Arbres de Décision

INTUITION

Principe : réfléchir dimension par dimension, probablement le plus proche d'une décision humaine.
Les arbres s'utilisent dans des problèmes de régression et de classification.
Ils acceptent les données de types différents : qualitatives/quantitatives, discrètes/continues.
Ils sont très facilement **interprétables**.

RAPPEL: VARIABLES DISCRÈTES

On dit qu'une variable est **discrète** lorsqu'elle prend un nombre dénombrable de valeurs. (Dénombrable = fini.) Par exemple : nombre de pièces d'un appartement (1,2,3,4,5...), nom de l'auteur d'un livre (Victor Hugo, Stendhal, Balzac...), match à domicile (oui/non), etc.

RAPPEL: VARIABLES CONTINUES

On dit qu'une variable est **continue** lorsqu'elle prend un nombre non dénombrable de valeurs (i.e. qu'on ne peut pas compter, même si on a un temps infini). Par exemple : prix d'un appartement (tous les prix possibles entre 0 et l'infini).

Remarque 1 : on peut **discrétiser** une variable continue. Exemple du prix d'un appartement : classe 1 (entre 0 et 500), classe 2 (entre 500 et 700), classe 3 (entre 700 et 800), etc.

PHASE D'APPRENTISSAGE : VARIABLES DISCRÈTES

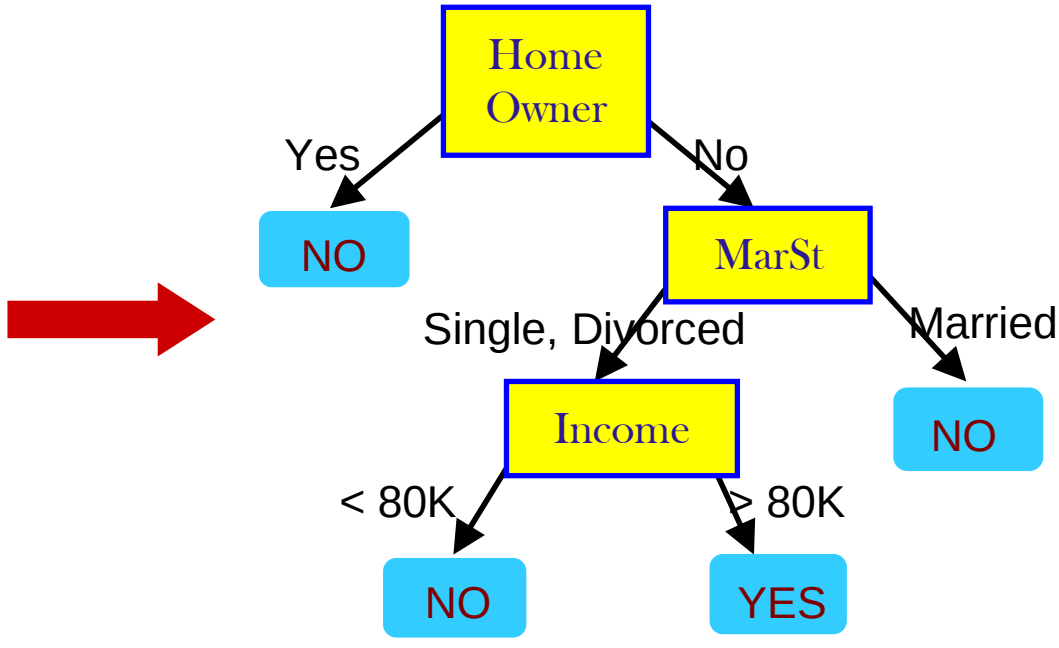
Match à domicile ?	Balance positive ?	Mauvaises conditions climatiques ?	Match précédent gagné ?	Match gagné
V	V	F	F	V
F	F	V	V	V
V	V	V	F	V
V	V	F	V	V
F	V	V	V	F
F	F	V	F	F
V	F	F	V	F
V	F	V	F	F

Source : Cours de François Denis

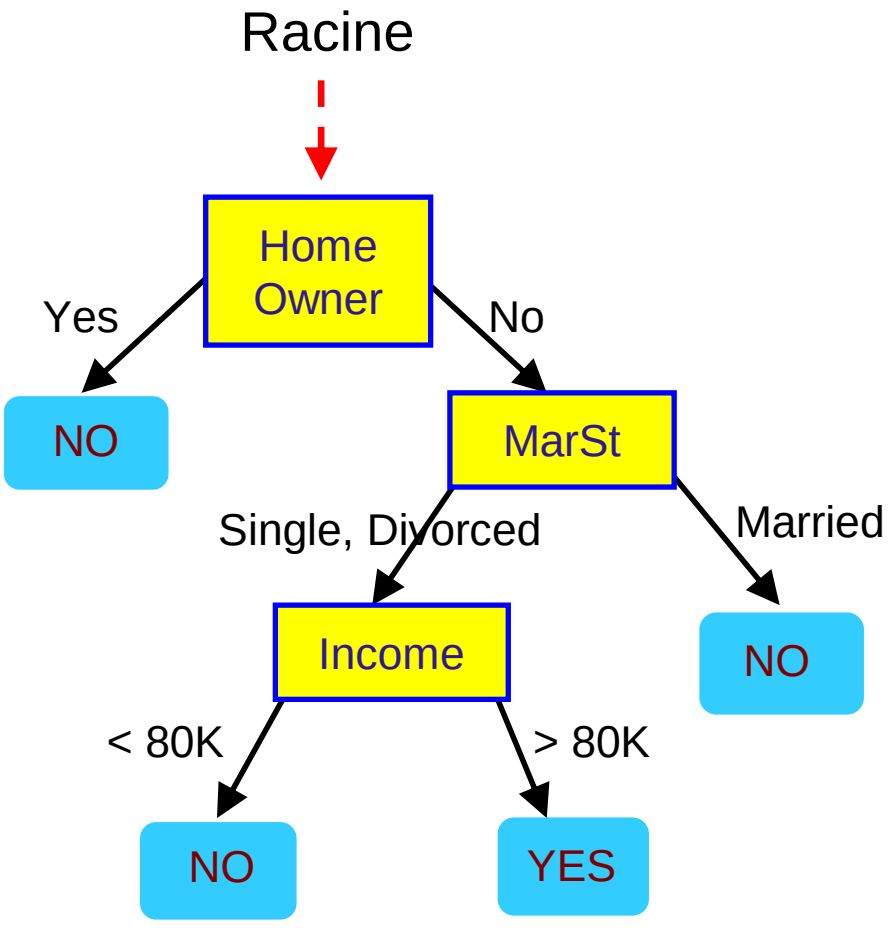
UN EXEMPLE D'ARBRE

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



CLASSIFICATION D'UNE DONNEE DE TEST



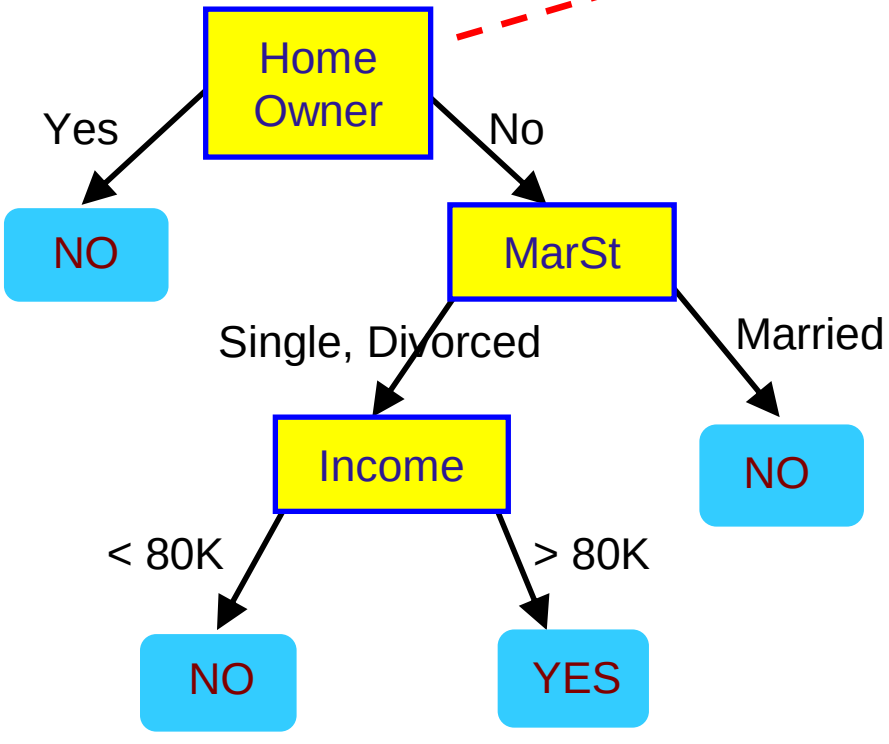
Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	**?**

CLASSIFICATION D'UNE DONNEE DE TEST

Test Data

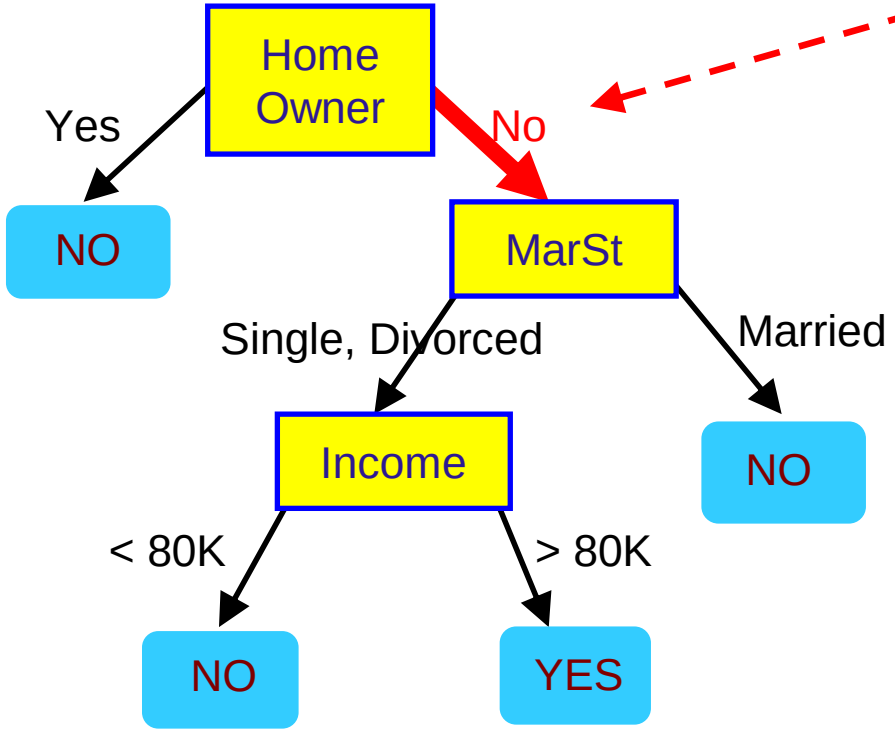
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



CLASSIFICATION D'UNE DONNEE DE TEST

Test Data

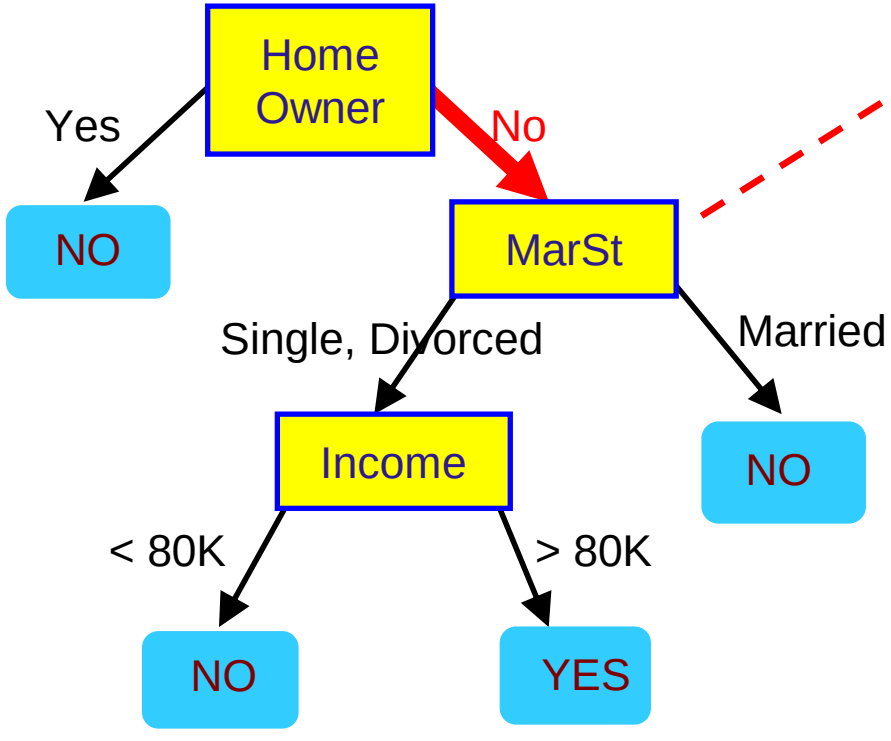
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



CLASSIFICATION D'UNE DONNEE DE TEST

Test Data

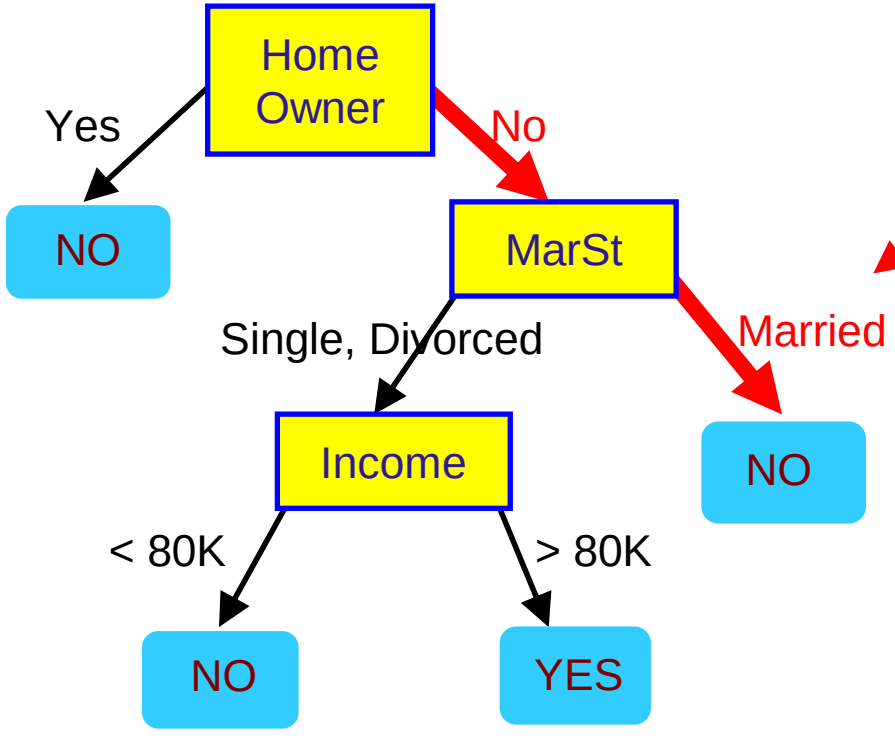
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



CLASSIFICATION D'UNE DONNEE DE TEST

Test Data

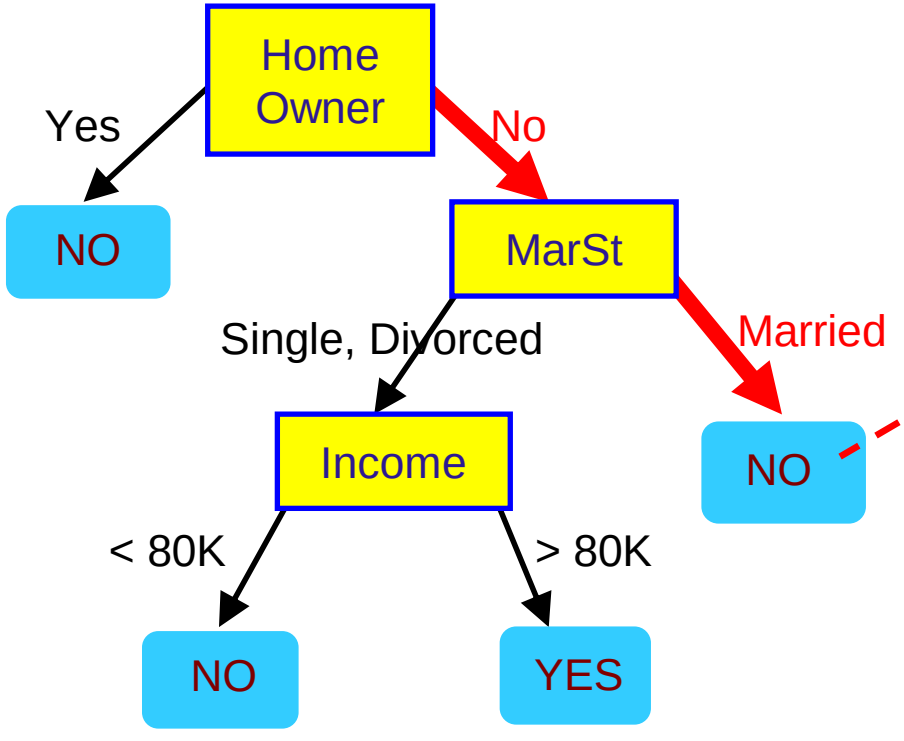
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



CLASSIFICATION D'UNE DONNEE DE TEST

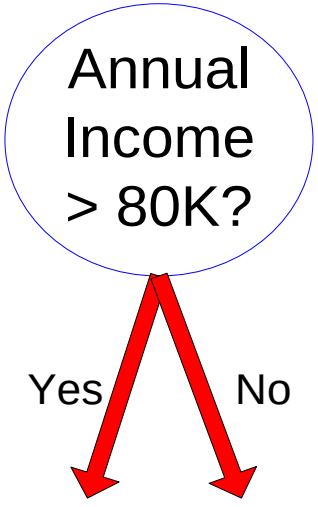
Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

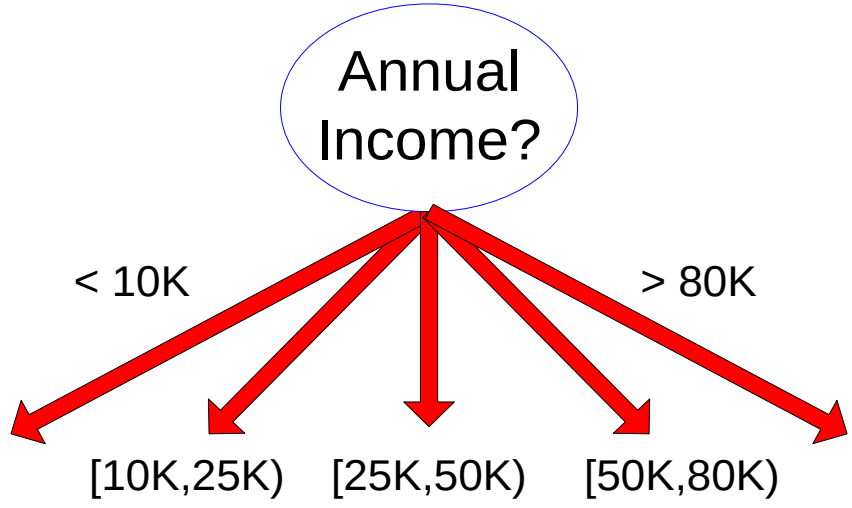


Assignation à la classe "No"

COUPURE BINAIRE OU N-AIRE



(i) Binary split

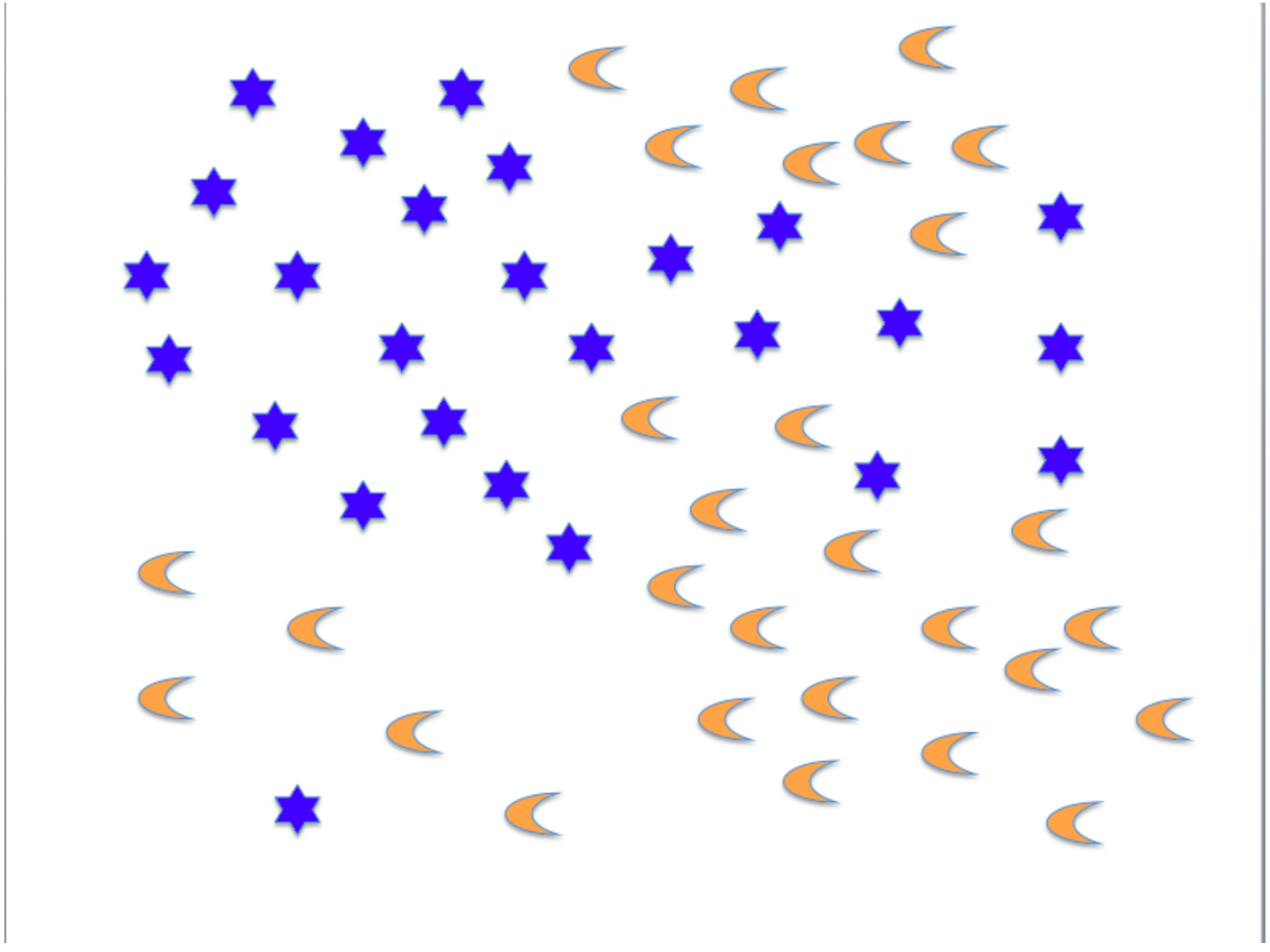


(ii) Multi-way split

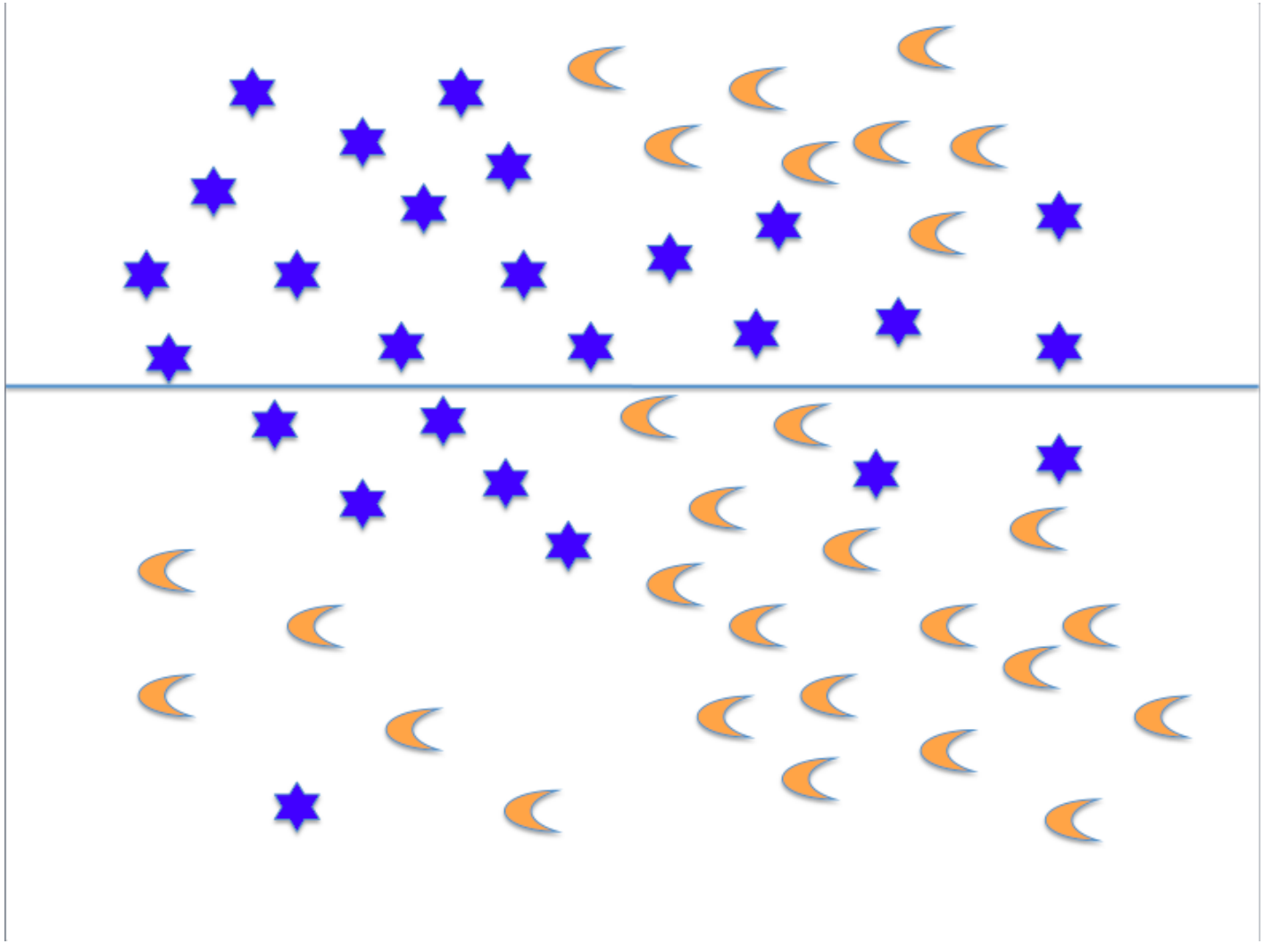
COUPURE POUR ATTRIBUTS CONTINUS

- Peut être traité de plusieurs manières.
  - Discrétiser pour former un attribut catégorique ordinal. Les intervalles peuvent être égaux, ou déterminés en fonction des fréquences.
    - Statique – discrétiser une seule fois au tout début.
    - Dynamique – répétition à chaque nœud.

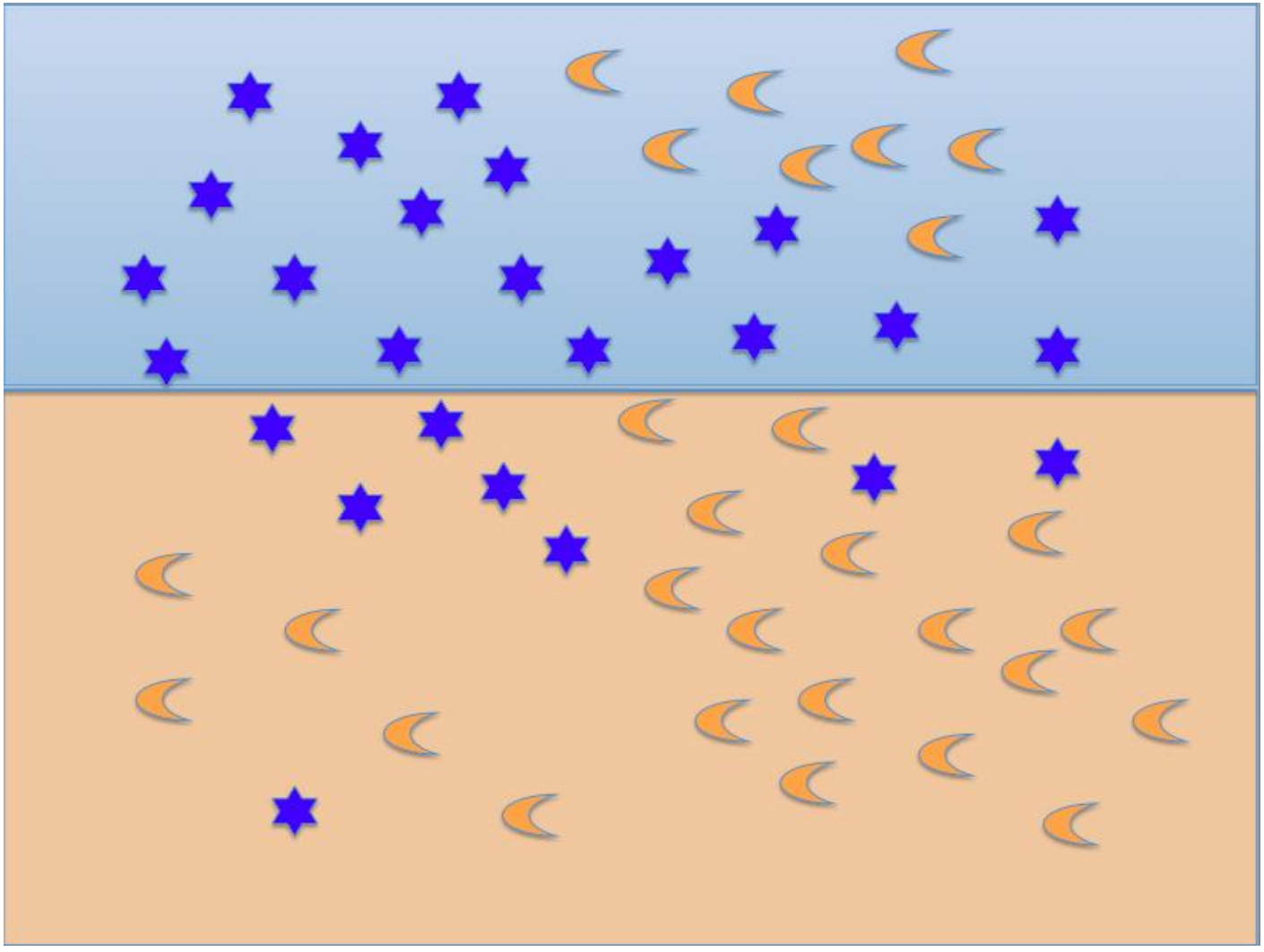
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



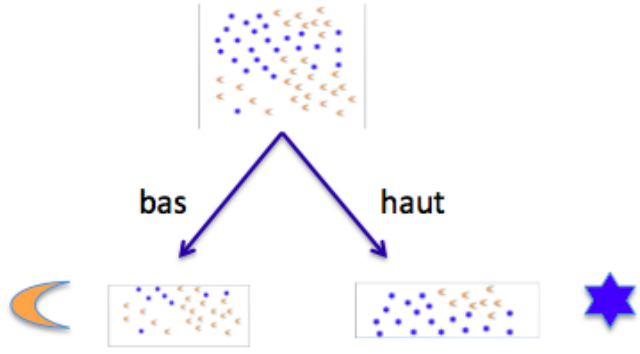
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



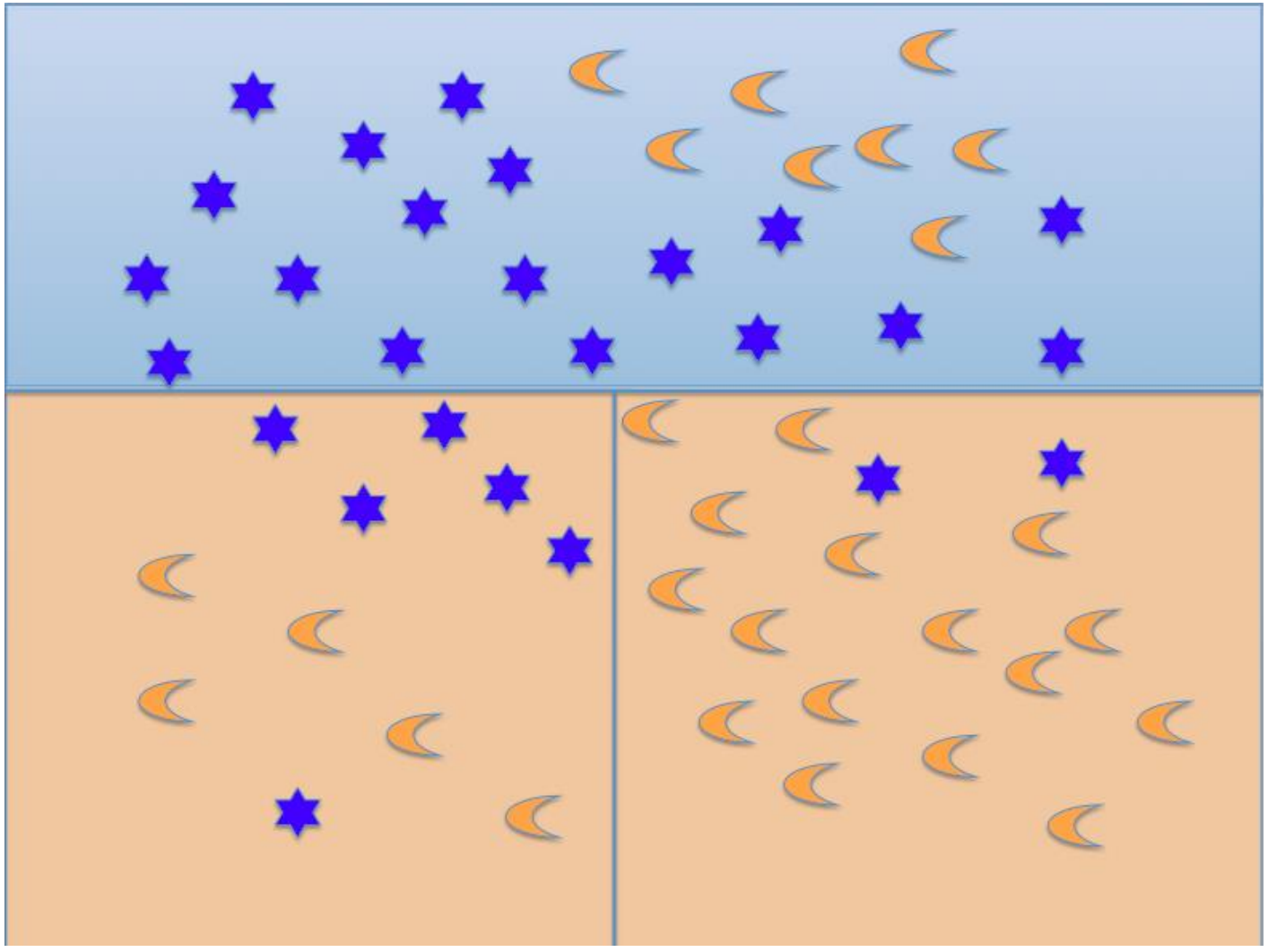
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



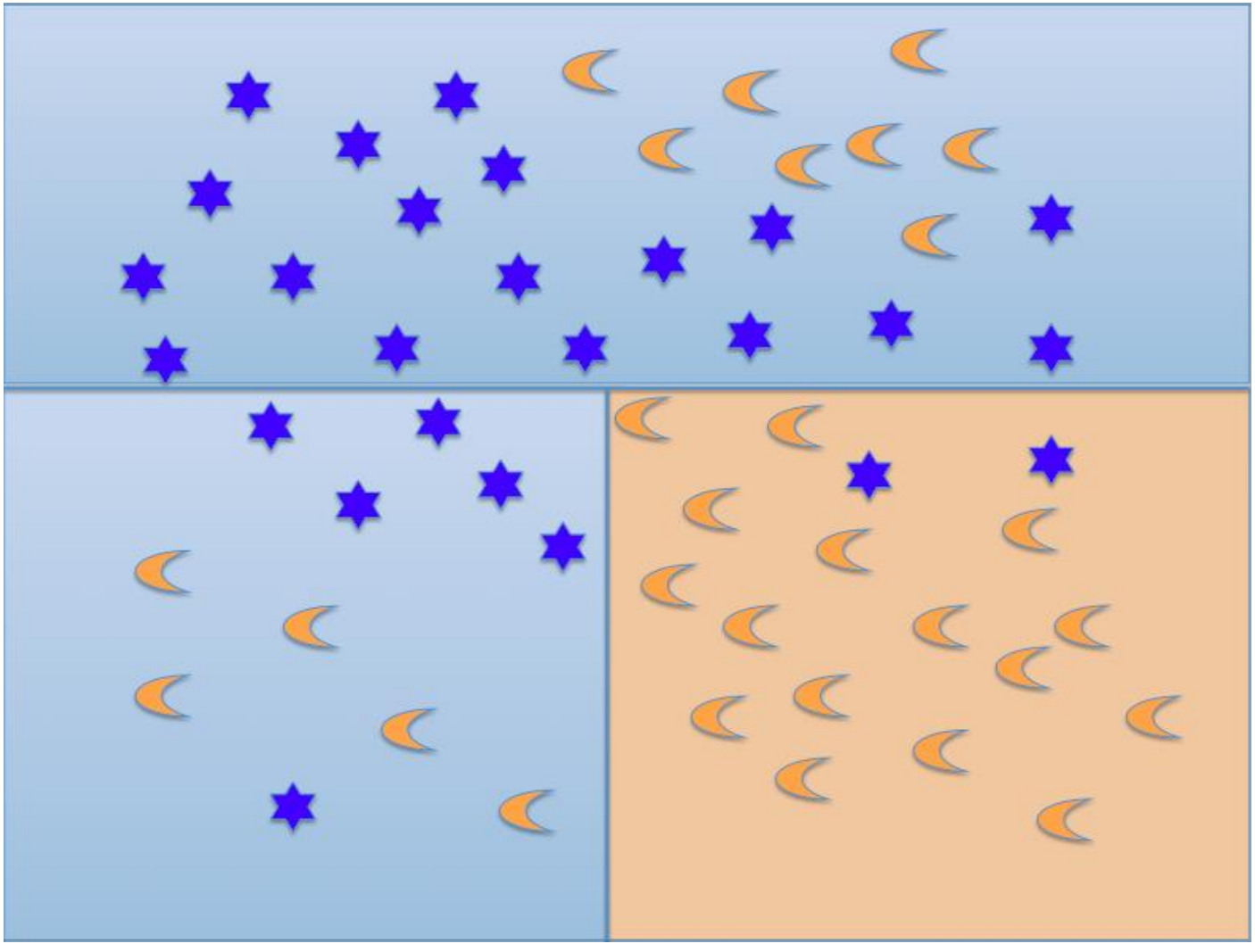
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



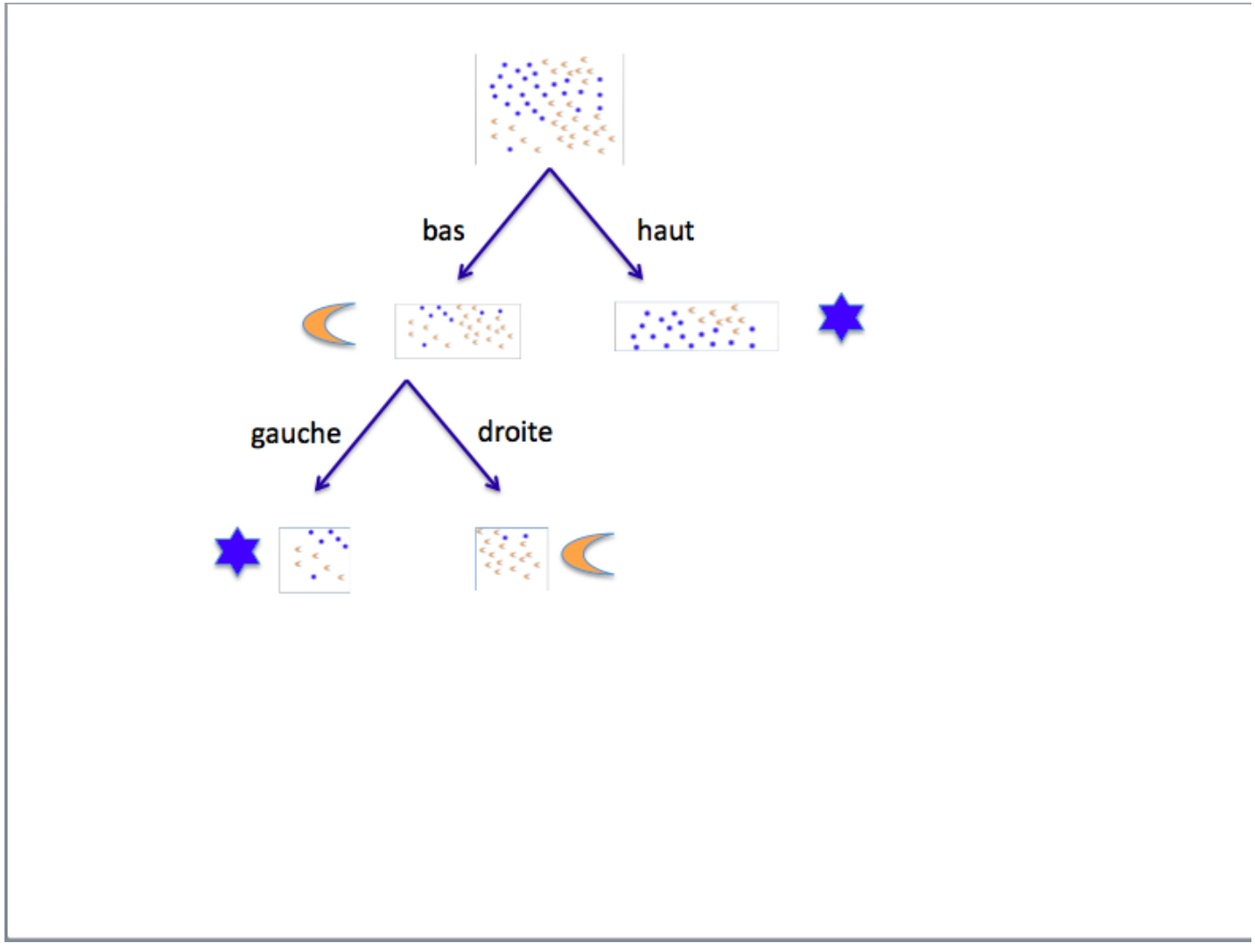
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



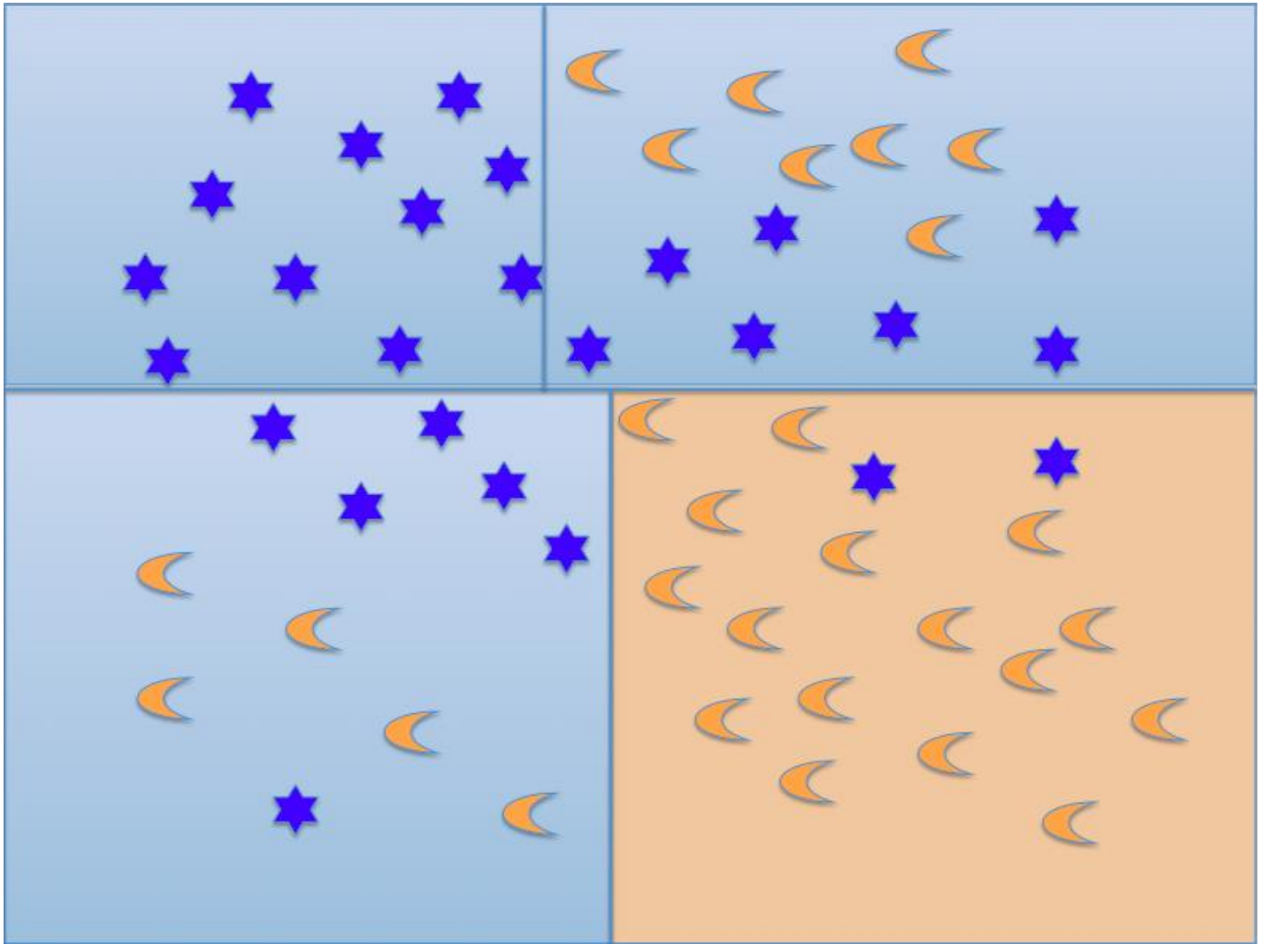
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



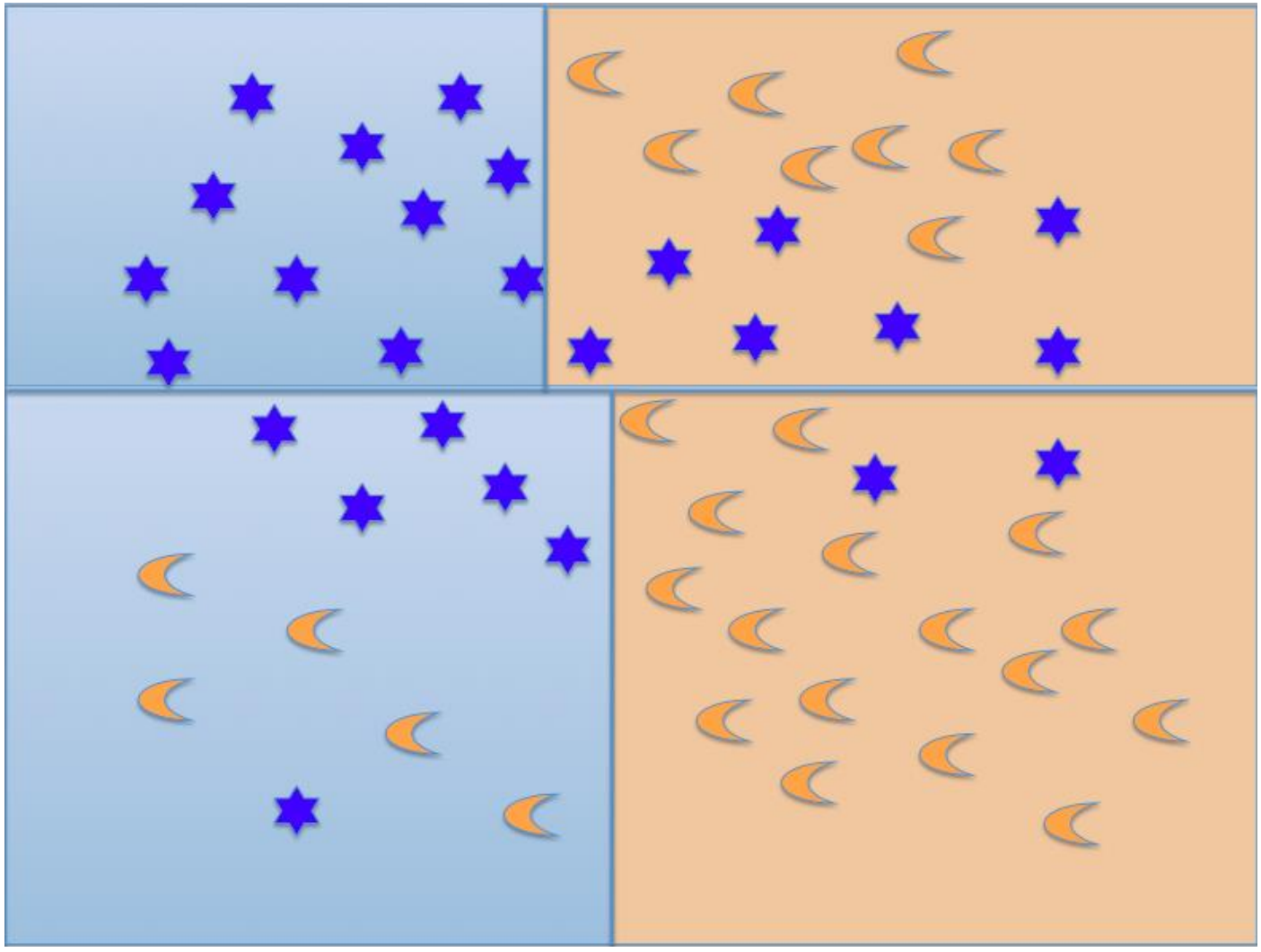
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



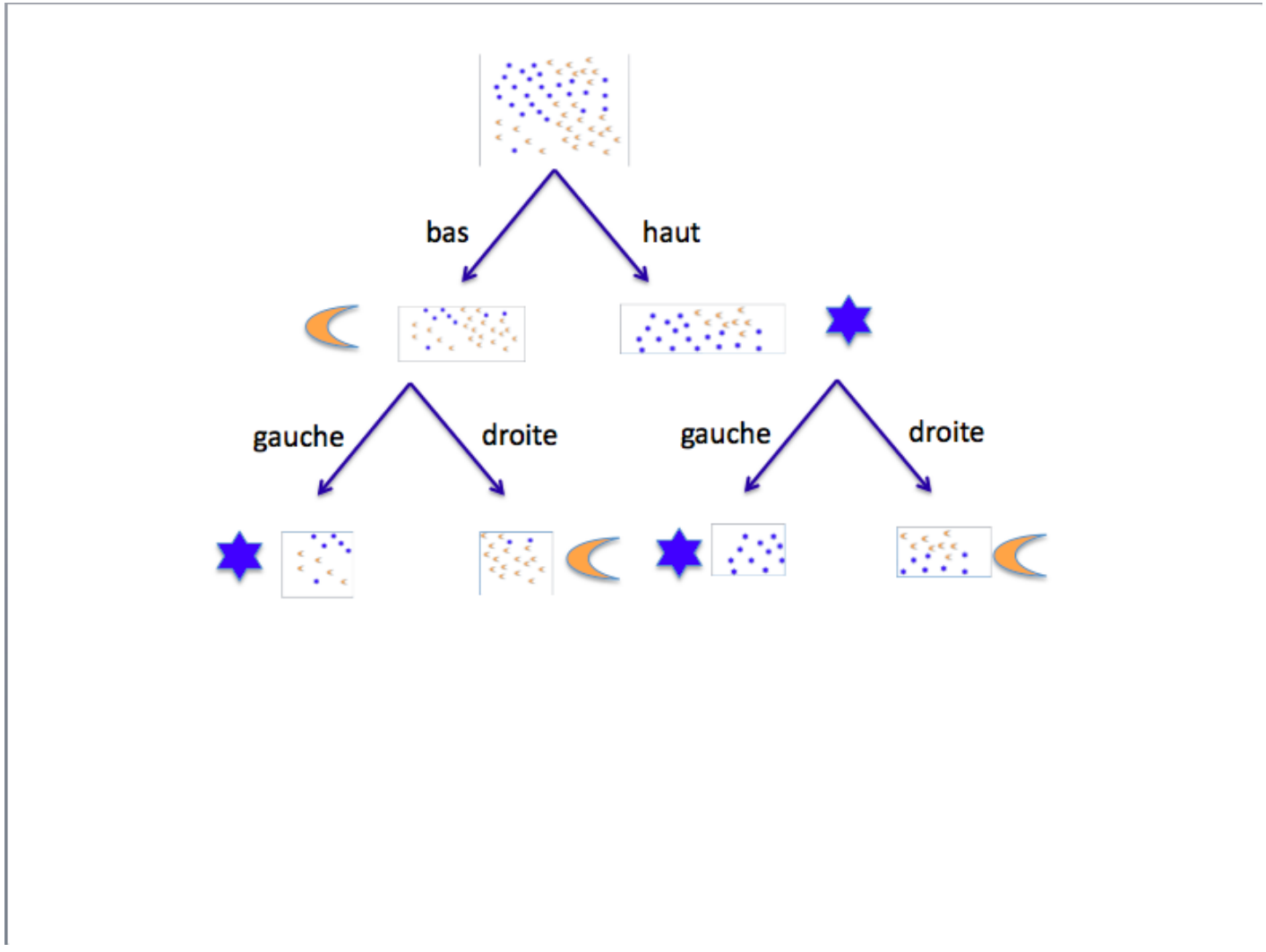
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



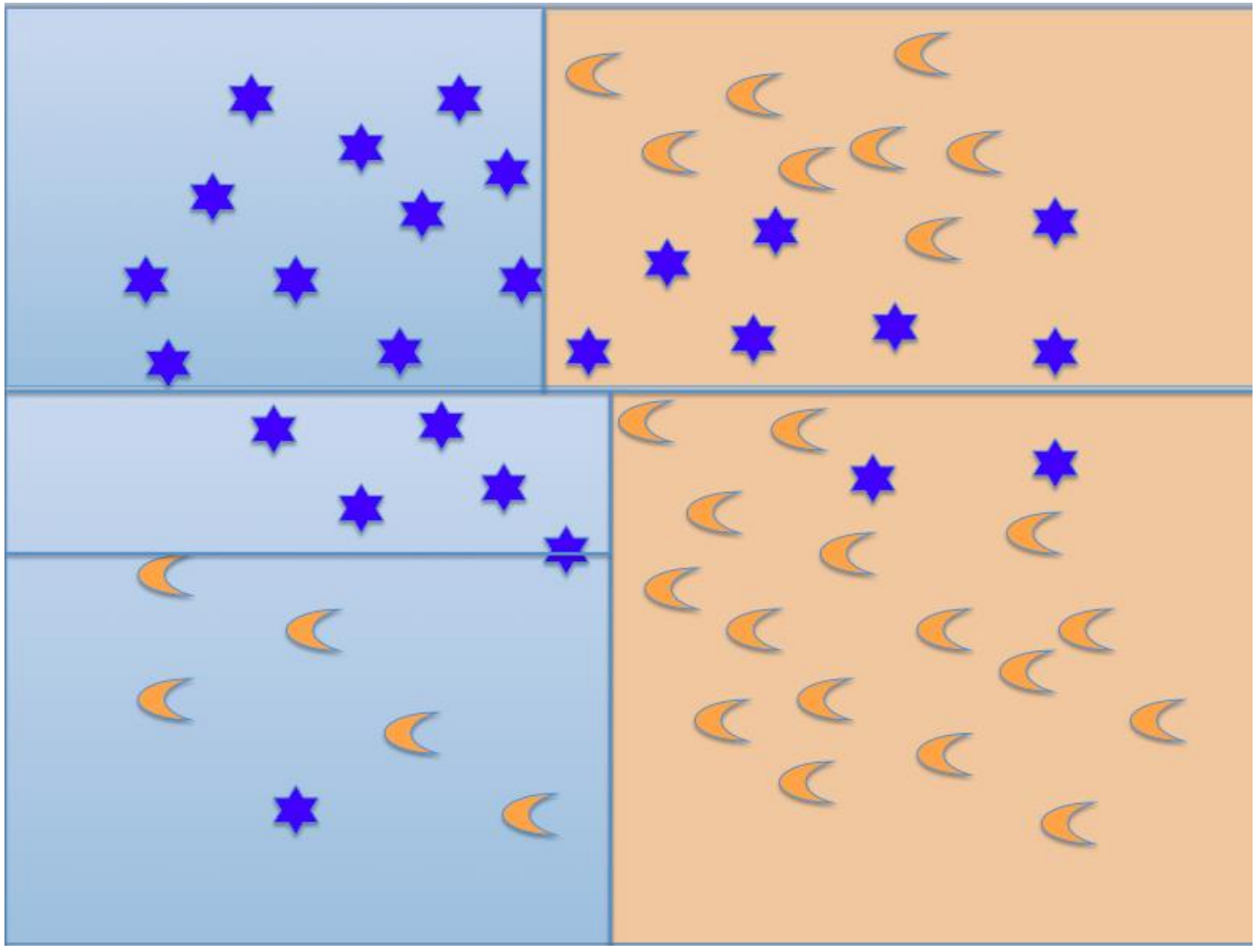
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



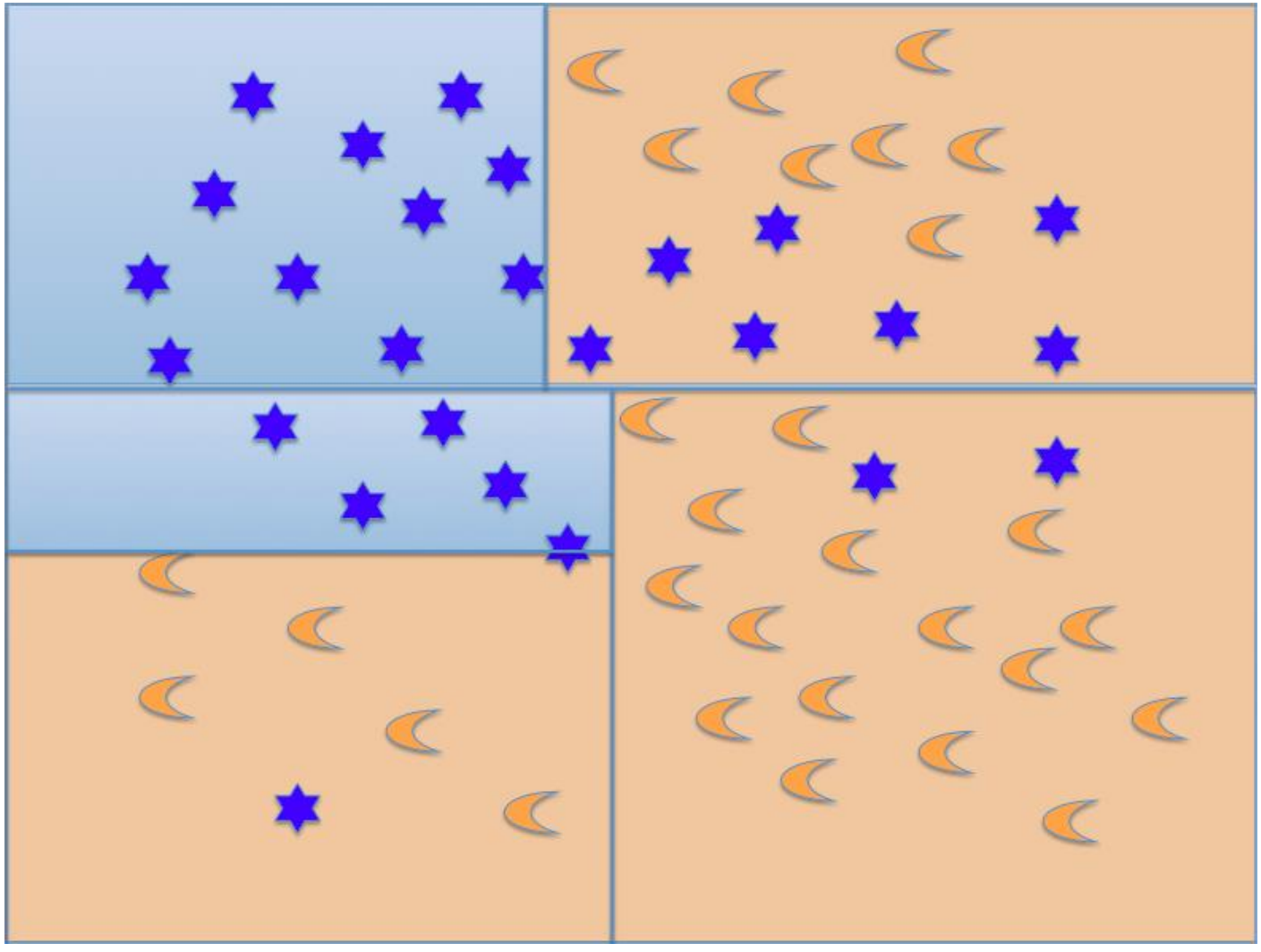
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



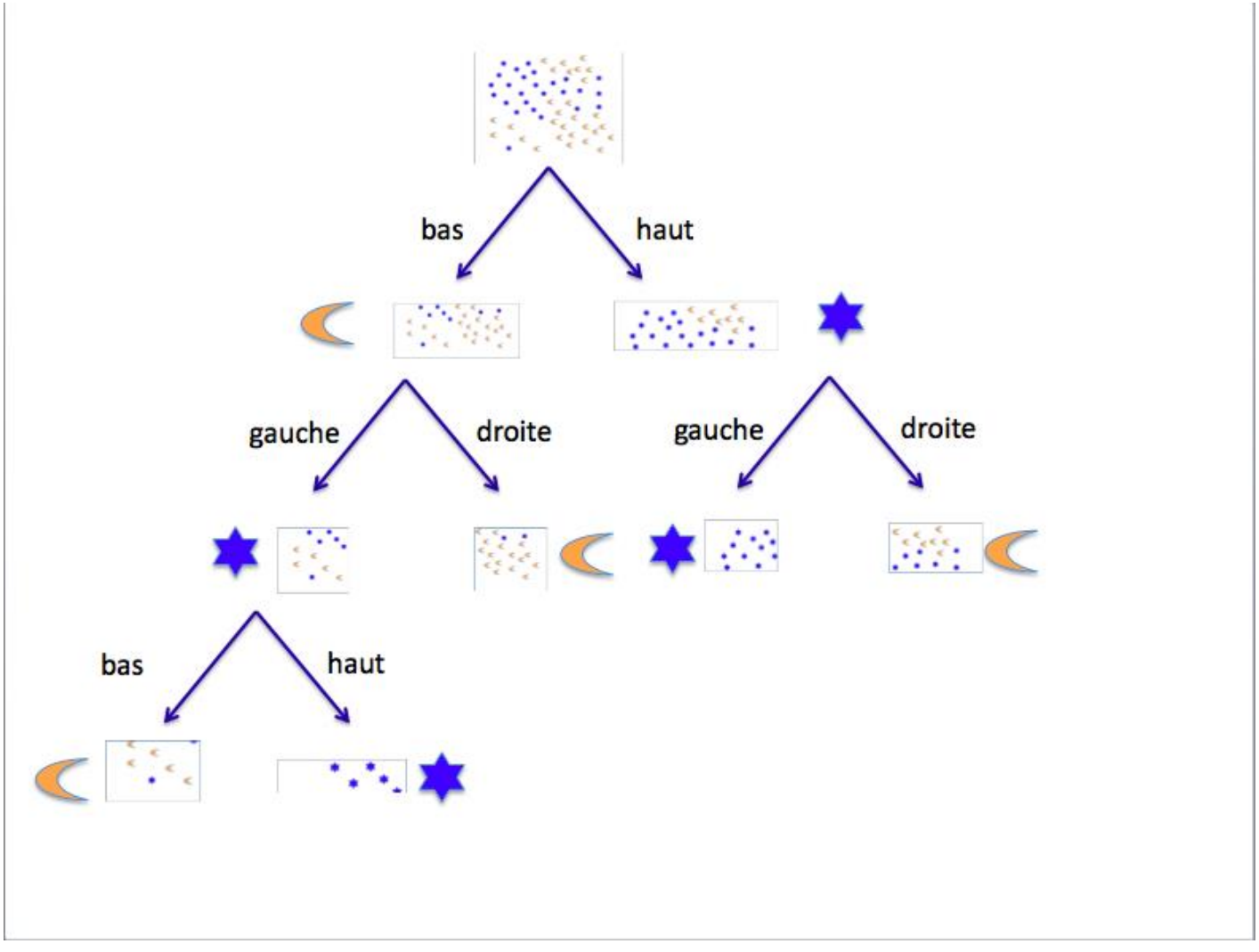
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



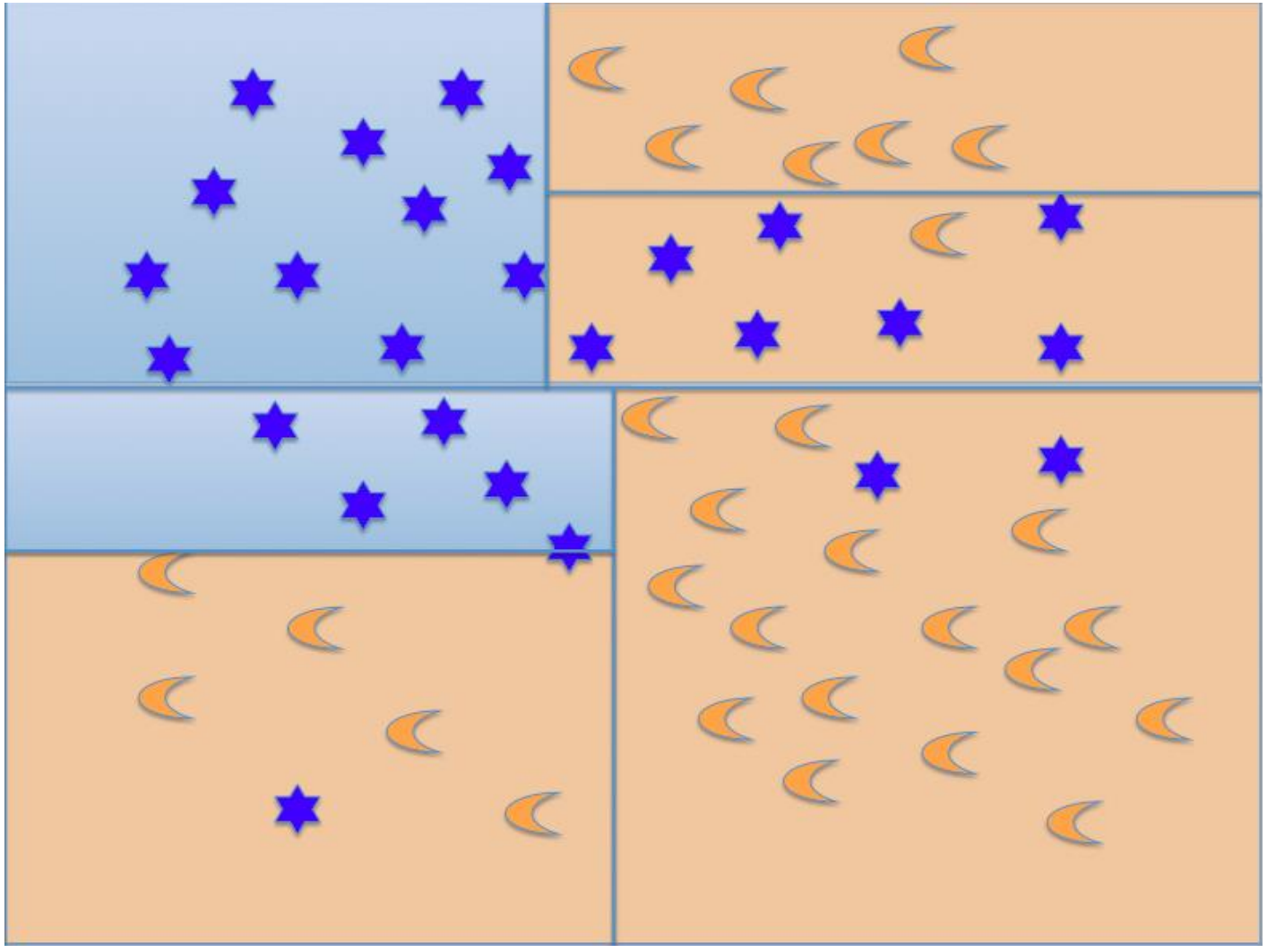
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



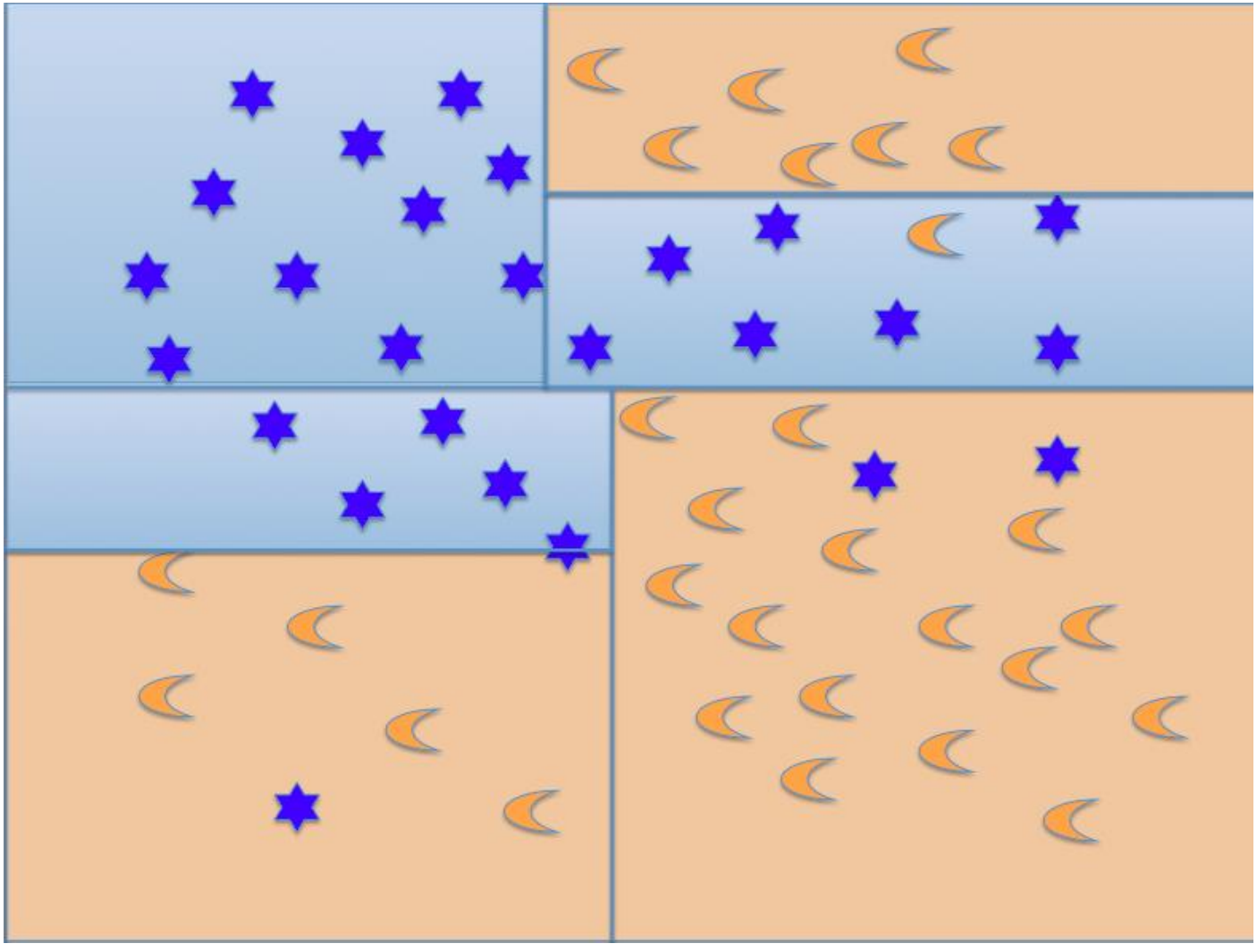
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



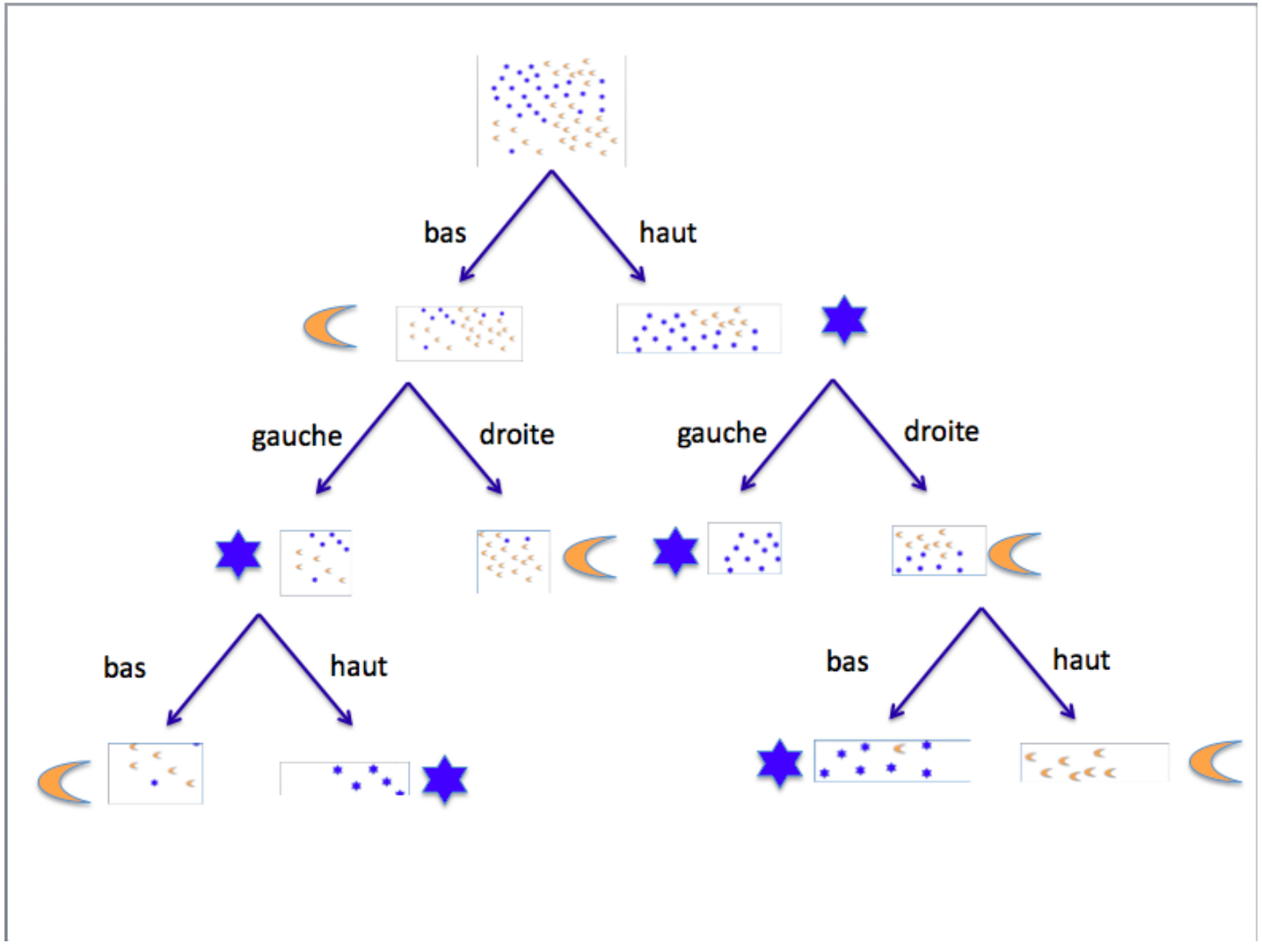
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



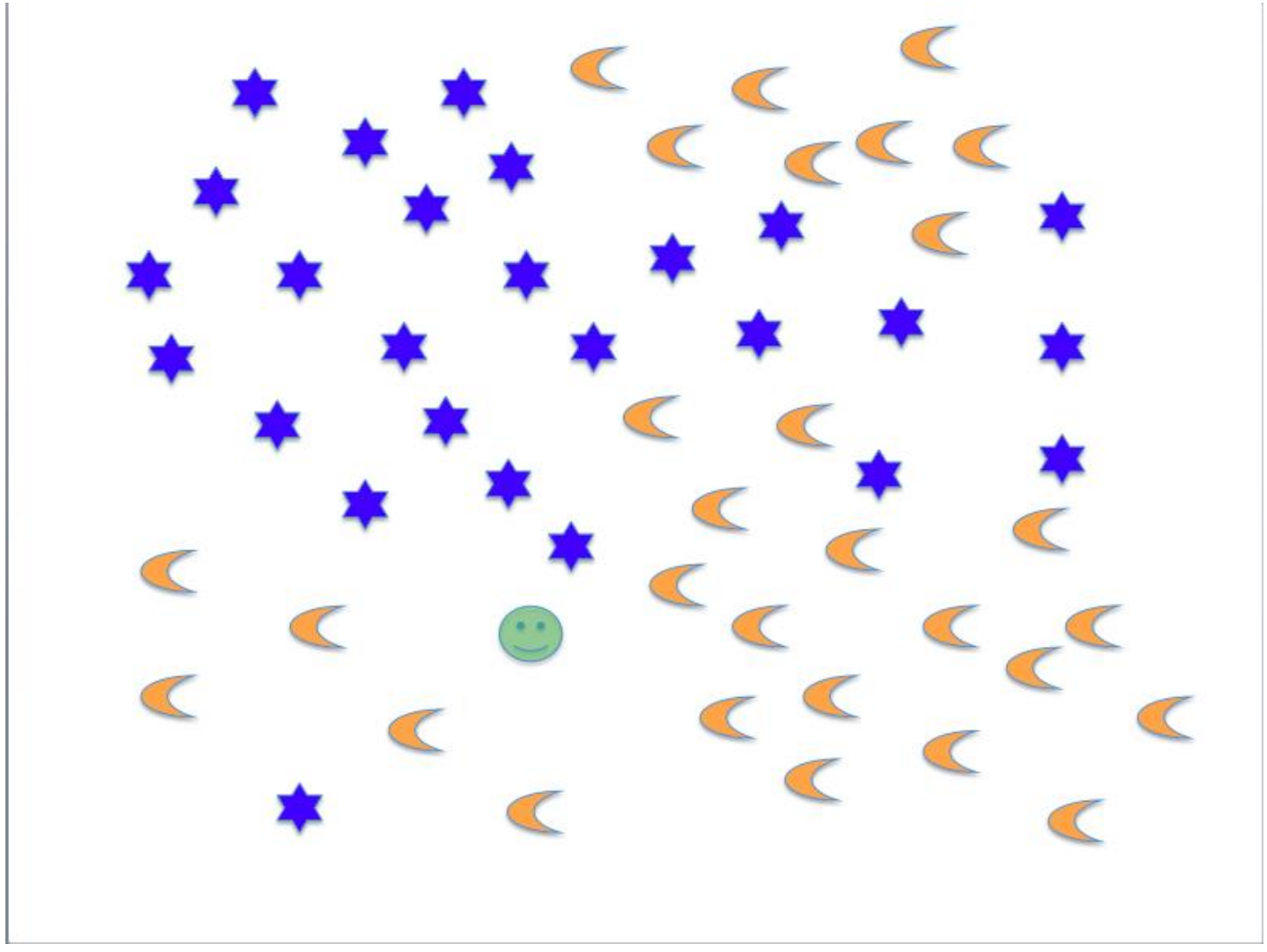
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



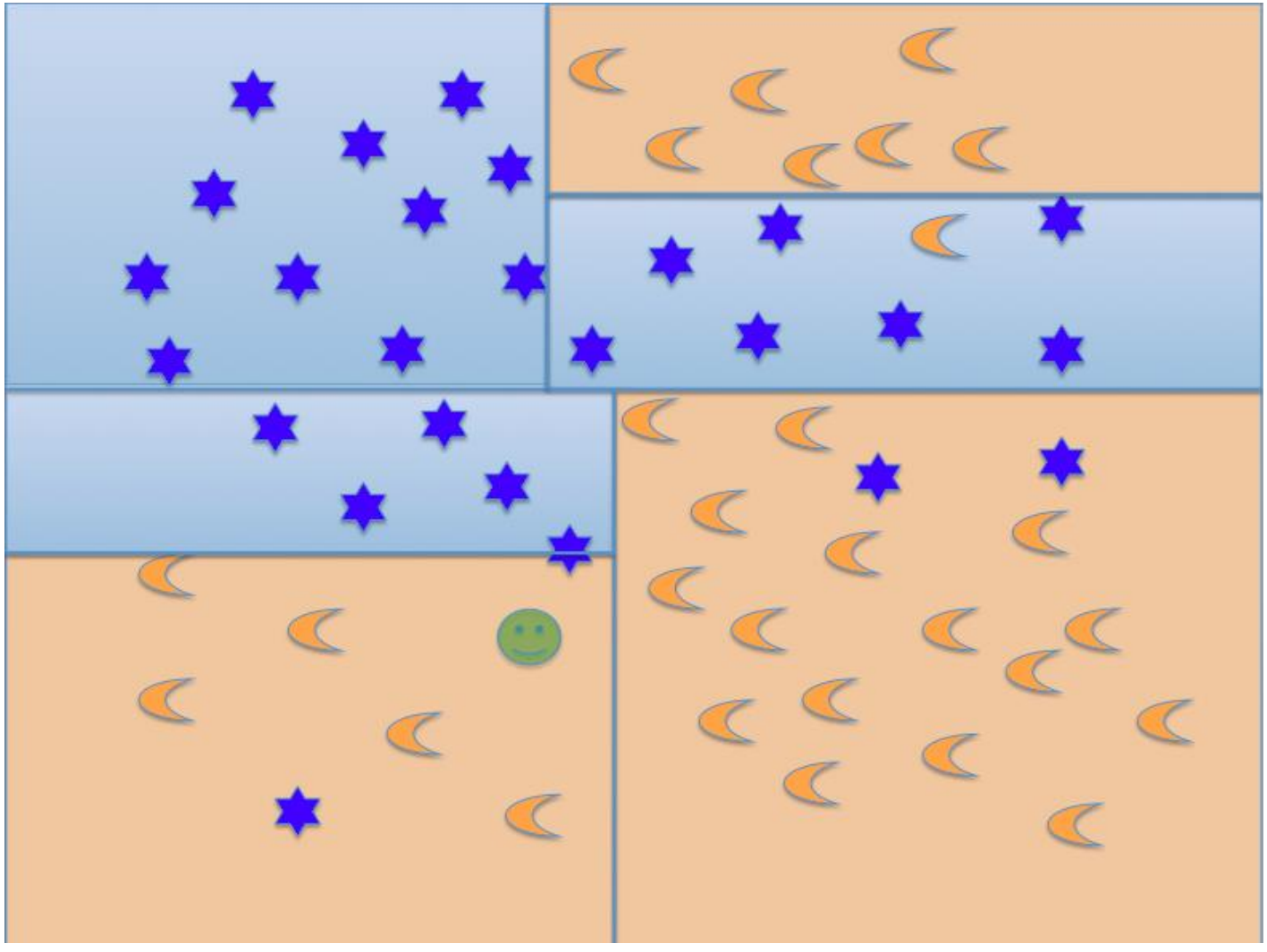
PHASE D'APPRENTISSAGE : VARIABLES CONTINUES



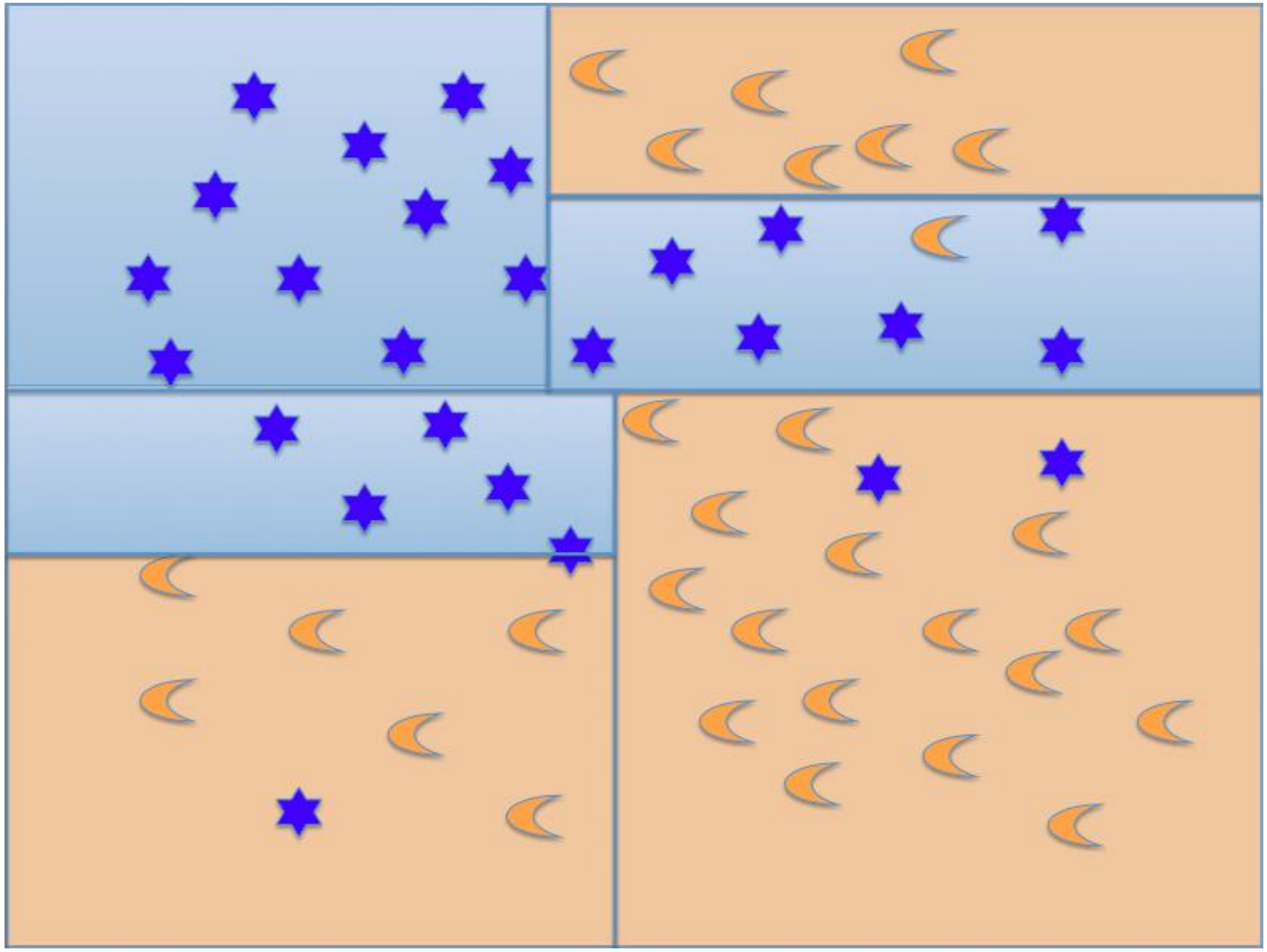
PHASE DE TEST : VARIABLES CONTINUES



PHASE DE TEST : VARIABLES CONTINUES



PHASE DE TEST : VARIABLES CONTINUES



VOCABULAIRE

Noeud : endroit de coupure

Racine : noeud initial, aucune coupure n'a été faite.

Feuille : extrémité de l'arbre (noeud qui n'est pas divisé).

Profondeur : nombre de niveaux de l'arbre

Taille minimale des feuilles : nombre minimal de points de l'ensemble d'apprentissage toléré dans une feuille.

Critère de séparation : critère selon lequel on choisit la variable/la coupure.

ALGORITHME

Algorithme : Arbre de decision.

Avec Ensemble d'apprentissage $(x_1, y_1), (x_2, y_2)..., (x_n, y_n)$.

1 : **Initialisation** : Arbre = Arbre vide ; N = racine.
2 : **while** N existe **do**
3 : **if** noeud N est terminal **then**
4 : N est une feuille. Donner une classe à N.
5 : **else**
6 : Créer des noeuds fils (selon la meilleure coupure)
7 : **end if**
8 : N = noeud non encore exploré s'il existe.
9 : **end while**

TESTER SI UN NOEUD EST TERMINAL

Deux critères :

Profondeur de l'arbre : si l'on a atteint la profondeur maximale, on arrête.

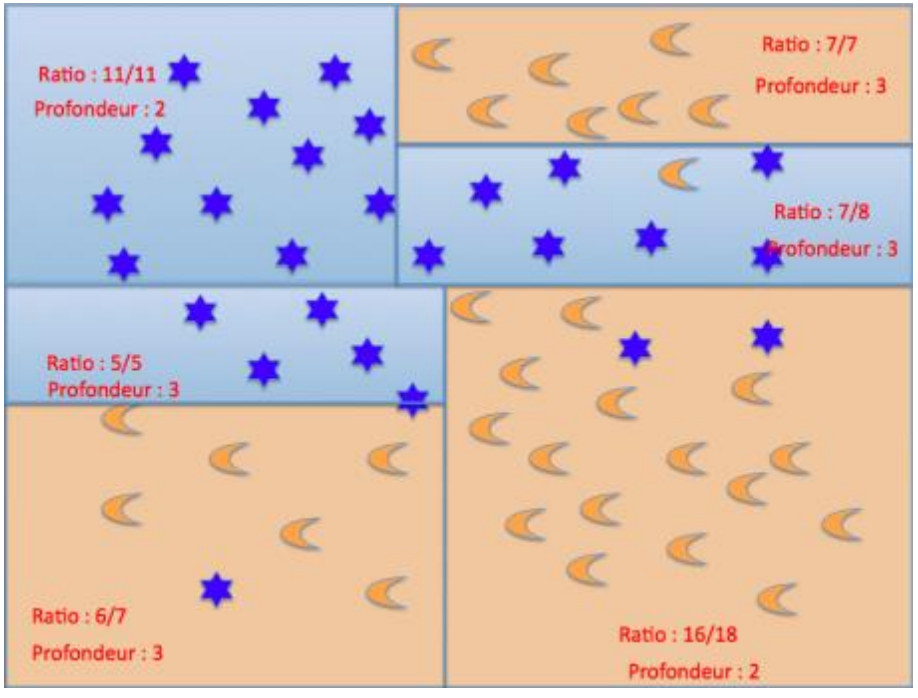
Proportion des classes dans le noeud : si le quotient classe majoritaire / total des points du noeud est supérieure à une valeur, on arrête.

TESTER SI UN NOEUD EST TERMINAL

Deux critères :

Profondeur de l'arbre : si l'on a atteint la profondeur maximale, on arrête.

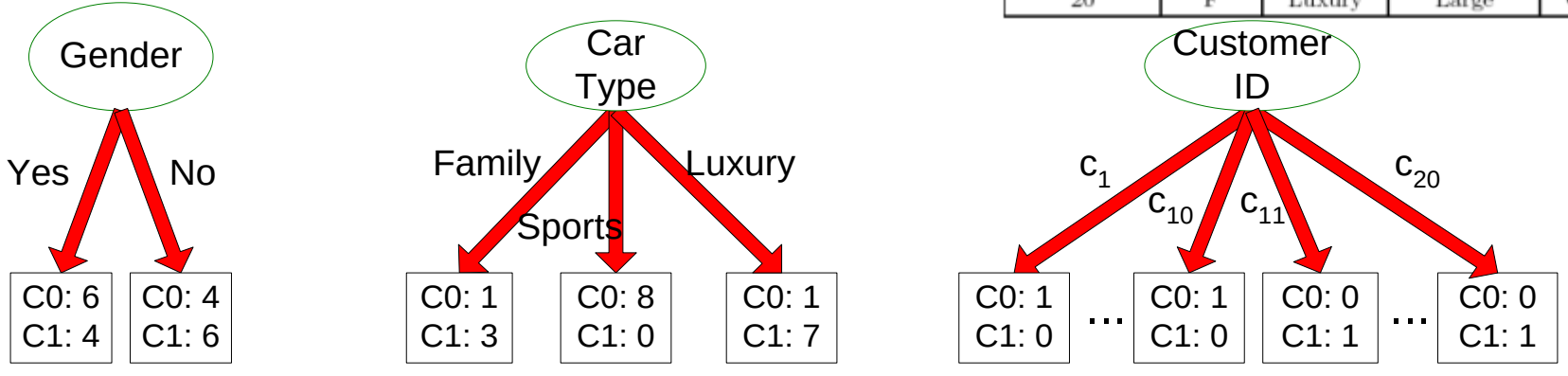
Proportion des classes dans le noeud : si le quotient classe majoritaire / total des points du noeud est supérieur à une valeur, on arrête.



COMMENT DETERMINER LA MEILLEURE COUPURE

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

Avant coupure : 10 exemples de classe C0 et 10 exemples de classe C1.



Quelle condition de test est la meilleure ?

COMMENT DETERMINER LA MEILLEURE COUPURE

- L’approche gloutonne :
  - Les nœuds avec une distribution de classe plus pure sont préférés
- Besoin d'une mesure d'impureté :

C0: 5 C1: 5

Degré élevé d'impureté

C0: 9 C1: 1

Faible degré d'impureté

CHOIX DE LA VARIABLE DE COUPURE

Objectif: Choisir la variable qui maximise le gain en **pureté** ou en information.
On n'a le droit que de segmenter que sur **une variable à la fois** (un axe).
Cas discret : la segmentation se fait sur toutes les valeurs possibles de la variable.

Cas continu : la segmentation est forcément simple : de la forme "variable > valeur, variable <= valeur".

Trois critères couramment utilisés :

- **l'indice de Gini**
- **l'entropie de Shannon**
- **l'erreur de classification**

L'approche est gloutonne : on teste chaque variable et chaque coupure possible et on choisit la meilleure.

MESURES D'IMPURETE (POUR UN NOEUD)

- Gini Index $$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

- Entropie $$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

- Erreur de classification $$Error(t) = 1 - \max_i P(i \mid t)$$

TROUVER LA MEILLEURE COUPURE

1. Calculer la mesure d'impureté (P) avant de réaliser une coupure
2. Calculer la mesure d'impureté (M) après avoir réalisé la coupure, c'est à dire :
    - Calcul de la mesure d'impureté pour chaque nœud enfant ; M est l'impureté pondérée des enfants
3. Choisir la condition de test d'attribut qui produit le gain le plus élevé

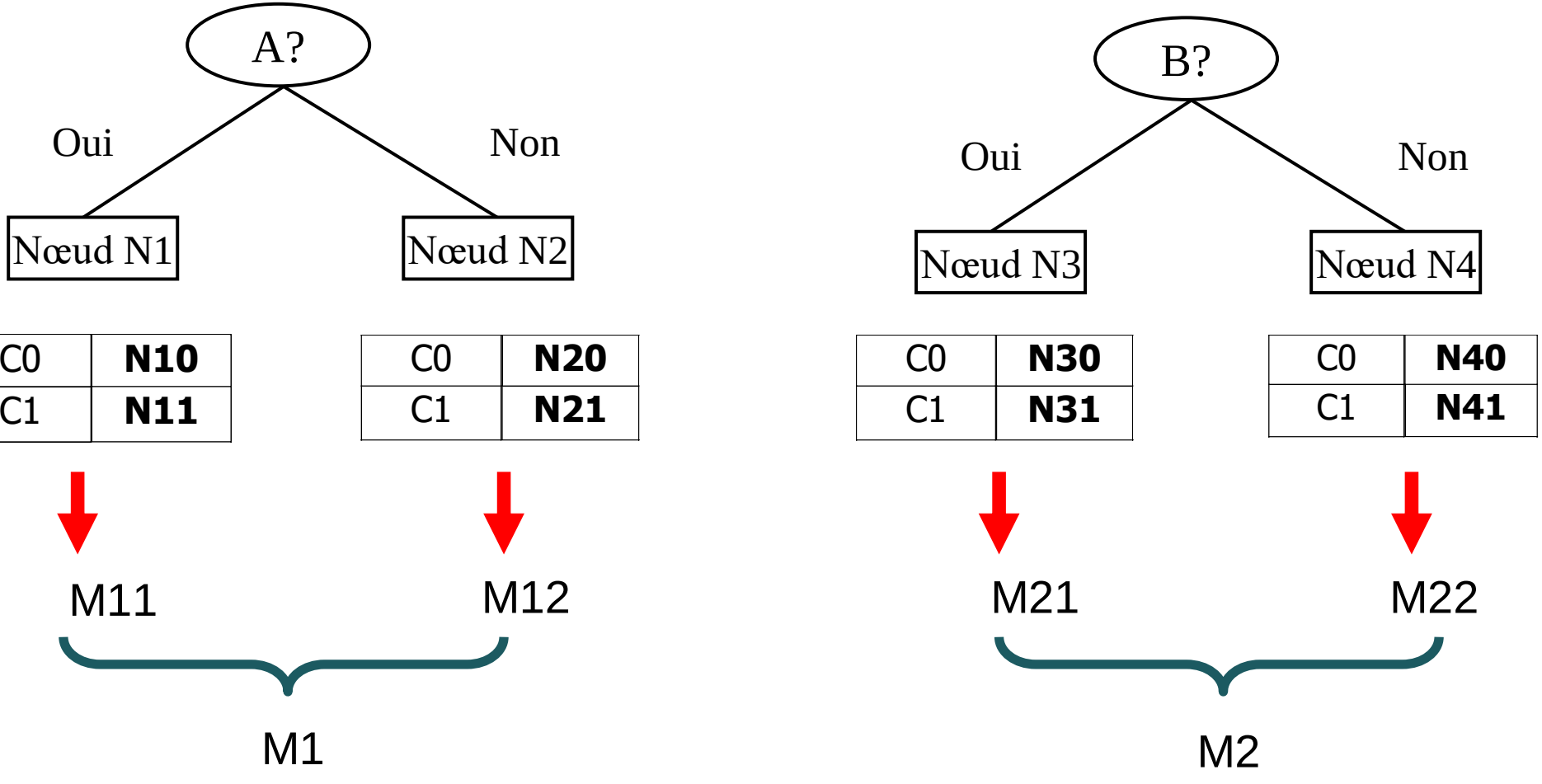
$$\text{Gain} = P - M$$

TROUVER LA MEILLEURE COUPURE

Avant de couper:

C0	**N00**
C1	**N01**

→ P



Gain = P – M1 vs P – M2

IMPURETE PAR L'INDEX DE GINI

- Pour un nœud t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

avec : $p(j \mid t)$ la fréquence relative de la classe j pour le nœud t.

  - La valeur maximale est atteinte lorsque les exemples sont répartis uniformément entre toutes les classes (= 1 - 1/nc ; nc = nombre de classes) ; ce qui implique une information la moins intéressantes.
  - La valeur minimale est atteinte lorsque tous les exemples du nœud appartiennent à la même catégorie (=0).

Measure of Impurity: GINI

- Pour un nœud t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

avec : $p(j \mid t)$ la fréquence relative de la classe j pour le nœud t.

  - Pour un problème à deux classes (p, 1 – p) :
    - GINI = $1 - p^2 - (1 - p)^2 = 2p\,(1\text{-}p)$

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

C1	**3**
C2	**3**
Gini=0.500	

CALCUL DE L'INDEX DE GINI POUR UN NOEUD

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

CALCUL DE L'INDEX DE GINI POUR PLUSIEURS NOEUDS

- Quand un nœud *p* est divisé en *k* partitions (enfants)

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

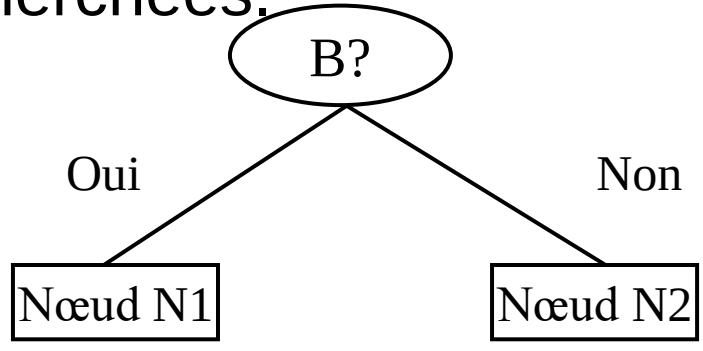
avec : n_i = nombre d'exemples pour l'enfant *i*,

n = nombre d'exemples pour le parent *p*.

- Choisir l'attribut qui minimise l'indice de Gini pondéré des enfants.
- L'indice de Gini est utilisé dans les algorithmes d'arbre de décision tels que : CART, SLIQ, SPRINT

ATTRIBUTS BINAIRES : CALCUL DE L'INDEX DE GINI PONDERE

- Coupures pour 2 partitions
- L'effet de pondération des partitions :
  - Des partitions plus grandes et plus pures sont recherchées.



	Parent
C1	**7**
C2	**5**
Gini = 0.486	

Gini(N1)
= 1 – $(5/6)^2$ – $(1/6)^2$
= 0.278

Gini(N2)
= 1 – $(2/6)^2$ – $(4/6)^2$
= 0.444

	N1	**N2**
C1	**5**	**2**
C2	**1**	**4**
Gini=0.361		

Gini pondéré pour N1 et N2
= 6/12 * 0.278 +
6/12 * 0.444
= 0.361

Gain = 0.486 – 0.361 = 0.125

ATTRIBUTS CATEGORIQUES : CALCUL DU GINI

- Pour chaque valeur distincte, compter pour chaque classe.
- Utiliser la matrice de comptage pour prendre des décisions.

Coupures multiples

	CarType		
	Family	Sports	Luxury
C1	1	8	1
C2	3	0	7
Gini	0.163		

Coupures binaires
(trouver la meilleure partition des valeurs)

	CarType	
	{Sports, Luxury}	{Family}
C1	9	1
C2	7	3
Gini	0.468	

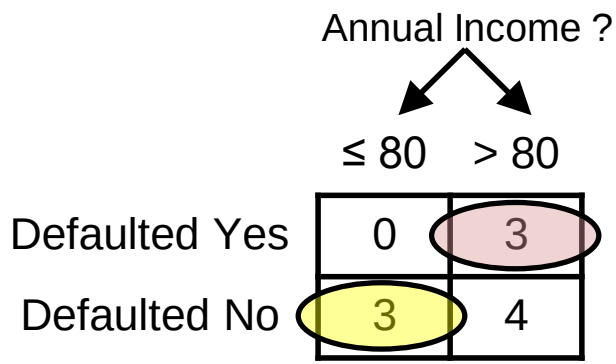
	CarType	
	{Sports}	{Family, Luxury}
C1	8	2
C2	0	10
Gini	0.167	

Lequel de ces choix est le meilleur ?

ATTRIBUTS CONTINUS : CALCUL DU GINI

- Utiliser des décisions binaires basées sur une seule valeur
- Plusieurs choix pour la valeur de coupure
  - Nombre de valeurs de coupure possibles = Nombre de valeurs distinctes
- Chaque valeur de coupure est associée à une matrice de comptage.
  - Nombre de classes dans chacune des partitions, A < v et A ≥ v
- Méthode simple pour choisir le meilleur *v* :
  - Pour chaque v, parcourir les données pour calculer la matrice de comptage et son indice de Gini.

ID	Home Owner	Marital Status	Annual Income	Defaulted
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



ATTRIBUTS CONTINUS : CALCUL DU GINI

- Pour chaque attribut :
  - Ordonner l'attribut selon ses valeurs
  - Parcourir linéairement ces valeurs, en mettant à jour à chaque fois la matrice de comptage et en calculant l'indice de Gini
  - Choisir la valeur qui a l'indice de Gini le plus faible

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Annual Income										
→	60	70	75	85	90	95	100	120	125	220

ATTRIBUTS CONTINUS : CALCUL DU GINI

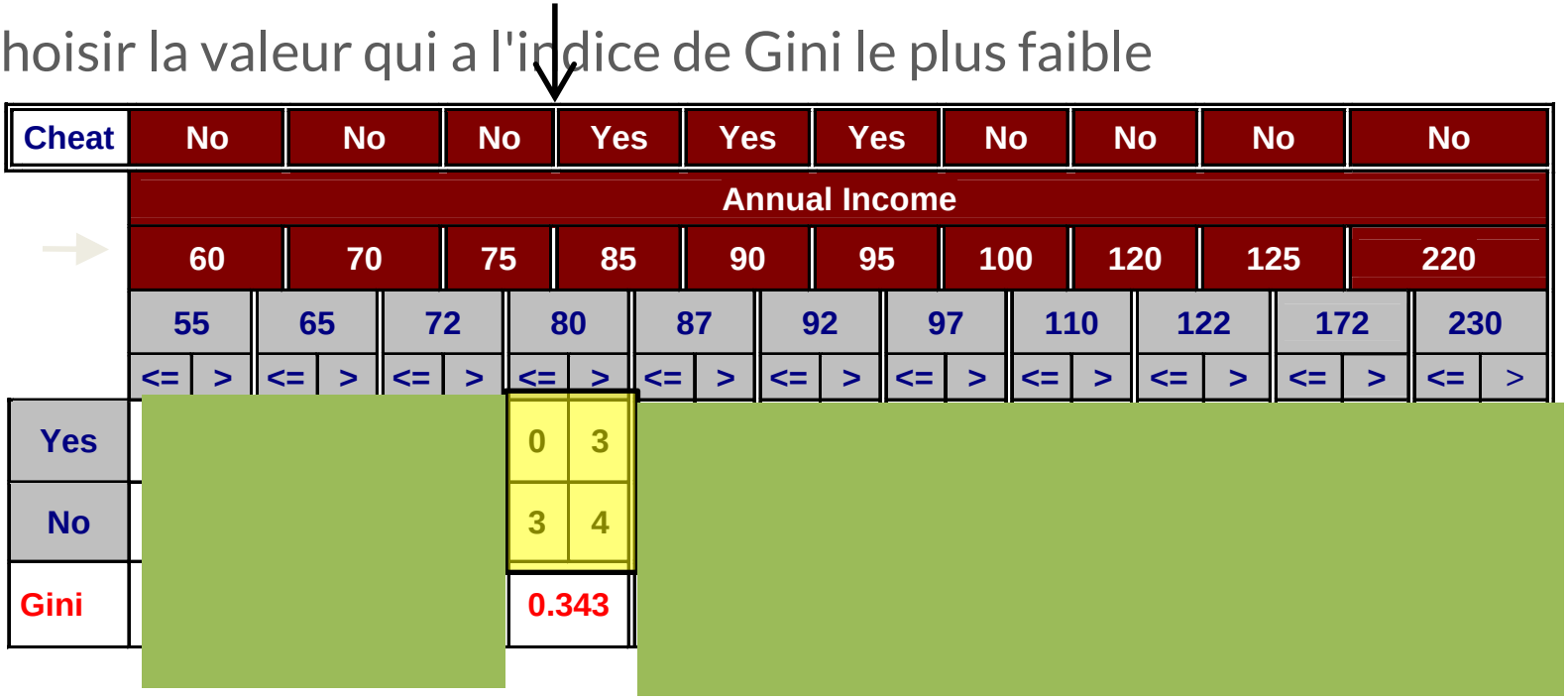
- Pour chaque attribut :
  - Ordonner l'attribut selon ses valeurs
  - Parcourir linéairement ces valeurs, en mettant à jour à chaque fois la matrice de comptage et en calculant l'indice de Gini
  - Choisir la valeur qui a l'indice de Gini le plus faible

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
Annual Income										
→	60	70	75	85	90	95	100	120	125	220

55		65		72		80		87		92		97		110		122		172		230	
<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>

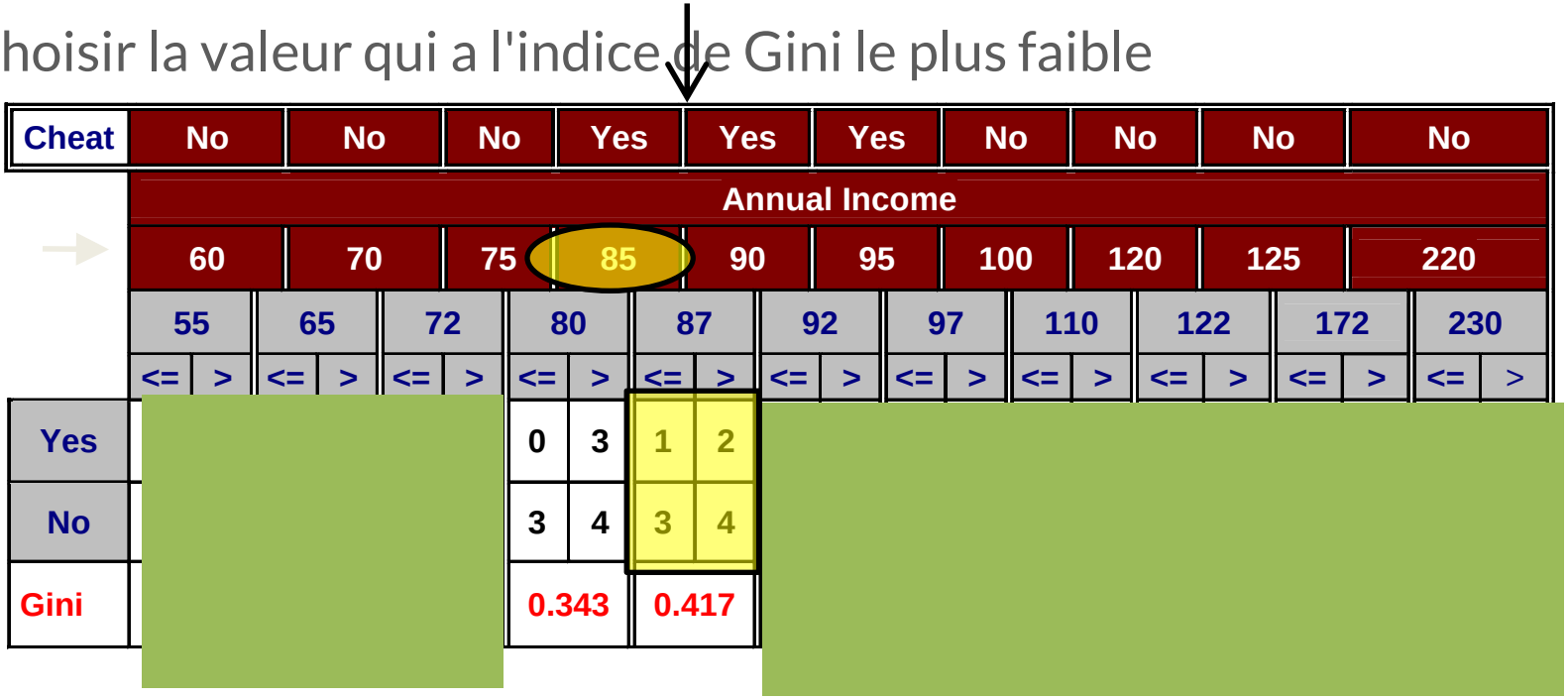
ATTRIBUTS CONTINUS : CALCUL DU GINI

- Pour chaque attribut :
  - Ordonner l'attribut selon ses valeurs
  - Parcourir linéairement ces valeurs, en mettant à jour à chaque fois la matrice de comptage et en calculant l'indice de Gini
  - Choisir la valeur qui a l'indice de Gini le plus faible



ATTRIBUTS CONTINUS : CALCUL DU GINI

- Pour chaque attribut :
  - Ordonner l'attribut selon ses valeurs
  - Parcourir linéairement ces valeurs, en mettant à jour à chaque fois la matrice de comptage et en calculant l'indice de Gini
  - Choisir la valeur qui a l'indice de Gini le plus faible



ATTRIBUTS CONTINUS : CALCUL DU GINI

- Pour chaque attribut :
  - Ordonner l'attribut selon ses valeurs
  - Parcourir linéairement ces valeurs, en mettant à jour à chaque fois la matrice de comptage et en calculant l'indice de Gini
  - Choisir la valeur qui a l'indice de Gini le plus faible

Cheat	No		No		No		Yes		Yes		Yes		No		No		No		No	
	Annual Income																			
→	60		70		75		85		90		95		100		120		125		220	
	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		***0.300***		0.343		0.375		0.400		0.420	

IMPURETE PAR L'ENTROPIE

- Entropie pour un nœud t :

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

avec : $p(j \mid t)$ la fréquence relative de la classe j pour le nœud t.

- La valeur maximale est atteinte lorsque les exemples sont répartis uniformément entre toutes les classes (= log(nc) ; nc = nombre de classes) ; ce qui implique une information la moins intéressante.
- La valeur minimale est atteinte lorsque tous les exemples du nœud appartiennent à la même catégorie (=0).

Le calcul de l'impureté par l'entropie ressemble à celui de l'indice de GINI.

CALCUL DE L'ENTROPIE (UN SEUL NOEUD)

$$Entropy(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropie = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropie = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Entropie = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

CALCUL DU GAIN APRES UNE COUPURE

- Gain :

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

le nœud parent p est divisé en k partitions (enfants) ;

avec : n_i = nombre d'exemples pour l'enfant I ;
n = nombre d'exemples pour le parent p.

  - Choisir la coupure qui réduit le plus le GAIN (maximisation du GAIN).
  - Utilisé par les algorithmes ID3 et C4.5.

QUOTIENT DE GAIN

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO} \qquad SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

le nœud parent *p* est divisé en *k* partitions (enfants) ;

avec : n_i = nombre d'exemples pour l'enfant *I* ;
n = nombre d'exemples pour le parent *p*.

- Adapte le Gain en fonction de l'entropie du partitionnement (SplitINFO).

  Pénalise les régions avec peu d'exemples.

- Utilisé dans C4.5.

L'ERREUR DE CLASSIFICATION

- Erreur de classification au nœud t:

$$Error(t) = 1 - \max_i P(i \mid t)$$

  - La valeur maximale est atteinte lorsque les exemples sont répartis uniformément entre toutes les classes (= 1 -1/nc) ; ce qui implique une information la moins intéressante.
  - La valeur minimale est atteinte lorsque tous les exemples du nœud appartiennent à la même catégorie (=0).

ERREUR DE CLASSIFICATION (POUR UN NOEUD)

$$Error(t) = 1 - \max_i P(i \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Erreur = 1 – max (0, 1) = 1 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Erreur = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

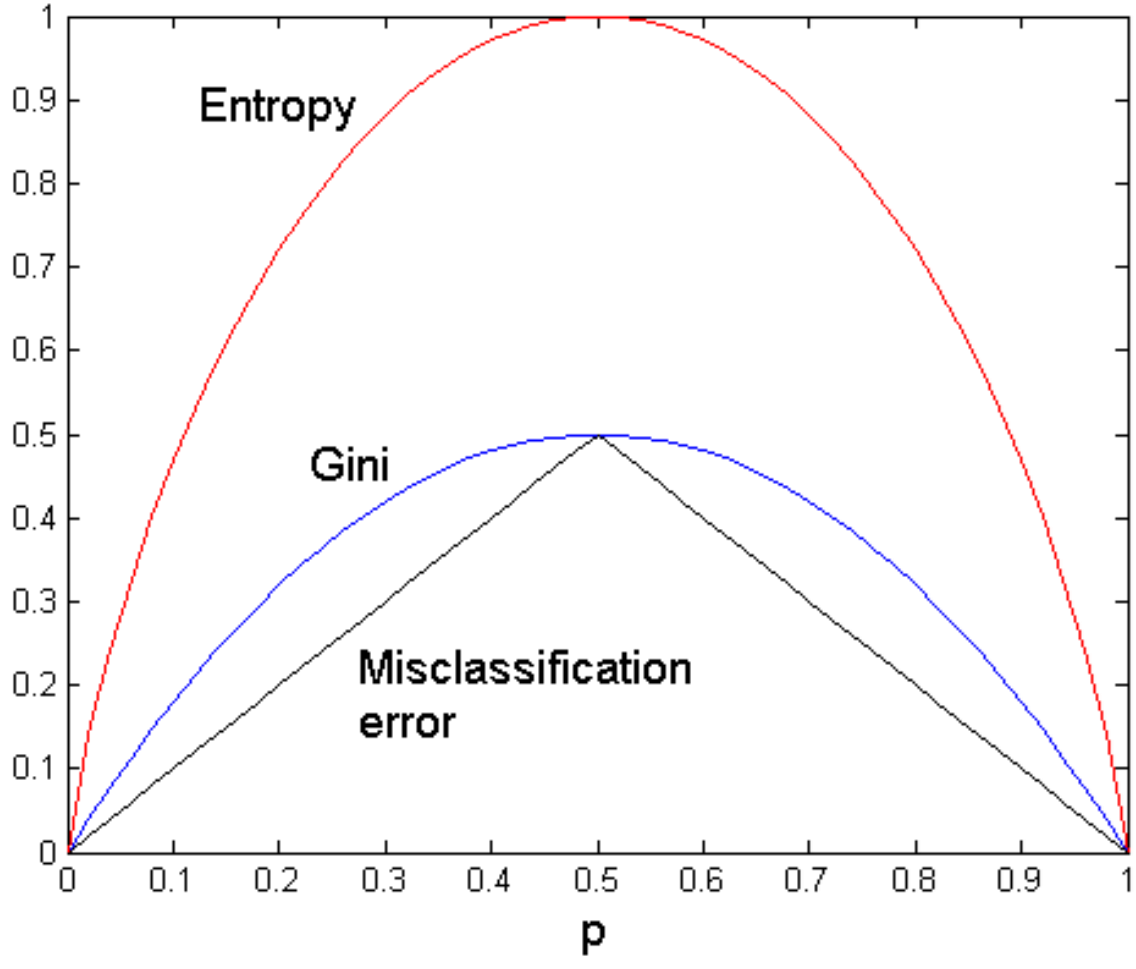
C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

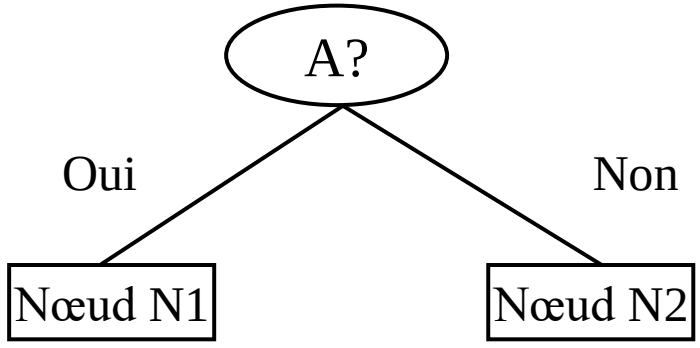
Erreur = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

COMPARAISON DES MESURES D'IMPURETE

Pour un problème à 2 classes :



ERREUR DE CLASSIFICATION ET GINI



	Parent
C1	**7**
C2	**3**
Gini = 0.42	

Gini(N1)
= 1 – $(3/3)^2$ – $(0/3)^2$
= 0

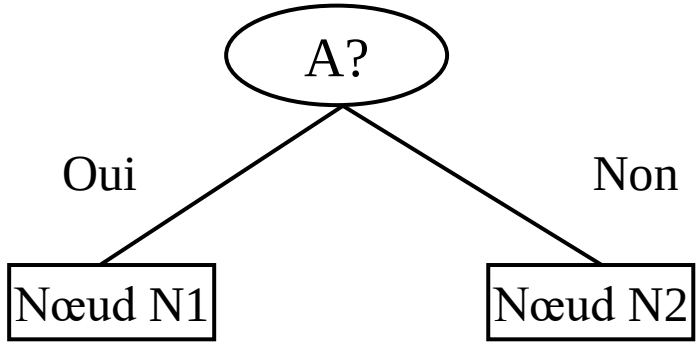
Gini(N2)
= 1 – $(4/7)^2$ – $(3/7)^2$
= 0.489

	N1	N2
C1	**3**	**4**
C2	**0**	**3**
Gini=0.342		

Gini(Children)
= 3/10 * 0
+ 7/10 * 0.489
= 0.342

Le Gini est amélioré, mais l'erreur reste la même (= 3/10)

ERREUR DE CLASSIFICATION ET GINI



	Parent
C1	**7**
C2	**3**
Gini = 0.42	

	N1	N2
C1	**3**	**4**
C2	**0**	**3**
Gini=0.342		

	N1	N2
C1	**3**	**4**
C2	**1**	**2**
Gini=0.416		

L’erreur dans les trois cas = 0.3 !

CONCLUSION

Avantages:

Très interprétables !

Faciles à implémenter

Inconvénients:

Forte élasticité aux exemples : si l'on change un exemple, l'arbre peut changer complètement. On risque donc de faire du **sur-apprentissage**.

Peuvent nécessiter beaucoup de calculs à cause de l'approche gloutonne.