



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Il clustering gerarchico agglomerativo prevede che si parta da n cluster contenenti un solo elemento ciascuno.

Avendo 6 punti, avremo 6 cluster di partenza.

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Il clustering gerarchico agglomerativo prevede che si parta da n cluster contenenti un solo elemento ciascuno.

Avendo 6 punti, avremo 6 cluster di partenza.

$$c_1 = \{A(1, 2)\}$$

$$c_2 = \{B(2, 5)\}$$

$$c_3 = \{C(3, 5)\}$$

$$c_4 = \{D(6, 3)\}$$

$$c_5 = \{E(9, 3)\}$$

$$c_6 = \{F(10, 4)\}$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

$\{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)\}$

Il clustering gerarchico agglomerativo prevede che si parta da n cluster contenenti un solo elemento ciascuno.

Avendo 6 punti, avremo 6 cluster di partenza.

$$c_1 = \{A(1, 2)\}$$

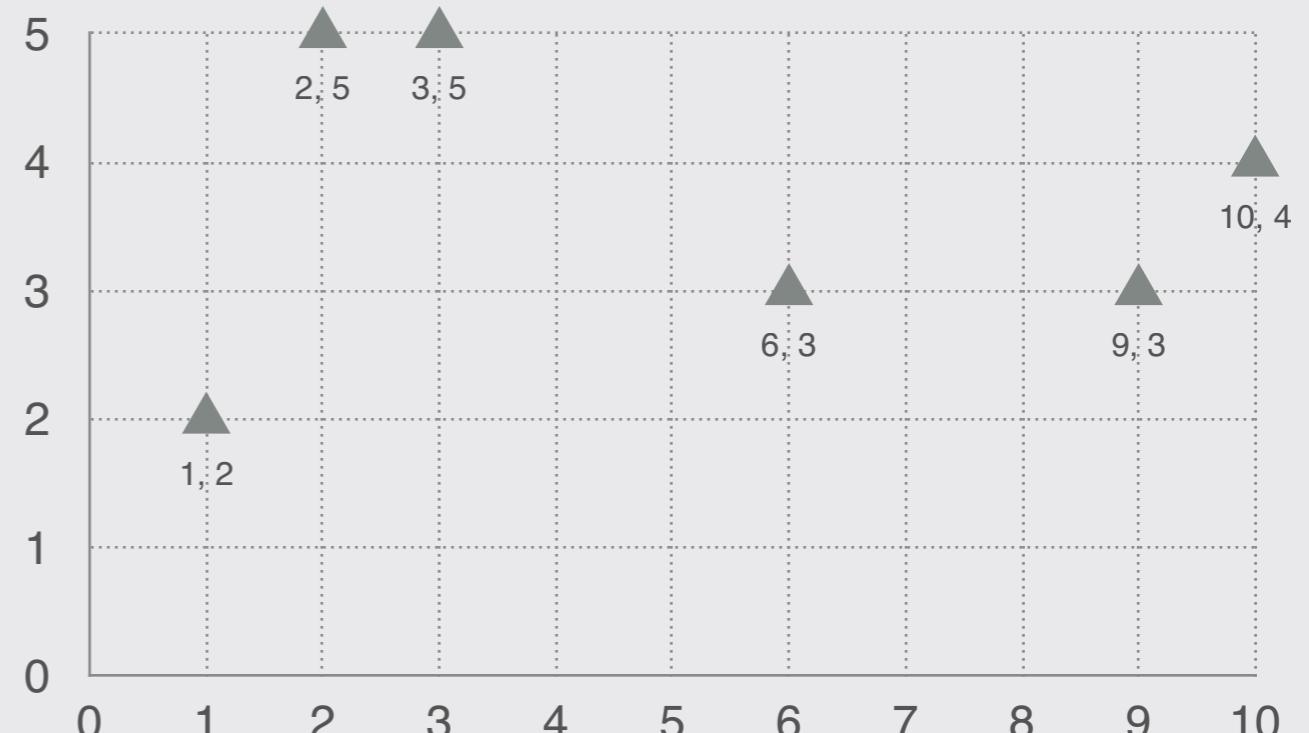
$$c_2 = \{B(2, 5)\}$$

$$c_3 = \{C(3, 5)\}$$

$$c_4 = \{D(6, 3)\}$$

$$c_5 = \{E(9, 3)\}$$

$$c_6 = \{F(10, 4)\}$$



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

$\{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)\}$

Il clustering gerarchico agglomerativo prevede che si parta da n cluster contenenti un solo elemento ciascuno.

Avendo 6 punti, avremo 6 cluster di partenza.

$$c_1 = \{A(1, 2)\}$$

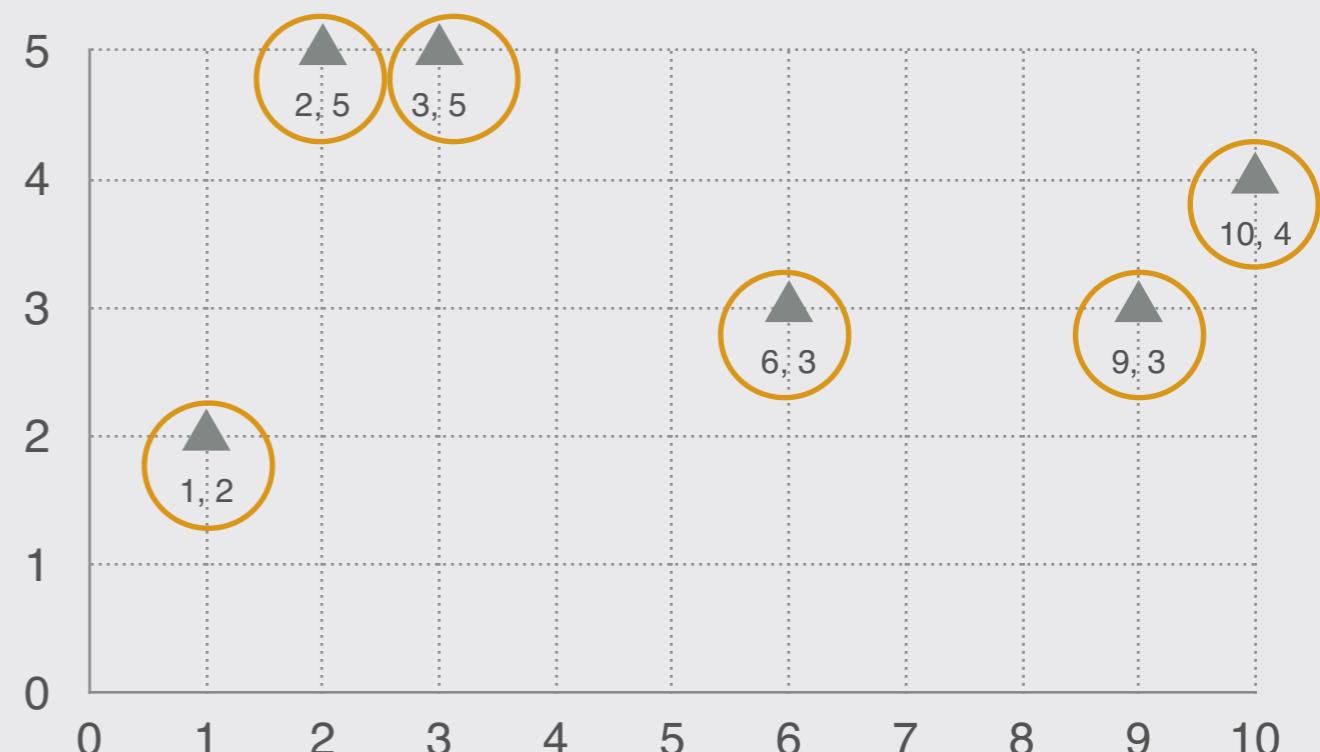
$$c_2 = \{B(2, 5)\}$$

$$c_3 = \{C(3, 5)\}$$

$$c_4 = \{D(6, 3)\}$$

$$c_5 = \{E(9, 3)\}$$

$$c_6 = \{F(10, 4)\}$$



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\					
B	\	\				
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\					
B	\	\				
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

Distanza euclidea:

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16				
B	\	\				
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(A, B) = \sqrt{(2 - 1)^2 + (5 - 2)^2} = \sqrt{10} = 3.16$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6			
B	\	\				
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(A, B) = \sqrt{(2 - 1)^2 + (5 - 2)^2} = \sqrt{10} = 3.16$$

$$d(A, C) = \sqrt{(3 - 1)^2 + (5 - 2)^2} = \sqrt{13} = 3.6$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09		
B	\	\				
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(A, B) = \sqrt{(2 - 1)^2 + (5 - 2)^2} = \sqrt{10} = 3.16$$

$$d(A, C) = \sqrt{(3 - 1)^2 + (5 - 2)^2} = \sqrt{13} = 3.6$$

$$d(A, D) = \sqrt{(6 - 1)^2 + (3 - 2)^2} = \sqrt{26} = 5.09$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	
B	\	\				
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(A, B) = \sqrt{(2 - 1)^2 + (5 - 2)^2} = \sqrt{10} = 3.16$$

$$d(A, C) = \sqrt{(3 - 1)^2 + (5 - 2)^2} = \sqrt{13} = 3.6$$

$$d(A, D) = \sqrt{(6 - 1)^2 + (3 - 2)^2} = \sqrt{26} = 5.09$$

$$d(A, E) = \sqrt{(9 - 1)^2 + (3 - 2)^2} = \sqrt{65} = 8.06$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\				
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(A, B) = \sqrt{(2 - 1)^2 + (5 - 2)^2} = \sqrt{10} = 3.16$$

$$d(A, C) = \sqrt{(3 - 1)^2 + (5 - 2)^2} = \sqrt{13} = 3.6$$

$$d(A, D) = \sqrt{(6 - 1)^2 + (3 - 2)^2} = \sqrt{26} = 5.09$$

$$d(A, E) = \sqrt{(9 - 1)^2 + (3 - 2)^2} = \sqrt{65} = 8.06$$

$$d(A, F) = \sqrt{(10 - 1)^2 + (4 - 2)^2} = \sqrt{85} = 9.21$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1			
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(B, C) = \sqrt{(3 - 2)^2 + (5 - 5)^2} = \sqrt{1} = 1$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47		
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(B, C) = \sqrt{(3 - 2)^2 + (5 - 5)^2} = \sqrt{1} = 1$$

$$d(B, D) = \sqrt{(6 - 2)^2 + (3 - 5)^2} = \sqrt{20} = 4.47$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(B, C) = \sqrt{(3 - 2)^2 + (5 - 5)^2} = \sqrt{1} = 1$$

$$d(B, D) = \sqrt{(6 - 2)^2 + (3 - 5)^2} = \sqrt{20} = 4.47$$

$$d(B, E) = \sqrt{(9 - 2)^2 + (3 - 5)^2} = \sqrt{53} = 7.28$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\			
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(B, C) = \sqrt{(3 - 2)^2 + (5 - 5)^2} = \sqrt{1} = 1$$

$$d(B, D) = \sqrt{(6 - 2)^2 + (3 - 5)^2} = \sqrt{20} = 4.47$$

$$d(B, E) = \sqrt{(9 - 2)^2 + (3 - 5)^2} = \sqrt{53} = 7.28$$

$$d(B, F) = \sqrt{(10 - 2)^2 + (4 - 5)^2} = \sqrt{65} = 8.06$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6		
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(C, D) = \sqrt{(6 - 3)^2 + (3 - 5)^2} = \sqrt{13} = 3.6$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6	6.32	
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(C, D) = \sqrt{(6 - 3)^2 + (3 - 5)^2} = \sqrt{13} = 3.6$$

$$d(C, E) = \sqrt{(9 - 3)^2 + (3 - 5)^2} = \sqrt{40} = 6.32$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6	6.32	7.07
D	\	\	\	\		
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(C, D) = \sqrt{(6 - 3)^2 + (3 - 5)^2} = \sqrt{13} = 3.6$$

$$d(C, E) = \sqrt{(9 - 3)^2 + (3 - 5)^2} = \sqrt{40} = 6.32$$

$$d(C, F) = \sqrt{(10 - 3)^2 + (4 - 5)^2} = \sqrt{50} = 7.07$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6	6.32	7.07
D	\	\	\	\	3	
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(D, E) = \sqrt{(9 - 6)^2 + (3 - 3)^2} = \sqrt{9} = 3$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6	6.32	7.07
D	\	\	\	\	3	4.12
E	\	\	\	\	\	
F	\	\	\	\	\	\

$$d(D, E) = \sqrt{(9 - 6)^2 + (3 - 3)^2} = \sqrt{9} = 3$$

$$d(D, F) = \sqrt{(10 - 6)^2 + (4 - 3)^2} = \sqrt{17} = 4.12$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Calcoliamo la distanza euclidea tra ciascuna coppia di cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6	6.32	7.07
D	\	\	\	\	3	4.12
E	\	\	\	\	\	2
F	\	\	\	\	\	\

$$d(E, F) = \sqrt{(10 - 9)^2 + (4 - 3)^2} = \sqrt{2} = 1.41$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

A questo punto, scegliamo la distanza minore e uniamo i due punti in un nuovo cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6	6.32	7.07
D	\	\	\	\	3	4.12
E	\	\	\	\	\	1.41
F	\	\	\	\	\	\

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

A questo punto, scegliamo la distanza minore e uniamo i due punti in un nuovo cluster.

	A	B	C	D	E	F
A	\	3.16	3.6	5.09	8.06	9.21
B	\	\	1	4.47	7.28	8.06
C	\	\	\	3.6	6.32	7.07
D	\	\	\	\	3	4.12
E	\	\	\	\	\	1.41
F	\	\	\	\	\	\

$$c1 = \{A(1, 2)\}$$

$$c2 = \{B(2, 5), C(3, 5)\}$$

$$c3 = \{D(6, 3)\}$$

$$c4 = \{E(9, 3)\}$$

$$c5 = \{F(10, 4)\}$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

A questo punto, scegliamo la distanza minore e uniamo i due punti in un nuovo cluster.

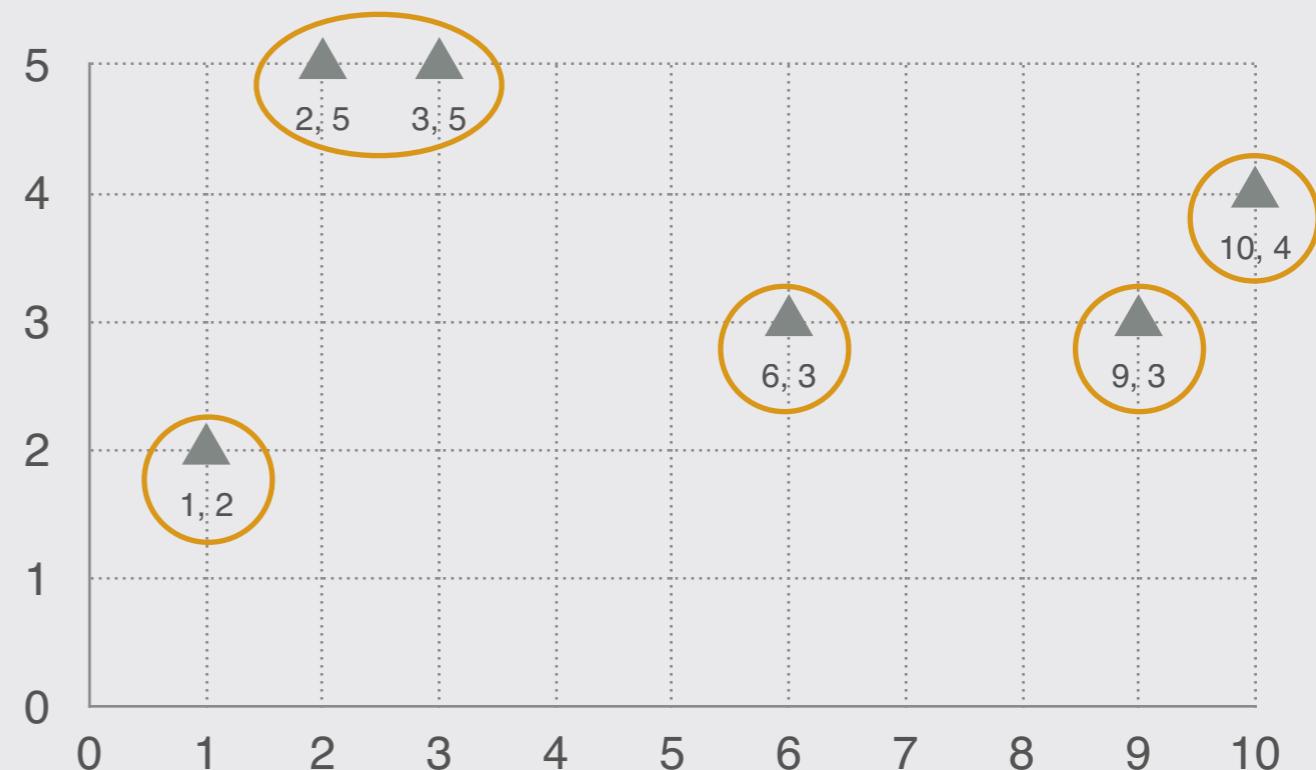
$$c1 = \{A(1, 2)\}$$

$$c2 = \{B(2, 5), C(3, 5)\}$$

$$c3 = \{D(6, 3)\}$$

$$c4 = \{E(9, 3)\}$$

$$c5 = \{F(10, 4)\}$$



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

A questo punto, scegliamo la distanza minore e uniamo i due punti in un nuovo cluster.

$$c1 = \{A(1, 2)\}$$

$$c2 = \{B(2, 5), C(3, 5)\}$$

$$c3 = \{D(6, 3)\}$$

$$c4 = \{E(9, 3)\}$$

$$c5 = \{F(10, 4)\}$$

Come calcoliamo la distanza tra
questo cluster e gli altri visto che
ora ci sono due punti?



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

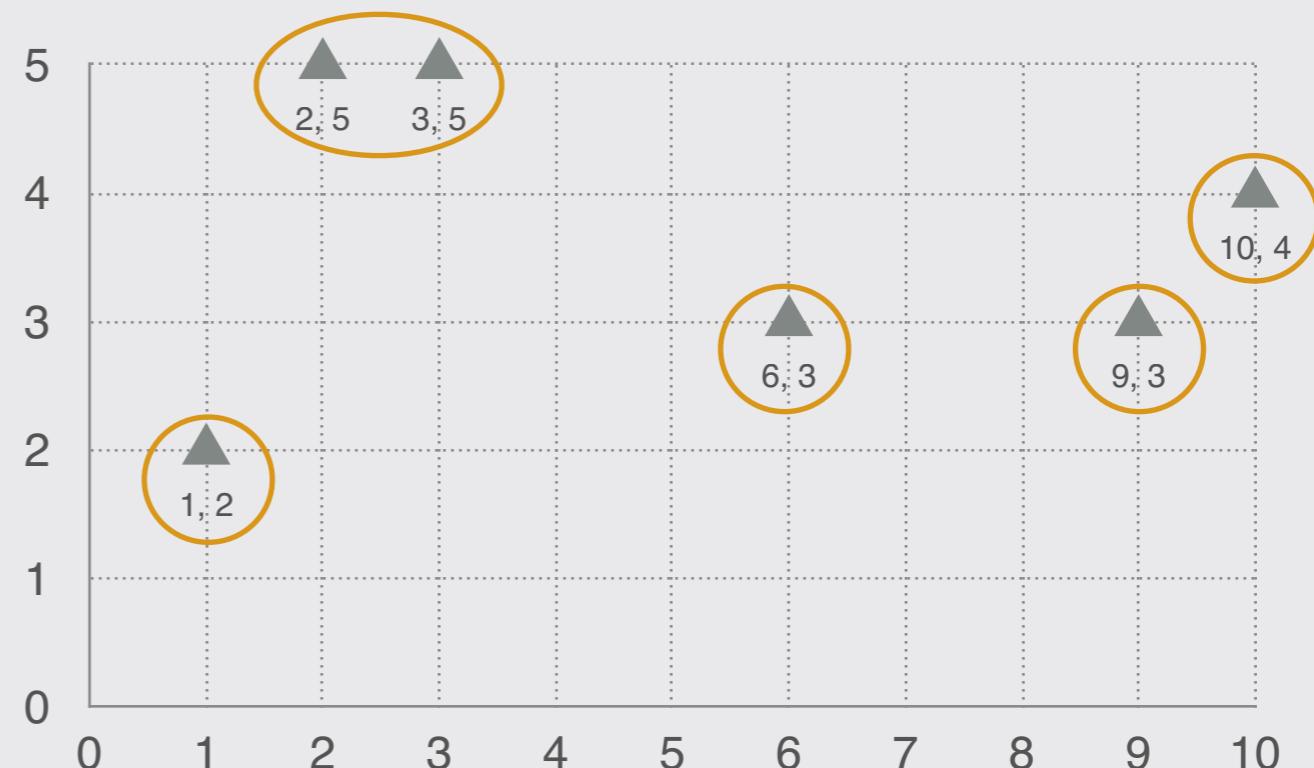
Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

A questo punto, scegliamo la distanza minore e uniamo i due punti in un nuovo cluster.

Una strategia veloce è quella di calcolare un **punto medio**, similmente a come si fa con il centroide nel k-means.



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

A questo punto, scegliamo la distanza minore e uniamo i due punti in un nuovo cluster.

Una strategia veloce è quella di calcolare un **punto medio**, similmente a come si fa con il centroide nel k-means.

Il punto medio sarà dato da:

$$c_2 = \left(\frac{2+3}{2} + \frac{5+5}{2} \right) = (2.5, 5)$$



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (B,C) consideriamo il punto medio (2.5, 5).

	A	(B,C)	D	E	F
A	\		5.09	8.06	9.21
(B,C)	\	\			
D	\	\	\	3	4.12
E	\	\	\	\	1.41
F	\	\	\	\	\

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (B,C) consideriamo il punto medio (2.5, 5).

	A	(B,C)	D	E	F
A	\	3.35	5.09	8.06	9.21
(B,C)	\	\			
D	\	\	\	3	4.12
E	\	\	\	\	1.41
F	\	\	\	\	\

$$d(A, c_2) = \sqrt{(2.5 - 1)^2 + (5 - 2)^2} = 3.35$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (B,C) consideriamo il punto medio (2.5, 5).

	A	(B,C)	D	E	F
A	\	3.35	5.09	8.06	9.21
(B,C)	\	\	4.03		
D	\	\	\	3	4.12
E	\	\	\	\	1.41
F	\	\	\	\	\

$$d(A, c_2) = \sqrt{(2.5 - 1)^2 + (5 - 2)^2} = 3.35$$

$$d(c_2, D) = \sqrt{(6 - 2.5)^2 + (3 - 5)^2} = 4.03$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (B,C) consideriamo il punto medio (2.5, 5).

	A	(B,C)	D	E	F
A	\	3.35	5.09	8.06	9.21
(B,C)	\	\	4.03	6.8	
D	\	\	\	3	4.12
E	\	\	\	\	1.41
F	\	\	\	\	\

$$d(A, c_2) = \sqrt{(2.5 - 1)^2 + (5 - 2)^2} = 3.35$$

$$d(c_2, D) = \sqrt{(6 - 2.5)^2 + (3 - 5)^2} = 4.03$$

$$d(c_2, E) = \sqrt{(9 - 2.5)^2 + (3 - 5)^2} = 6.8$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (B,C) consideriamo il punto medio (2.5, 5).

	A	(B,C)	D	E	F
A	\	3.35	5.09	8.06	9.21
(B,C)	\	\	4.03	6.8	7.56
D	\	\	\	3	4.12
E	\	\	\	\	1.41
F	\	\	\	\	\

$$d(A, c_2) = \sqrt{(2.5 - 1)^2 + (5 - 2)^2} = 3.35$$

$$d(c_2, D) = \sqrt{(6 - 2.5)^2 + (3 - 5)^2} = 4.03$$

$$d(c_2, E) = \sqrt{(9 - 2.5)^2 + (3 - 5)^2} = 6.8$$

$$d(c_2, F) = \sqrt{(10 - 2.5)^2 + (4 - 5)^2} = 7.56$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (B,C) consideriamo il punto medio (2.5, 5).

	A	(B,C)	D	E	F
A	\	3.35	5.09	8.06	9.21
(B,C)	\	\	4.03	6.8	7.56
D	\	\	\	3	4.12
E	\	\	\	\	1.41
F	\	\	\	\	\

$$d(A, c_2) = \sqrt{(2.5 - 1)^2 + (5 - 2)^2} = 3.35$$

$$d(c_2, D) = \sqrt{(6 - 2.5)^2 + (3 - 5)^2} = 4.03$$

$$d(c_2, E) = \sqrt{(9 - 2.5)^2 + (3 - 5)^2} = 6.8$$

$$d(c_2, F) = \sqrt{(10 - 2.5)^2 + (4 - 5)^2} = 7.56$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

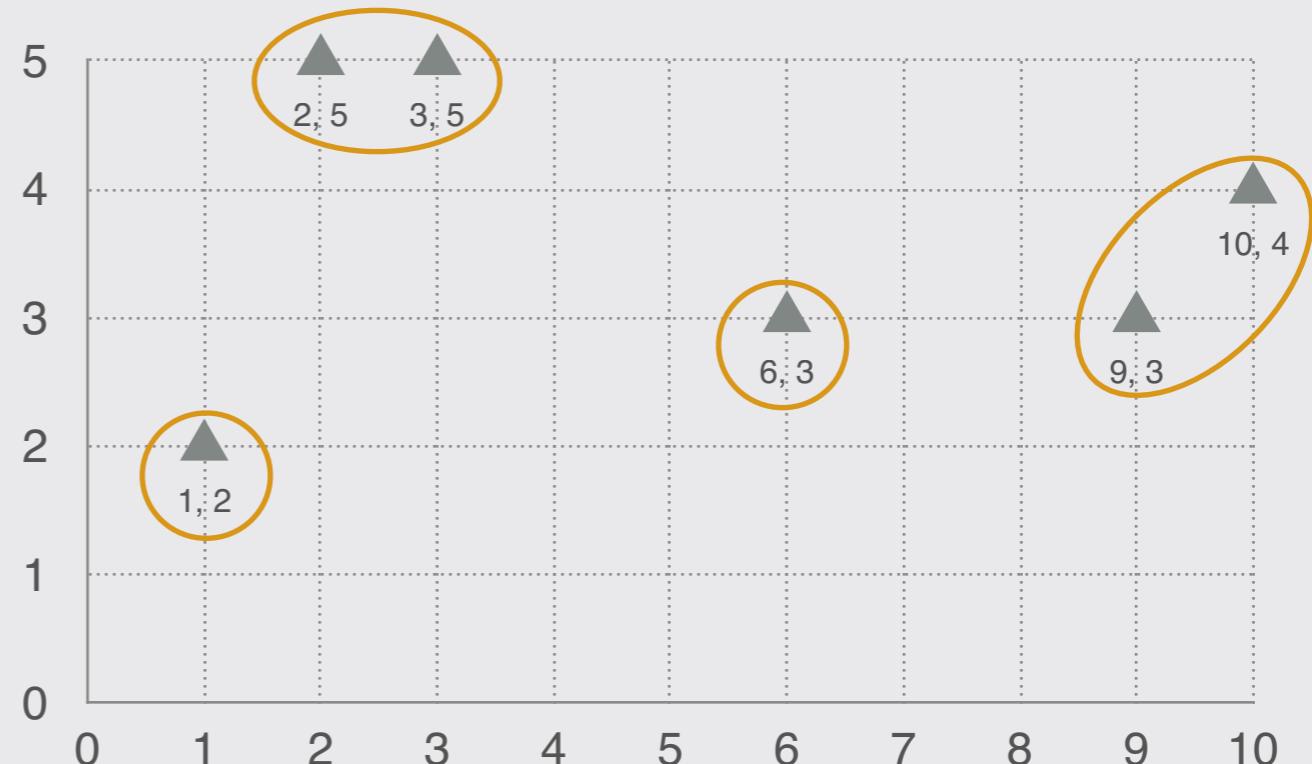
Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Per il prossimo raggruppamento, scegliamo di nuovo il minimo.

- c1 = {A(1, 2)}
- c2 = {B(2, 5), C(3,5)}
- c3 = {D(6, 3)}
- c4 = {E(9, 3), F(10, 4)}



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Per il prossimo raggruppamento, scegliamo di nuovo il minimo.

$$c_1 = \{A(1, 2)\}$$

$$c_2 = \{B(2, 5), C(3, 5)\}$$

$$c_3 = \{D(6, 3)\}$$

$$c_4 = \{E(9, 3), F(10, 4)\}$$

Calcoliamo il nuovo punto medio per c4:

$$c_4 = \left(\frac{9 + 10}{2} + \frac{3 + 4}{2} \right) = (9.5, 3.5)$$



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (E,F) consideriamo il punto medio (9.5, 3.5).

	A	(B,C)	D	(E,F)
A	\	3.35	5.09	
(B,C)	\	\	4.03	
D	\	\	\	
(E,F)	\	\	\	\

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (E,F) consideriamo il punto medio (9.5, 3.5).

	A	(B,C)	D	(E,F)
A	\	3.35	5.09	8.86
(B,C)	\	\	4.03	
D	\	\	\	
(E,F)	\	\	\	\

$$d(A, c_4) = \sqrt{(9.5 - 1)^2 + (3.5 - 1)^2} = 8.86$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (E,F) consideriamo il punto medio (9.5, 3.5).

	A	(B,C)	D	(E,F)
A	\	3.35	5.09	8.86
(B,C)	\	\	4.03	7.15
D	\	\	\	
(E,F)	\	\	\	\

$$d(A, c_4) = \sqrt{(9.5 - 1)^2 + (3.5 - 1)^2} = 8.86$$

$$d(c_2, c_4) = \sqrt{(9.5 - 2.5)^2 + (3.5 - 5)^2} = 7.15$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (E,F) consideriamo il punto medio (9.5, 3.5).

	A	(B,C)	D	(E,F)
A	\	3.35	5.09	8.86
(B,C)	\	\	4.03	7.15
D	\	\	\	3.53
(E,F)	\	\	\	\

$$d(A, c_4) = \sqrt{(9.5 - 1)^2 + (3.5 - 1)^2} = 8.86$$

$$d(c_2, c_4) = \sqrt{(9.5 - 2.5)^2 + (3.5 - 5)^2} = 7.15$$

$$d(D, c_4) = \sqrt{(9.5 - 6)^2 + (3.5 - 3)^2} = 3.53$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Scegliamo il minimo per la prossima aggregazione.

	A	(B,C)	D	(E,F)
A	\	3.35	5.09	8.86
(B,C)	\	\	4.03	7.15
D	\	\	\	3.53
(E,F)	\	\	\	\

$$c1 = \{A(1, 2), B(2, 5), C(3, 5)\}$$

$$c2 = \{D(6, 3)\}$$

$$c3 = \{E(9, 3), F(10, 4)\}$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

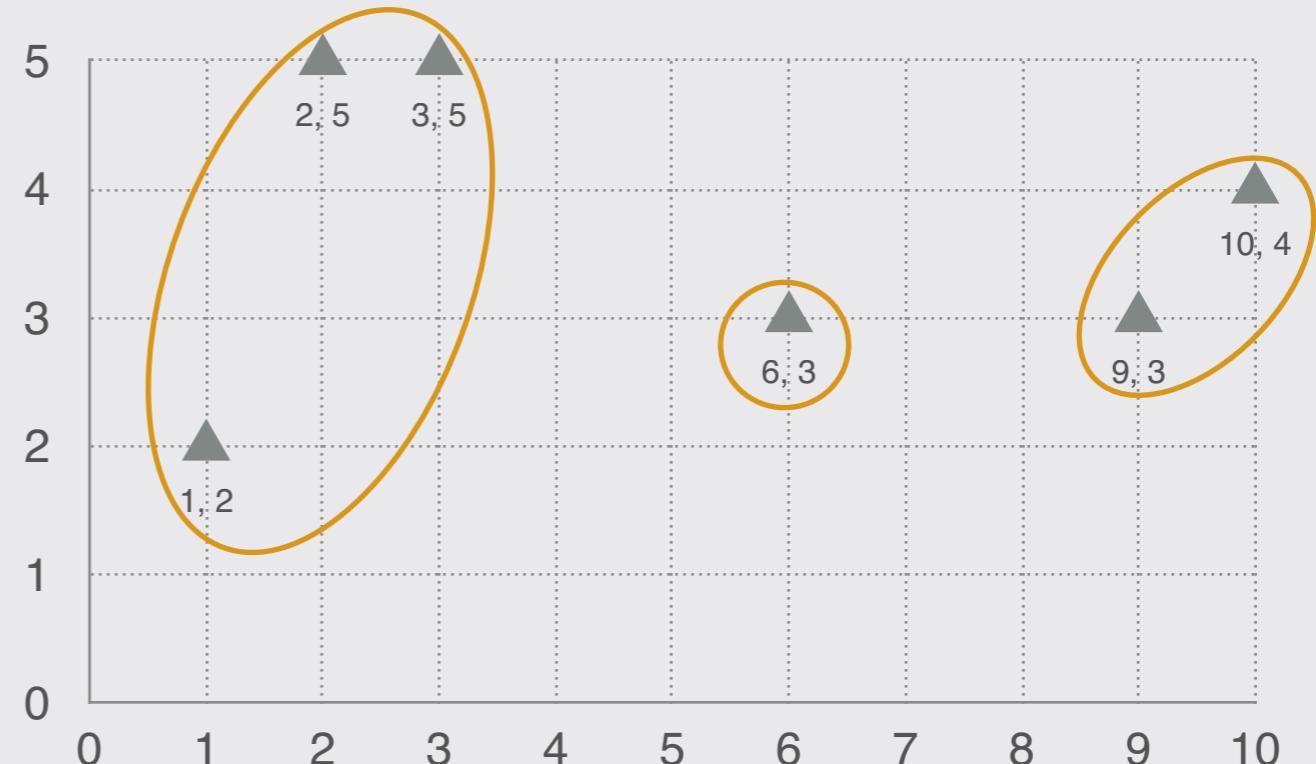
Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

$$c1 = \{A(1, 2), B(2, 5), C(3, 5)\}$$

$$c2 = \{D(6, 3)\}$$

$$c3 = \{E(9, 3), F(10, 4)\}$$



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

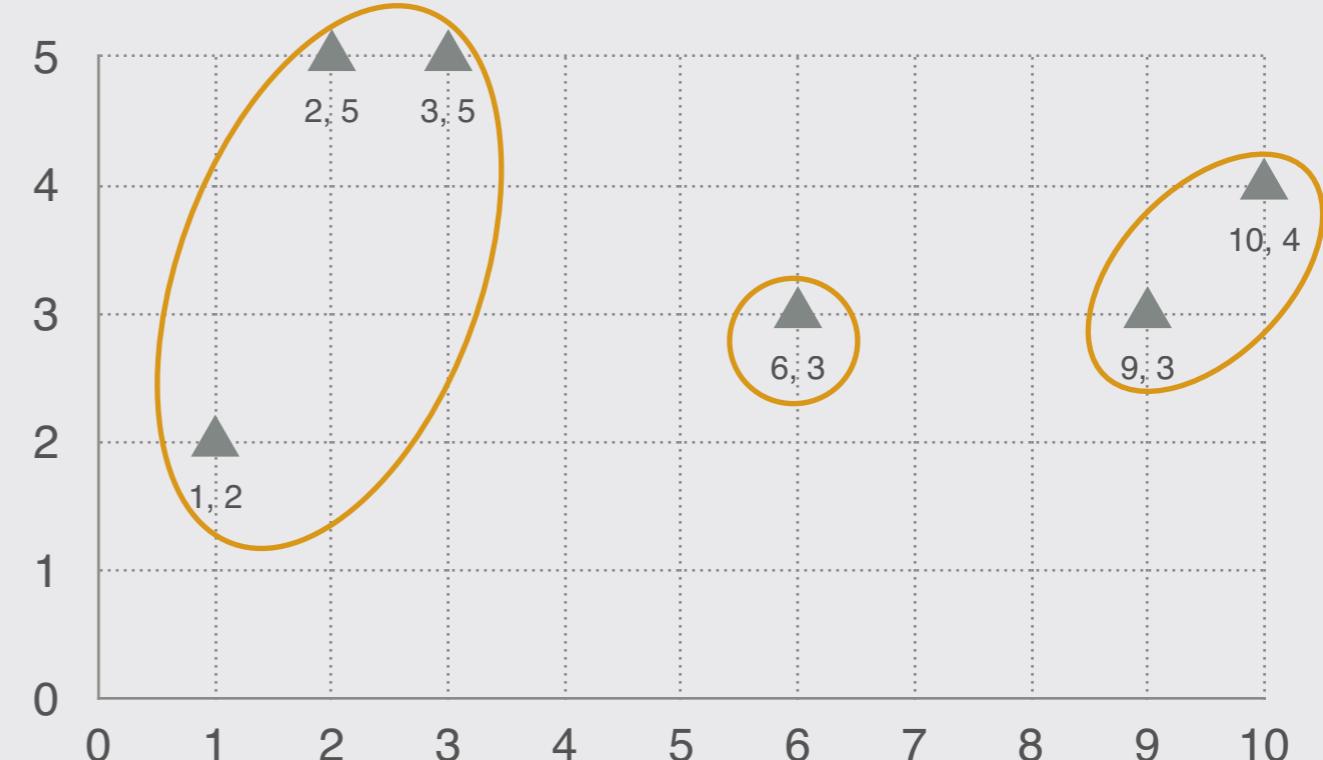
$$c1 = \{A(1, 2), B(2, 5), C(3, 5)\}$$

$$c2 = \{D(6, 3)\}$$

$$c3 = \{E(9, 3), F(10, 4)\}$$

Calcoliamo il nuovo punto medio per c3:

$$c_1 = \left(\frac{1 + 2 + 3}{3} + \frac{2 + 5 + 5}{3} \right) = (2, 4)$$



Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (D,E,F) consideriamo il punto medio (2, 4).

	(A,B,C)	D	(E,F)
(A,B,C)	\		
D	\	\	3.53
(E,F)	\	\	\

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (D,E,F) consideriamo il punto medio (2, 4).

	(A,B,C)	D	(E,F)
(A,B,C)	\	4.12	
D	\	\	3.53
(E,F)	\	\	\

$$d(c_1, D) = \sqrt{(6 - 2)^2 + (3 - 4)^2} = 4.12$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Ricalcoliamo le distanze con il nuovo cluster, ricordando che per (D,E,F) consideriamo il punto medio (2, 4).

	(A,B,C)	D	(E,F)
(A,B,C)	\	4.12	7.51
D	\	\	3.53
(E,F)	\	\	\

$$d(c_1, D) = \sqrt{(6 - 2)^2 + (3 - 4)^2} = 4.1$$

$$d(c_1, c_4) = \sqrt{(9.5 - 2)^2 + (3.5 - 4)^2} = 7.51$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Scegliamo il minimo per la prossima aggregazione.

	(A,B,C)	D	(E,F)
(A,B,C)	\	4.12	7.51
D	\	\	3.53
(E,F)	\	\	\

$$c1 = \{A(1, 2), B(2, 5), C(3, 5)\}$$

$$c2 = \{D(6, 3), E(9, 3), F(10, 4)\}$$

Clustering - Esercizio 2

Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

$$c1 = \{A(1, 2), B(2, 5), C(3, 5)\}$$

$$c2 = \{D(6, 3), E(9, 3), F(10, 4)\}$$

Abbiamo raggiunto due cluster!



Clustering - Esercizio 2

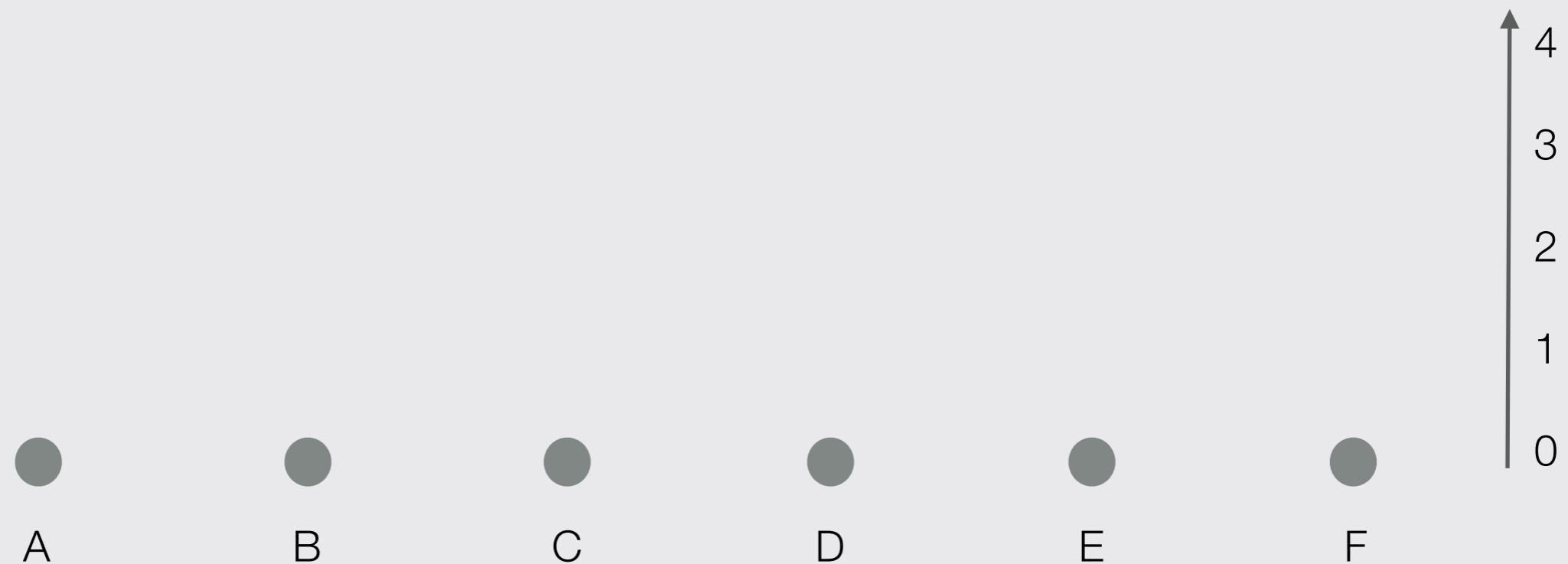
Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Costruiamo il dendrogramma che schematizza le operazioni che abbiamo fatto.



Clustering - Esercizio 2

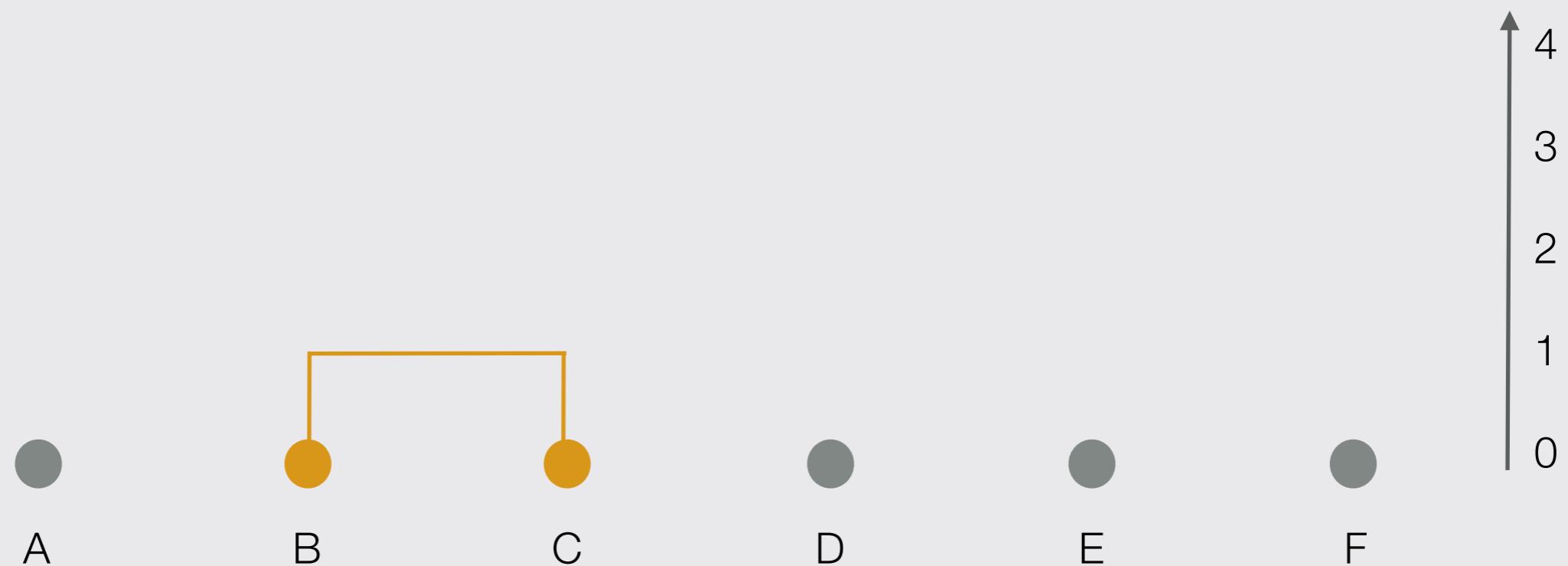
Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Costruiamo il dendrogramma che schematizza le operazioni che abbiamo fatto.



Clustering - Esercizio 2

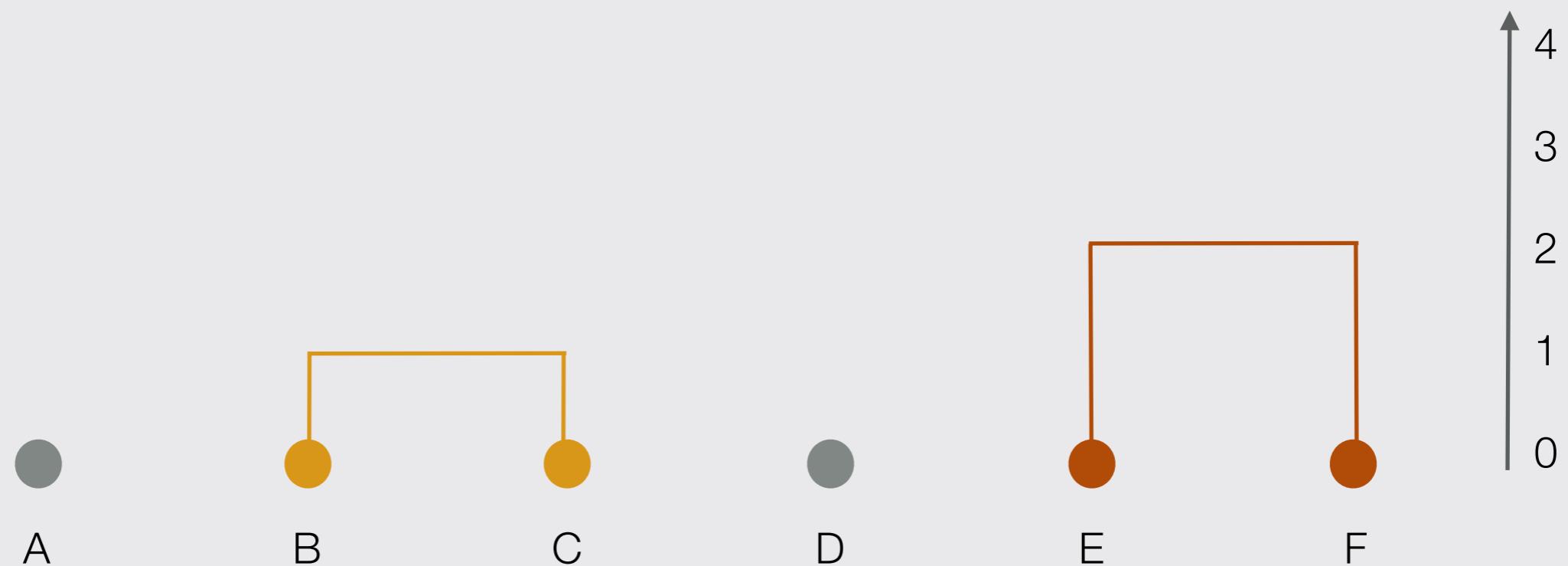
Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Costruiamo il dendrogramma che schematizza le operazioni che abbiamo fatto.



Clustering - Esercizio 2

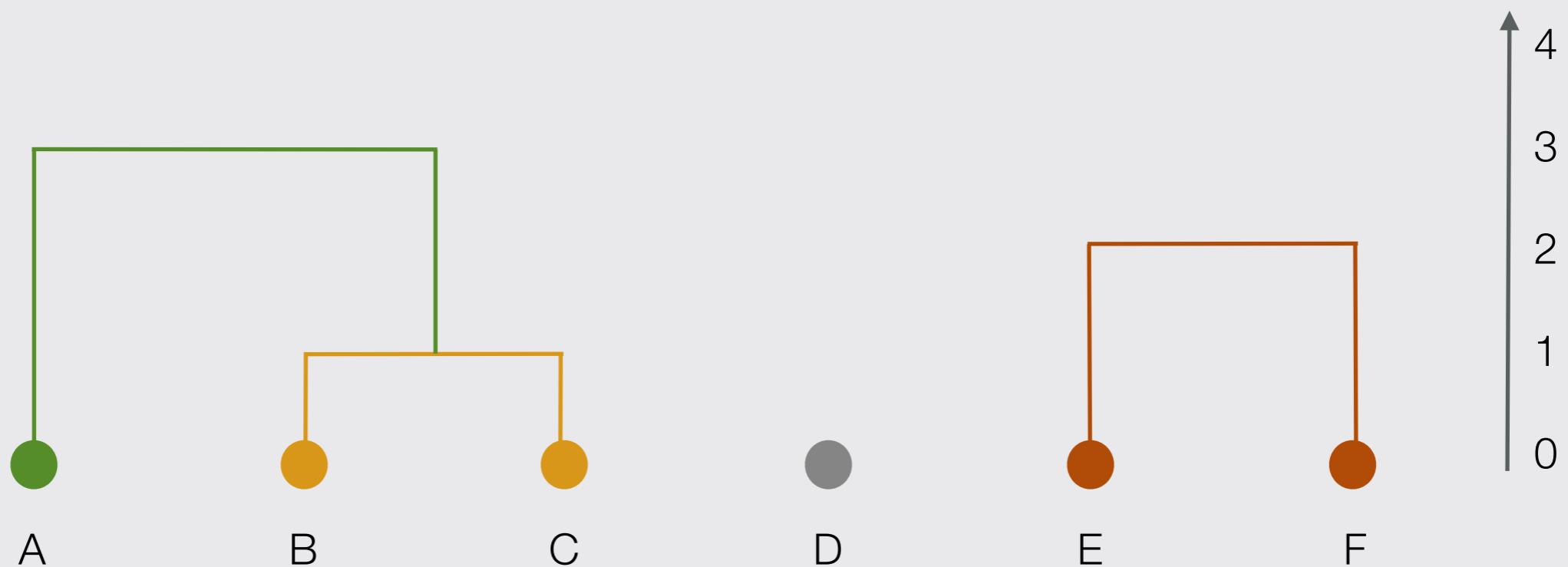
Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Costruiamo il dendrogramma che schematizza le operazioni che abbiamo fatto.



Clustering - Esercizio 2

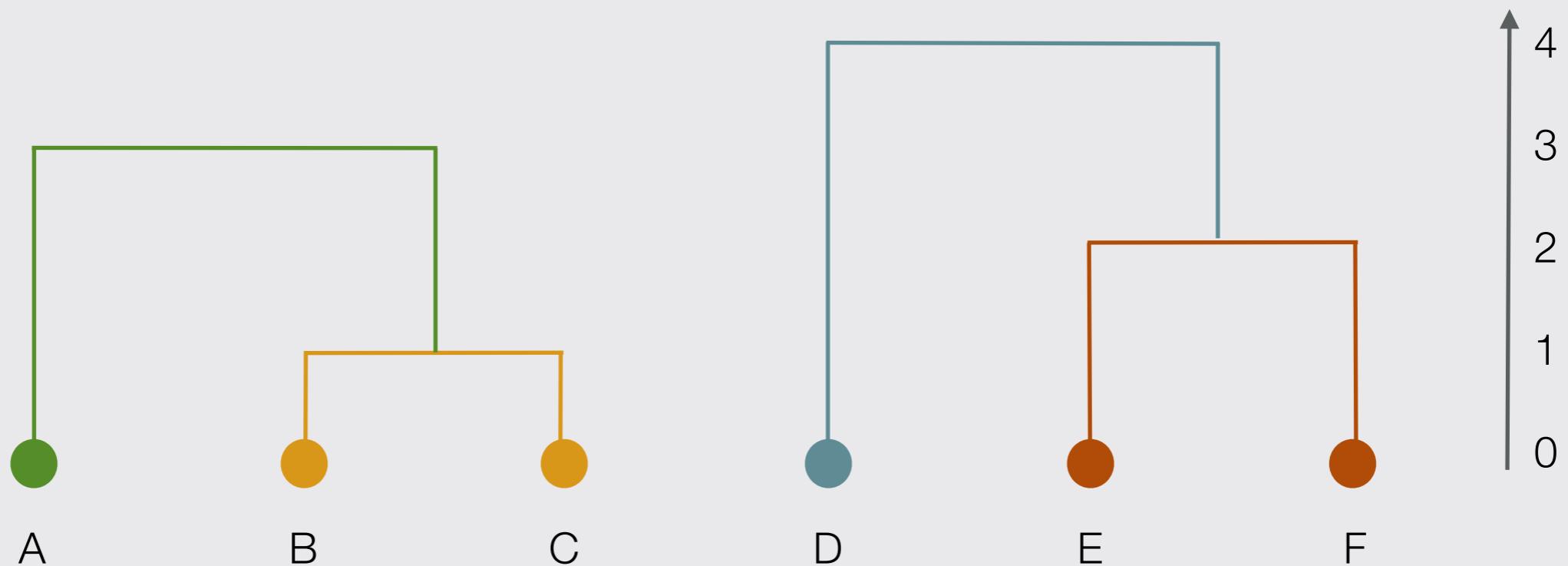
Suddividere i seguenti punti in 2 cluster, utilizzando un metodo di clustering gerarchico agglomerativo.

Utilizzare la distanza euclidea per valutare la distanza tra due cluster.

Mostrare il dendrogramma risultante.

{A(1, 2) B(2, 5) C(3, 5) D(6, 3) E(9, 3) F(10, 4)}

Costruiamo il dendrogramma che schematizza le operazioni che abbiamo fatto.





Dubbi?

Domande?

Perplessità?





UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA

Laurea triennale in Informatica



Viviana Pentangelo
✉ tutoratofia@gmail.com

Fondamenti di Intelligenza Artificiale

Help Teaching - Esercitazione 3

