

Machine Learning^{4DS}

Notions de base sur l'apprentissage supervisé
CLASSIFICATION

Mohamed Heny SELMI

medheny.selmi@esprit.tn

Enseignant et Responsable option Data Science à ESPRIT

Diversité de données



Structuration de données



Données issues d'un SGBDR

Données structurées selon un processus BI

Data Lake

Matrice Documents Termes

Data Hub

Fichiers tabulaires

Images en niveaux de gris [0:255, 256 niveaux de gris possibles]

N scènes issues d'une vidéo

Tableau de données



Individus, observations, objets, enregistrements, etc.

Variables, caractères, attributs, Descripteurs, champs, etc.

age	sexe	typedouleur	sucré	tauxmax	angine	depression	coeur
70	masculin	D	A	109	non	24	presence
67	feminin	C	A	160	non	16	absence
57	masculin	B	A	141	non	3	presence
64	masculin	D	A	105	oui	2	absence
74	feminin	B	A	121	oui	2	absence
65	masculin	D	A	140	non	4	absence
56	masculin	C	B	142	oui	6	presence
59	masculin	D	A	142	oui	12	presence
60	masculin	D	A	170	non	12	presence
63	feminin	D	A	154	non	40	presence
59	masculin	D	A	161	non	5	absence
53	masculin	D	A	111	oui	0	absence
44	masculin	C	A	180	non	0	absence
61	masculin	A	A	145	non	26	presence
57	feminin	D	A	159	non	0	absence
71	feminin	D	A	125	non	16	absence
46	masculin	D	A	120	oui	18	presence
53	masculin	D	B	155	oui	31	presence
64	masculin	A	A	144	oui	18	absence
40	masculin	A	A	178	oui	14	absence
67	masculin	D	A	129	oui	26	presence
48	masculin	B	A	180	non	2	absence
43	masculin	D	A	181	non	12	absence

Statut des variables dans le processus supervisé

Variables prédictives
Descripteurs
Variables exogènes

De type quelconque

Temperature	Sun	Heat	Rain	Quality
3064	1201	10	361	medium
3000	1053	11	338	bad
3155	1133	19	393	medium
3085	970	4	467	bad
3245	1258	36	294	good
3267	1386	35	225	good
3080	966	13	417	bad
2974	1189	12	488	bad
3038	1103	14	677	bad
3318	1310	29	427	medium
3317	1362	25	326	good
3182	1171	28	326	bad
2998	1102	9	349	bad
3221	1424	21	382	good
3019	1230	16	275	medium
3022	1285	9	303	medium
3094	1329	11	339	medium
3009	1210	15	536	bad
3227	1331	21	414	medium
3308	1366	24	282	good
3212	1289	17	302	medium
3361	1444	25	253	good
3061	1175	12	261	medium
3478	1317	42	259	good
3126	1248	11	315	medium
3458	1508	43	286	good
3252	1361	26	346	medium
3052	1186	14	443	bad
3270	1399	24	306	good
3198	1259	20	367	good
2904	1164	6	311	bad
3247	1277	19	375	good
3083	1195	5	441	bad
3043	1208	14	371	bad

Variable à prédire
Variable cible
Variable décisionnelle
Décision
Classe
Variable endogène

De type :

Qualitative : Classification
Quantitative : Régression

Idée de base de l'apprentissage supervisé

Population Ω Y : variable à prédire
 X : série de variable prédictive, $X = (X_1, X_2, \dots, X_n)$

Objectif :

Se servir des données disponibles sur la population, objet d'étude afin de construire un **modèle** (une fonction de classement) tel que :

$$Y = f(X, \alpha)$$

Utiliser un échantillon Ω_a : un extrait de la population totale, pour choisir la meilleure fonction f et ses paramètres α en minimisant le taux d'erreur

Apprentissage bayésien

cas particulier du problème à 2 classes – Positifs vs. Négatifs

Apprentissage en 2 étapes à partir des données :

- estimer la probabilité d'affectation $P(Y / X)$
- prédire $[Y = +]$ si $P(Y = + / X) > P(Y = - / X)$



- $P(Y = + / X)$ est selon le cas appelé « score » ou « appétence » : c'est la « propension à être un positif »
- Cette méthode d'affectation minimise l'erreur de prédiction c'est un cas particulier du coût de mauvaise affectation

généralisation à K classes

Apprentissage en 2 étapes à partir des données :

- estimer la probabilité d'affectation $P(Y = y_k / X)$
- prédire $y_{k^*} = \arg \max P(Y = y_k / X)$



Remarque : Lorsque les X sont discrets, nous pouvons en déduire un modèle logique d'affectation.

SI $X_1 = X_2 = \dots, X_n =$ ALORS $Y =$



Minimiser l'erreur théorique

- ✓ Pas de solution directe pour les descripteurs continus (discrétisation ou hypothèse de distribution)
- ✓ Pas de sélection et d'évaluation des descripteurs (individuellement ou des groupes de variables – donc pas de sélection)
- ✓ Dès que le nombre de descripteurs augmente
 - Problème de calculabilité
- ✓ Nombre d'opérations énorme, ex. 10 descr. Binaires $\Rightarrow 2^{10}$ règles
 - Problème de fragmentation des données Plein de cases avec des 0, estimations peu fiables
- ✓ Cette approche n'est pas utilisable dans la pratique !

Evaluation

Evaluation de l'apprentissage



Le modèle exprime une connaissance

- Explication : comprendre la causalité pour mieux l'exploiter
- Validation : l'expert du métier peut évaluer la pertinence de l'expertise
- Amélioration : l'expert peut intervenir pour ajuster des paramètres calculés

Rapidité

- En apprentissage : pouvoir tester plusieurs pistes à savoir l'ajout de variables, le test de combinaison de variables, modification de paramètres
- En Prédiction (Classification ou Régression) : attribuer un label de la variable à prédire à un nouvel utilisateur

Précision

- Évaluer la précision et la qualité du modèle construit lors de son utilisation future

Matrice de confusion

Principe : confronter les valeurs réelles de la variable cible avec celles obtenus en application un modèle de prédiction

		Prédite		
		+	-	Total
Observée	+	a	b	a+b
	-	c	d	c+d
	Total	a+c	b+d	n

Indicateurs issus de la matrice de confusion

- Vrais positifs VP = a
- Faux positifs FP = c
- Taux d'erreur = $(c+b)/n$
- Sensibilité = Rappel = Taux de VP = $a/(a+b)$
- Précision = $a/(a+c)$
- Taux de FP = $c/(c+d)$
- Spécificité = $d/(c+d) = 1 - \text{Taux de FP}$

		Prédite		
		+	-	Total
Observée	+	a	b	a+b
	-	c	d	c+d
	Total	a+c	b+d	n

Utilité des indicateurs de performance issus de la matrice de confusion

Coûts de mauvaises affectation

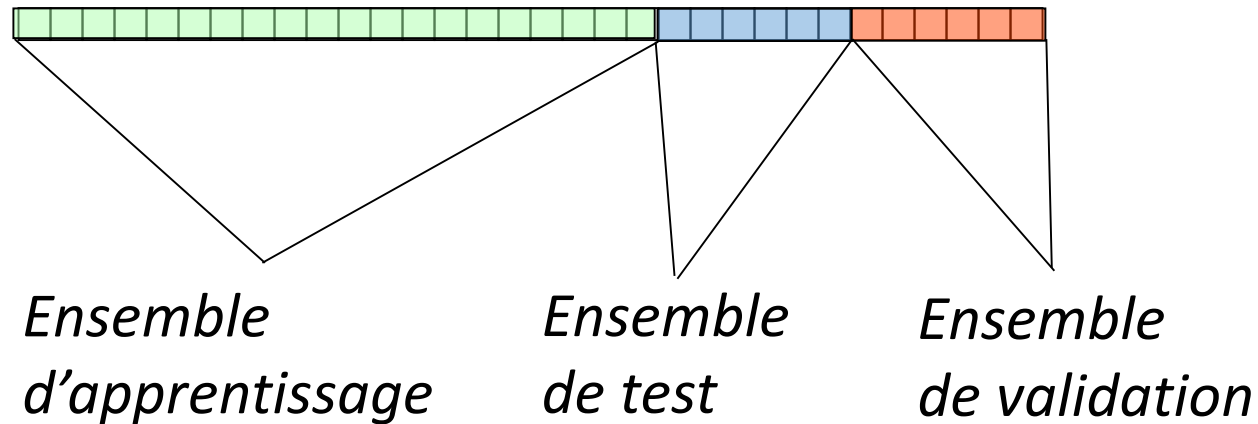
Comparaison de deux modèles de classification

		Prédite		
		+	-	Total
Observée	+	40	10	50
	-	20	30	50
	Total	60	40	100

		Prédite		
		+	-	Total
Observée	+	20	30	50
	-	0	50	50
	Total	20	80	100

Principe : Apprentissage et Test

Point de départ : Ensemble des données disponible



Subdivision aléatoire

- **Échantillon d'apprentissage**
 - Utilisé pour la construction du modèle 70%
- **Échantillon test**
 - Utilisé pour l'évaluation du modèle 30%
 - Calcul du Rappel, précision, taux d'erreur...

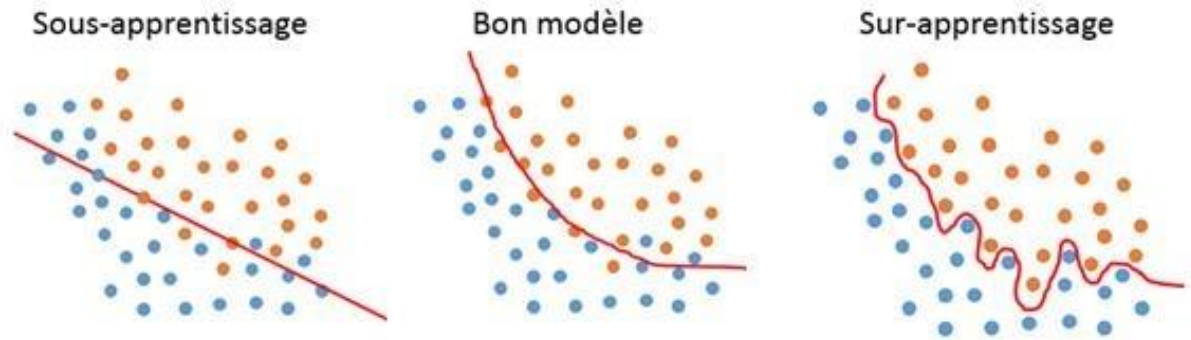
Les méthodes de ré-échantillonnage

Validation croisée

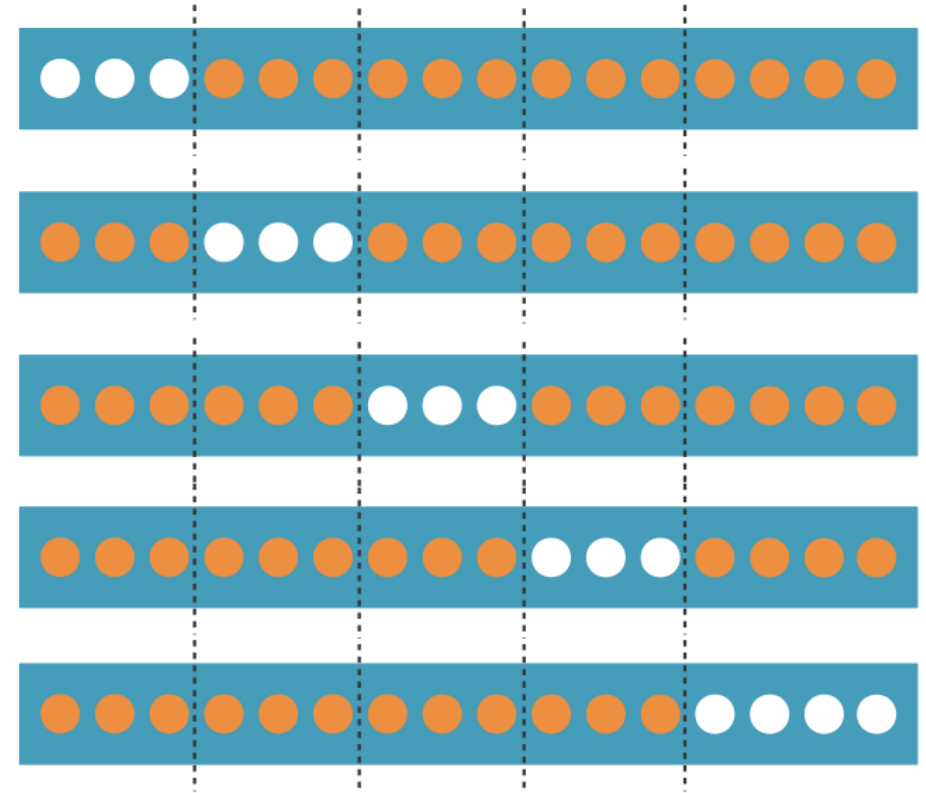
Leave-one-out

Bootstrap

La validation croisée

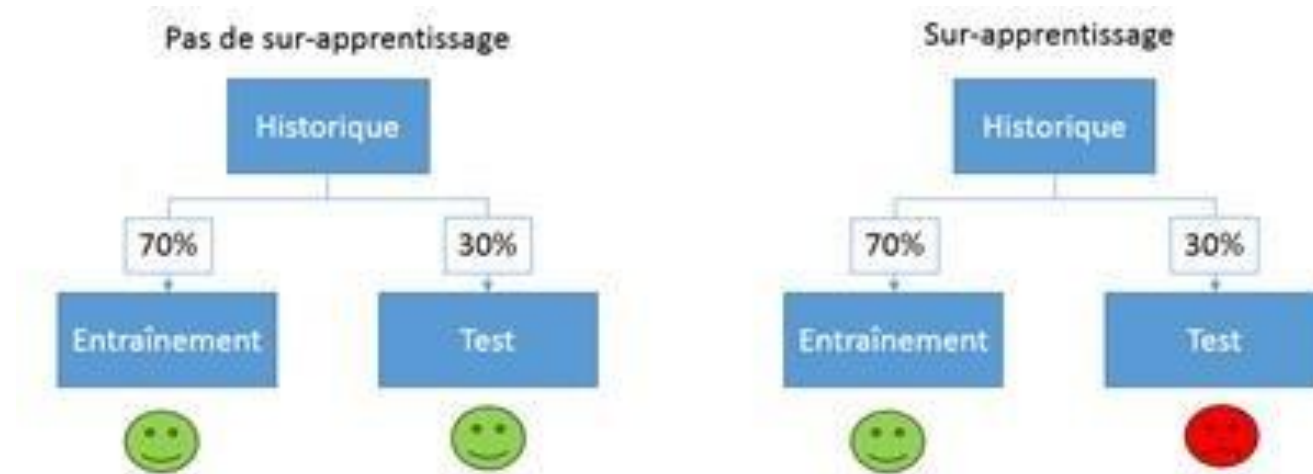


➔ Motivation, combattre **la malédiction de la dimensionnalité**



La validation croisée : variantes

- Variante 1 – Division entraînement test



- Variante 2 – Validation croisée à k plis

Variante 3 -LOO (Leave One Out)

Utilisation de l'ensemble de validation

- On règle les paramètres de l'algorithme d'apprentissage
 - E.g. : nb de couches cachées, nb de neurones, ...
 - en essayant de réduire l'erreur de test
- Pour avoir une estimation non optimiste de l'erreur,

il faut recourir à une base d'exemples non encore vus : la ***base de validation***

Principe de la validation croisée

Algorithme

- Subdiviser l'échantillon en K blocs
 - Pour chaque k :
 - Construire sur le modèle $M(X, n - n_k)$
 - Calculer l'erreur en test sur $n_k \rightarrow e_k$
 - e_{cv} = la moyenne des erreurs e_k
- *Utiliser la validation répétée ($B \times K$ -Fold Cross validation) améliore les caractéristiques de la modélisation*
 - *Sur les cas d'un fort sur-apprentissage (certaines méthodes mal paramétrées, ratio élevé de variables vs. individus, beaucoup de variables non pertinentes, etc.), la validation croisée (avec K élevé) a tendance à sous-estimer l'erreur !!!*

Leave-one-out

Algorithme

- Subdiviser l'échantillon en $K=n$ blocs
 - Pour chaque individu k :
 - Construire sur le modèle $M(X, n-1)$
 - Calculer l'erreur en test sur le $k^{\text{ième}}$ ind $\rightarrow e_k$ // $e_k = 1$ si erreur ou 0 si bonne classification
 - Calculer la moyenne e_{loo} des erreurs en test // e_{loo} **proportion des erreurs**
-
- Nettement plus coûteux en calcul que la K validation croisée sans être meilleur
 - Sous-estimation (dramatique) de l'erreur en cas de sur apprentissage

Bootstrap

Algorithme

- Répéter B fois (on parle de répliques)
 - Tirage avec remise d'un échantillon de taille $n \rightarrow \Omega_b$
 - Distinguer les individus non échantillonnés $\rightarrow \Omega(b)$
 - Apprentissage du modèle sur Ω_b
 - Erreur en resubstitution sur Ω_b [er(b)]
 - Erreur en test sur $\Omega(b)$ [et(b)]
 - Calcul de l'optimisme α_b

Sur l'échantillon complet, calculer l'erreur en resubstitution

Courbe ROC

Receiving Operating Characteristics

Nécessité d'évaluer les modèles de prédiction

Évaluer les performances d'un modèle de prédiction est primordial :

1. Pour savoir si le modèle est globalement significatif : Mon modèle traduit-il vraiment une causalité ?
2. Pour se donner une idée des performances en déploiement : Quelle sera la fiabilité (les coûts associés) lorsque j'utiliserai mon modèle ?
3. Pour comparer plusieurs modèles candidats : Lequel parmi plusieurs modèles sera le plus performant compte tenu de mes objectifs ?



- ✓ Le taux d'erreur semble être un indicateur synthétique pertinent.
- ✓ il indique (estime) la probabilité de mal classer un individu de la population.
- ✓ Les autres indicateurs sont très intéressants également (sensibilité/rappel, précision/spécificité) mais obligent à surveiller plusieurs valeurs simultanément.

Le taux d'erreur : un indicateur trop réducteur



M_1

	^positif	^négatif	Total
positif	40	10	50
négatif	10	40	50
Total	50	50	100

20%



M_2

	^positif	^négatif	Total
positif	30	20	50
négatif	5	45	50
Total	35	65	100

25%



Conclusion : Modèle 1 serait meilleur que Modèle 2

Cette conclusion -- basée sur l'échantillon test -- suppose que la matrice de coût de mauvais classement est symétrique et unitaire

Introduction d'une matrice de coût

Coût de mauvaise affectation non-symétrique

	^positif	^négatif
positif	0	1
négatif	10	0

	^positif	^négatif	Total
positif	40	10	50
négatif	10	40	50
Total	50	50	100

1,1

	^positif	^négatif	Total
positif	30	20	50
négatif	5	45	50
Total	35	65	100

0,7

Conclusion : Modèle 2 serait meilleur que Modèle 1 dans ce cas ???



Les matrices de coûts sont souvent le fruit d'opportunités conjoncturelles :
Faudrait-il tester toutes les matrices de coûts possibles pour comparer M1 et M2 ?

Peut-on bénéficier d'un dispositif qui permet de comparer globalement les modèles,
indépendamment de la matrice de coût de mauvaise affectation ?

Problème des distributions déséquilibrées

Lorsque les classes sont très déséquilibrées, la matrice de confusion et surtout le taux d'erreur donnent souvent une fausse idée de la qualité de l'apprentissage.

Exp. Marketing ciblé, Assurance, Crédit Bancaire, etc.

0.0627			
Confusion matrix			
	No	Yes	Sum
No	5435	39	5474
Yes	326	22	348
Sum	5761	61	5822

0.0650			
Confusion matrix			
	No	Yes	Sum
No	3731	31	3762
Yes	229	9	238
Sum	3960	40	4000



- ✓ Le classifieur par défaut (prédire systématiquement NO), propose un taux d'erreur de $238 / 4000 = 0.0595$
- ✓ Moralité : modéliser ne servirait à rien dès que les classes sont très déséquilibrées
- ✓ Cette anomalie est liée au fait que nous voulons absolument que le modèle réalise une affectation (positif ou négatif).
- ✓ Or dans de nombreux domaines, ce qui nous intéresse avant tout, c'est de mesurer la propension à être positif ou négatif !

Objectifs de la courbe ROC

La courbe ROC est un outil d'évaluation et de comparaison des modèles

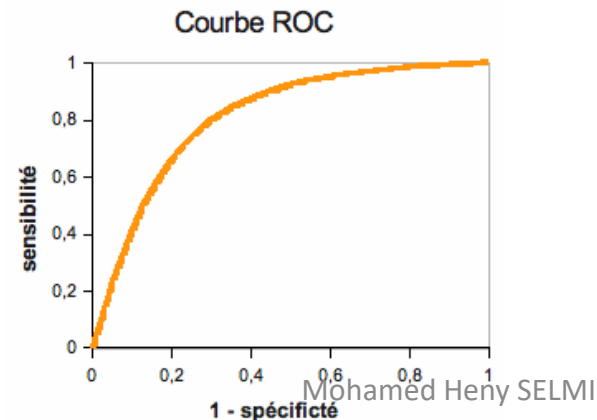
1. Indépendant des matrices de coûts de mauvaise affectation : Il permet de savoir si M1 sera toujours meilleur que M2 quelle que soit la matrice de coût
2. Opérationnel même dans le cas des distributions très déséquilibrées : Sans les effets pervers de la matrice de confusion liés à la nécessité de réaliser une affectation
3. Résultats valables même si l'échantillon test n'est pas représentatif : Tirage prospectif ou tirage rétrospectif : les indications fournies restent les mêmes
4. Un outil graphique qui permet de visualiser les performances : Un seul coup d'œil doit permettre de voir le(s) modèle(s) susceptible(s) de nous intéresser
5. Un indicateur synthétique associé : Aisément interprétable



Sa portée va largement au-delà des interprétations (indicateurs) issues de l'analyse de la matrice de confusion.

Cadre de l'utilisation de la courbe ROC

- Variable cible ayant deux modalités
- Une modalité d'intérêt : (+)



Principe de la courbe ROC

		Classe réelle	
		-	+
Classe prédite	-	True Negatives <i>(vrais négatifs)</i>	False Negatives <i>(faux négatifs)</i>
	+	False Positives <i>(faux positifs)</i>	True Positives <i>(vrais positifs)</i>

TVP = Rappel = Sensibilité = VP/Positifs (True Positif Rate)

TFP = 1 – Spécificité = FP/Négatifs (False Positif Rate)

Idée de base de la courbe ROC

$P(Y=+/X) \geq P(Y=-/X)$ équivaut à une règle d'affectation $P(Y=+/X) \geq 0.5$ (seuil = 0.5)

Cette règle d'affectation fournit une matrice de confusion MC_1 , et donc 2 indicateurs TVP_1 et TFP_1

Si nous choisissons un autre seuil (0.6 par ex.), nous obtiendrons MC_2 et donc TVP_2 et TF_2 Etc... MC_i , TVP_i , TFP_i

L'idée de la courbe ROC est de faire varier le « seuil » de 1 à 0 et, pour chaque cas, calculer le TVP et le TFP que l'on reporte dans un graphique : en abscisse le TFP, en ordonnée le TVP.

Construction de la courbe ROC

Classer les données selon un score décroissant



Card(Positif) = 6
Card(Négatif) = 14

Individu	Score (+)	Classe
1	1	+
2	0.95	+
3	0.9	+
4	0.85	-
5	0.8	+
6	0.75	-
7	0.7	-
8	0.65	+
9	0.6	-
10	0.55	-
11	0.5	-
12	0.45	+
13	0.4	-
14	0.35	-
15	0.3	-
16	0.25	-
17	0.2	-
18	0.15	-
19	0.1	-
20	0.05	-

Seuil = 1

	positif	négatif	Total
positif	1	5	6
négatif	0	14	14
Total	1	19	20

Seuil = 0,95

	positif	négatif	Total
positif	2	4	6
négatif	0	14	14
Total	2	18	20

Seuil = 0,9

	positif	négatif	Total
positif	3	3	6
négatif	0	14	14
Total	3	17	20

Seuil = 0,85

	positif	négatif	Total
positif	3	3	6
négatif	1	13	14
Total	4	16	20

Seuil = 0

	positif	négatif	Total
positif	6	0	6
négatif	14	0	14
Total	20	0	20

$$TVP = \frac{1}{6} = 0,2 ; TFP = \frac{0}{14} = 0$$

$$TVP = \frac{2}{6} = 0,33 ; TFP = \frac{0}{14} = 0$$

$$TVP = \frac{3}{6} = 0,5 ; TFP = \frac{0}{14} = 0$$

$$TVP = \frac{3}{6} = 0,5 ; TFP = \frac{1}{14} = 0,07$$

$$TVP = \frac{6}{6} = 1 ; TFP = \frac{14}{14} = 1$$

Schématisation de la courbe ROC

Objectif : Mettre en relation TFP (abscisse) et TVP (ordonnée)

Individu	Score (+)	Classe	TFP	TVP
			0	0.000
1	1	+	0.000	0.167
2	0.95	+	0.000	0.333
3	0.9	+	0.000	0.500
4	0.85	-	0.071	0.500
5	0.8	+	0.071	0.667
6	0.75	-	0.143	0.667
7	0.7	-	0.214	0.667
8	0.65	+	0.214	0.833
9	0.6	-	0.286	0.833
10	0.55	-	0.357	0.833
11	0.5	-	0.429	0.833
12	0.45	+	0.429	1.000
13	0.4	-	0.500	1.000
14	0.35	-	0.571	1.000
15	0.3	-	0.643	1.000
16	0.25	-	0.714	1.000
17	0.2	-	0.786	1.000
18	0.15	-	0.857	1.000
19	0.1	-	0.929	1.000
20	0.05	-	1.000	1.000

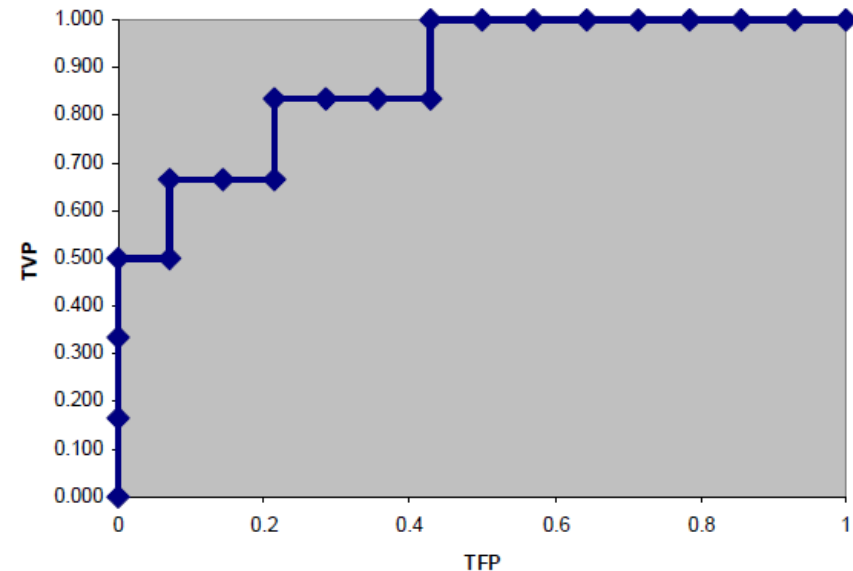
Calcul pratique :

$TFP(i) =$

Nombre de négatifs parmi les « i » premiers / (nombre total des négatifs)

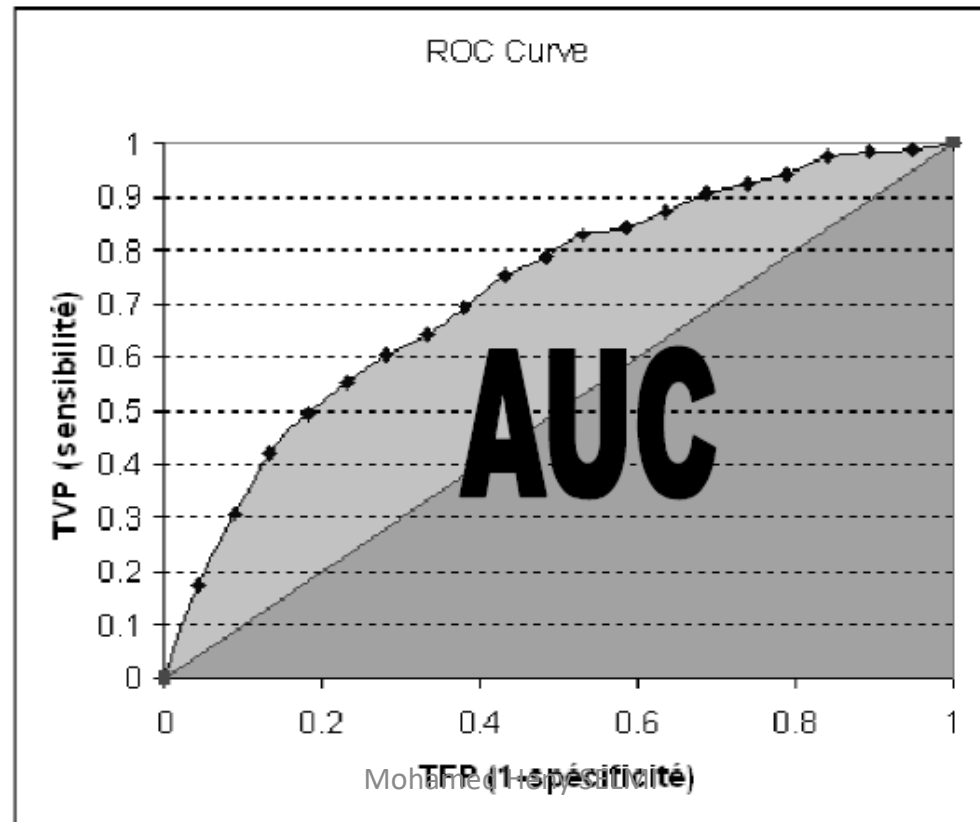
$TVP(i) =$

Nombre de positifs parmi les « i » premiers / (nombre total des positifs)



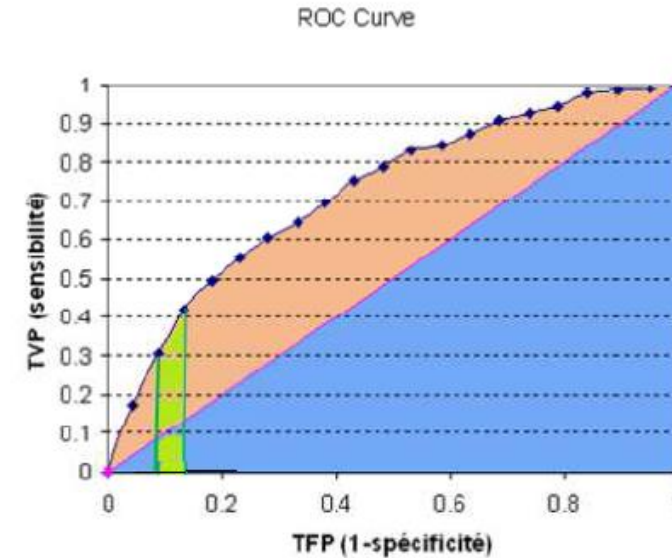
Interprétation basée sur l'AUC : aire sous la courbe

- ✓ AUC est la surface située sous la courbe ROC, c'est une mesure de la performance d'un score et la qualité de discrimination du modèle en traduisant la probabilité qu'un vrai positif aura un score supérieur au score d'un faux positif
- ✓ AUC varie entre 0 et 1, et en pratique elle est entre 0,5 et 1, car si $AUC < 0,5$ cela signifie que les scores ont été inversés
- ✓ AUC indique la probabilité pour que la fonction SCORE place un positif devant un négatif dans le meilleur des cas $AUC = 1$



AUC – Intégration avec la méthode des trapèzes

Individu	Score (+)	Classe	TFP	TVP	Largeur	Hauteur	Surface
			0	0.000			
1	1	+	0.000	0.167	0.000	0.083	0.000
2	0.95	+	0.000	0.333	0.000	0.250	0.000
3	0.9	+	0.000	0.500	0.000	0.417	0.000
4	0.85	-	0.071	0.500	0.071	0.500	0.036
5	0.8	+	0.071	0.667	0.000	0.583	0.000
6	0.75	-	0.143	0.667	0.071	0.667	0.048
7	0.7	-	0.214	0.667	0.071	0.667	0.048
8	0.65	+	0.214	0.833	0.000	0.750	0.000
9	0.6	-	0.286	0.833	0.071	0.833	0.060
10	0.55	-	0.357	0.833	0.071	0.833	0.060
11	0.5	-	0.429	0.833	0.071	0.833	0.060
12	0.45	+	0.429	1.000	0.000	0.917	0.000
13	0.4	-	0.500	1.000	0.071	1.000	0.071
14	0.35	-	0.571	1.000	0.071	1.000	0.071
15	0.3	-	0.643	1.000	0.071	1.000	0.071
16	0.25	-	0.714	1.000	0.071	1.000	0.071
17	0.2	-	0.786	1.000	0.071	1.000	0.071
18	0.15	-	0.857	1.000	0.071	1.000	0.071
19	0.1	-	0.929	1.000	0.071	1.000	0.071
20	0.05	-	1.000	1.000	0.071	1.000	0.071
AUC							0.881



Surface d'un trapèze

$$s_i = (TFP_i - TFP_{i-1}) \times \frac{TVP_i + TVP_{i-1}}{2}$$

$$\rightarrow AUC = \sum_i s_i$$



AUC – Statistique de Mann-Whitney

Individu	Score (+)	Classe	Rangs	Rangs +
1	1	+	20	20
2	0.95	+	19	19
3	0.9	+	18	18
4	0.85	-	17	0
5	0.8	+	16	16
6	0.75	-	15	0
7	0.7	-	14	0
8	0.65	+	13	13
9	0.6	-	12	0
10	0.55	-	11	0
11	0.5	-	10	0
12	0.45	+	9	9
13	0.4	-	8	0
14	0.35	-	7	0
15	0.3	-	6	0
16	0.25	-	5	0
17	0.2	-	4	0
18	0.15	-	3	0
19	0.1	-	2	0
20	0.05	-	1	0

Somme (Rang +)	95
U+	74

AUC	0.881
-----	-------

- ✓ Test de Mann-Whitney : montrer que deux distributions sont différentes (décalées).
- ✓ Statistique basée sur les rangs.
- ✓ Dans notre contexte, montrer que les « + » présentent en moyenne des scores plus élevés que les « - ».
- ✓ On peut en dériver un test statistique.

Somme des rangs des « + »

$$S_+ = \sum_{i:y_i=+} r_i = 20 + 19 + 18 + 16 + 13 + 9 = 95$$

Statistique de Mann-Whitney

$$U_+ = S_+ - \frac{n_+ (n_+ + 1)}{2} = 95 - \frac{6 \times 7}{2} = 74$$

AUC

$$AUC = \frac{U_+}{n_+ \times n_-} = \frac{74}{6 \times 14} = 0,881$$

Mohamed Heny SLM



AUC – Dénombrer les inversions *swaps*

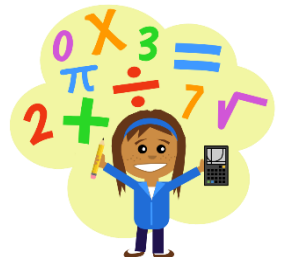
Individu	Score (+)	Classe	Nb de "-" devant un "+"
1	1	+	0
2	0.95	+	0
3	0.9	+	0
4	0.85	-	0
5	0.8	+	1
6	0.75	-	0
7	0.7	-	0
8	0.65	+	3
9	0.6	-	0
10	0.55	-	0
11	0.5	-	0
12	0.45	+	6
13	0.4	-	0
14	0.35	-	0
15	0.3	-	0
16	0.25	-	0
17	0.2	-	0
18	0.15	-	0
19	0.1	-	0
20	0.05	-	0

Swaps		10
AUC		0.881

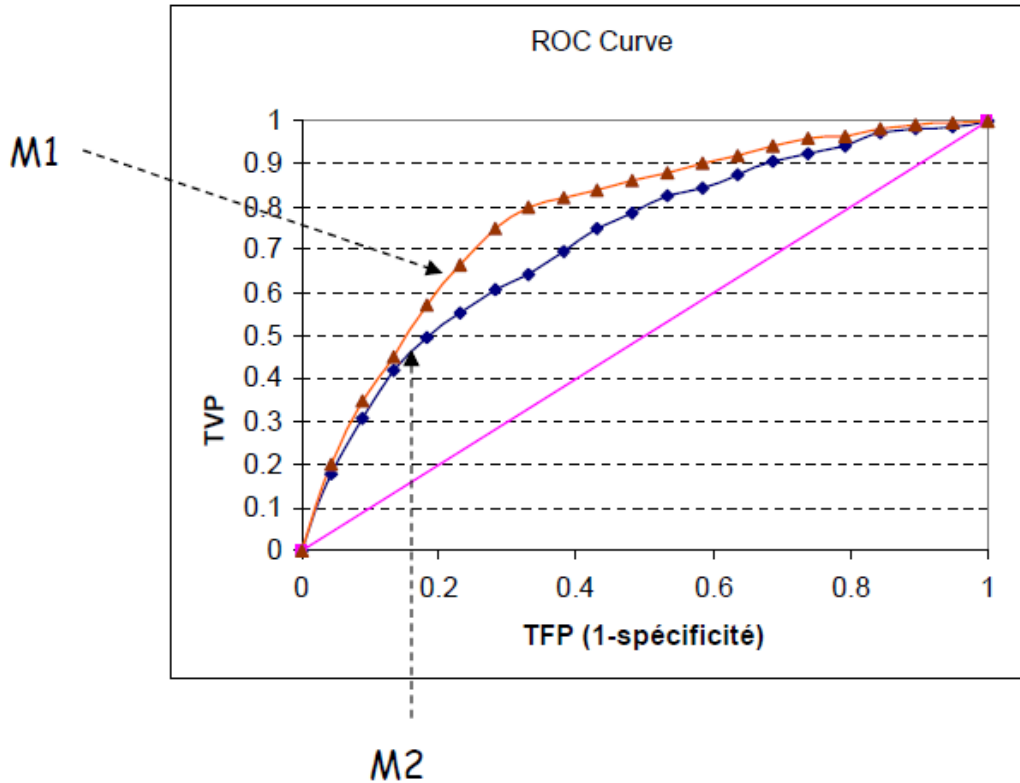
- Trier les individus selon un score décroissant.
- Pour chaque « + », compter le nombre de « - » qui le précède.
- Dans notre contexte, on souhaite que les scores élevés soient attribués aux « + » en priorité c.-à-d. les « + » sont peu précédés de « - ».

$$swaps = \sum_{i:y_i=+} c_i = 0 + 0 + 0 + 1 + 3 + 6 = 10$$

$$AUC = 1 - \frac{swaps}{n_+ \times n_-} = 1 - \frac{10}{6 \times 14} = 0,881$$



Interprétation : Dominance



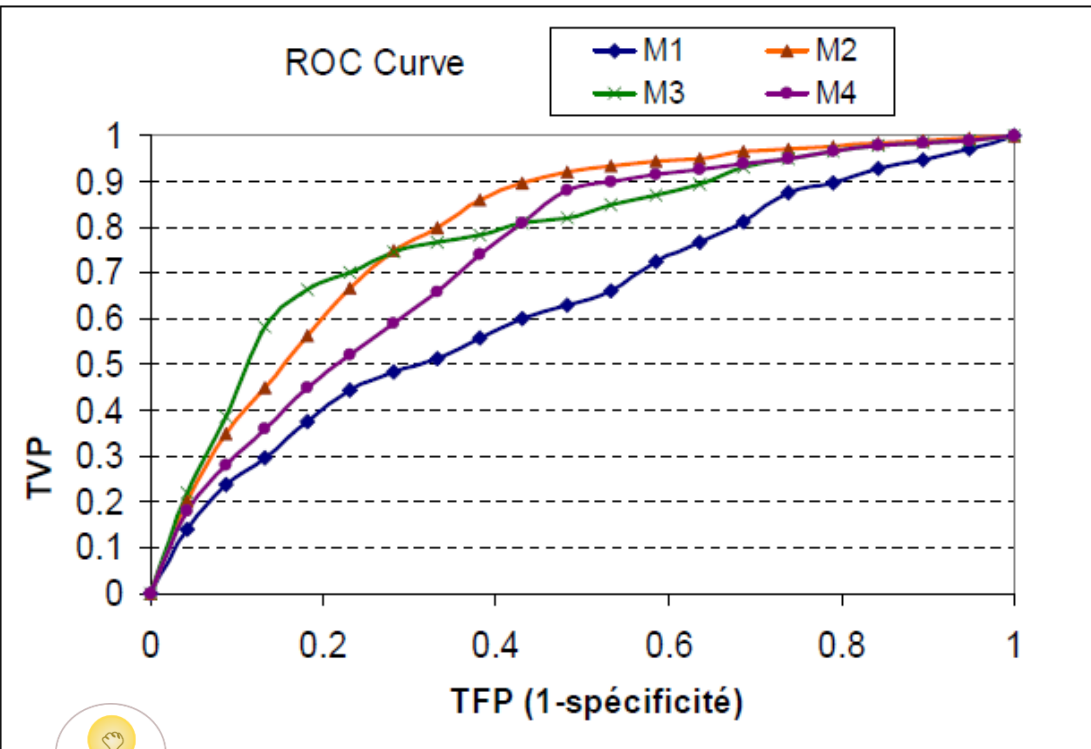
Comment montrer que M1 sera toujours meilleur que M2, quelle que soit la matrice de coût de mauvaise affectation utilisée ?



Cas Parfait :

La courbe de M1 est toujours « au-dessus » de celle de M2 : il ne peut pas exister de situation (matrice de coût de mauvais classement) où M2 serait un meilleur modèle de prédiction.

Interprétation : Dominance



Autres cas : Chevauchement entre les courbes ROC

Parmi un ensemble de modèles candidats, comment éliminer d'office ceux qui ne seront pas intéressants ?.

Autres cas : Enveloppe convexe

- Elle est formée par les courbes qui, à un moment ou à un autre, n'ont aucune courbe « au-dessus » d'elles.
- Les courbes situées sur cette enveloppe correspondent aux modèles qui sont potentiellement les plus performantes pour une matrice de coût donnée.
- Les modèles qui ne participent jamais à cette enveloppe peuvent être éliminés.
- Dans notre exemple, l'enveloppe convexe est formée par les courbes de M3 et M2.

- ✓ **M1** est dominé par tous les modèles, il peut être éliminé.
- ✓ M4 peut être meilleur que **M3** dans certains cas, mais dans ces cas là, il sera moins bon que **M2** : **M4** peut être éliminé.



Ce qu'on doit retenir

Dans de nombreuses applications, la courbe ROC fournit des informations plus intéressantes sur la qualité de l'apprentissage que le simple taux d'erreur.



C'est surtout vrai lorsque les classes sont très déséquilibrées, et lorsque le coût de mauvaise affectation est susceptible de modifications.



Il faut néanmoins que l'on ait une classe cible (positive) clairement identifiée et que la méthode d'apprentissage puisse fournir un SCORE proportionnel à $P(Y=+/X)$.

Classification

Arbre de Décision – Forêt aléatoire

Séparateur à Vaste Marge

k Plus Proches Voisins kNN

Réseaux de Neurones – Perceptrons Simples et Multicouches

Arbre de Décision

Exemple – information qualitative

client	<i>M</i>	<i>A</i>	<i>R</i>	<i>E</i>	<i>I</i>
1	moyen	moyen	village	oui	oui
2	élevé	moyen	bourg	non	non
3	faible	âgé	bourg	non	non
4	faible	moyen	bourg	oui	oui
5	moyen	jeune	ville	oui	oui
6	élevé	âgé	ville	oui	non
7	moyen	âgé	ville	oui	non
8	faible	moyen	village	non	non

Une banque dispose des informations suivantes sur un ensemble de clients:

M : moyenne des montants sur le compte client.

A : tranche d'âge du client.

R : localité de résidence du client.

E : valeur oui si le client a un niveau d'études supérieures.

I : classe oui correspond à un client qui effectue une consultation de ses comptes bancaires en utilisant Internet

Quelle est la variable à mettre comme racine de l'arbre?

M ?

A ?

R ?

E ?



Procédure *construire-arbre(X)*

SI tous les individus I appartiennent à la même modalité de la variable décisionnelle

ALORS créer un nœud feuille portant le nom de cette classe : Décision

SINON

- ✓ choisir le meilleur attribut pour créer un nœud // l'attribut qui sépare le mieux
- ✓ le test associé à ce nœud sépare X en des branches : $X_d \dots \dots \dots X_g$
 - ✓ *construire-arbre(X_d)*
 - ...
 - ...
 - ...
 - ✓ *construire-arbre(X_g)*

FIN

choix du meilleur attribut pour créer un nœud

- Il existe plusieurs méthodes pour choisir le meilleur attribut à placer dans un nœud :
 - ✓ Algorithme C4.5, C5.0
 - ✓ CHAID Chi-squared Automatic Interaction Detector
 - ✓ ID3 entropie de Shannon
 - ✓ **CART Classification and regression trees : Indice de GINI**
- **l'indice de GINI est le meilleur moyen pour la construction de l'arbre car il est le seul indice qui répond aux questions suivantes :**
 - ✓ Comment choisir la variable à segmenter parmi les variables explicatives disponibles ?
 - ✓ Lorsque la variable est continue, comment déterminer le seuil de coupe ?
 - ✓ Comment déterminer la bonne taille de l'arbre ?

Algorithme de CART

- ✓ Parmi les plus performants et plus répandus

- ✓ Accepte tout type de variables

- ✓ Utilise le **Critère de séparation : Indice de Gini** $I = 1 - \sum_i^n f_i^2$

Avec n : nombre de classes à prédire

f_i : fréquence de la classe dans le nœud

- ✓ *Plus l'indice de Gini est bas, plus le nœud est pure*

- ✓ En séparant 1 nœud en 2 nœuds fils on cherche la plus grande hausse de la pureté

- ✓ La variable la plus discriminante doit maximiser

$$IG(\text{avant séparation}) - [IG(\text{fils}_1) + \dots + IG(\text{fils}_n)]$$

Calcul de l'indice de gini

- Indice de Gini avant séparation au NIVEAU DE LA RACINE :

$\left\{ \begin{array}{l} I=\text{oui} : 3 \text{ clients} \\ I=\text{non} : 5 \text{ clients} \end{array} \right.$

8 clients

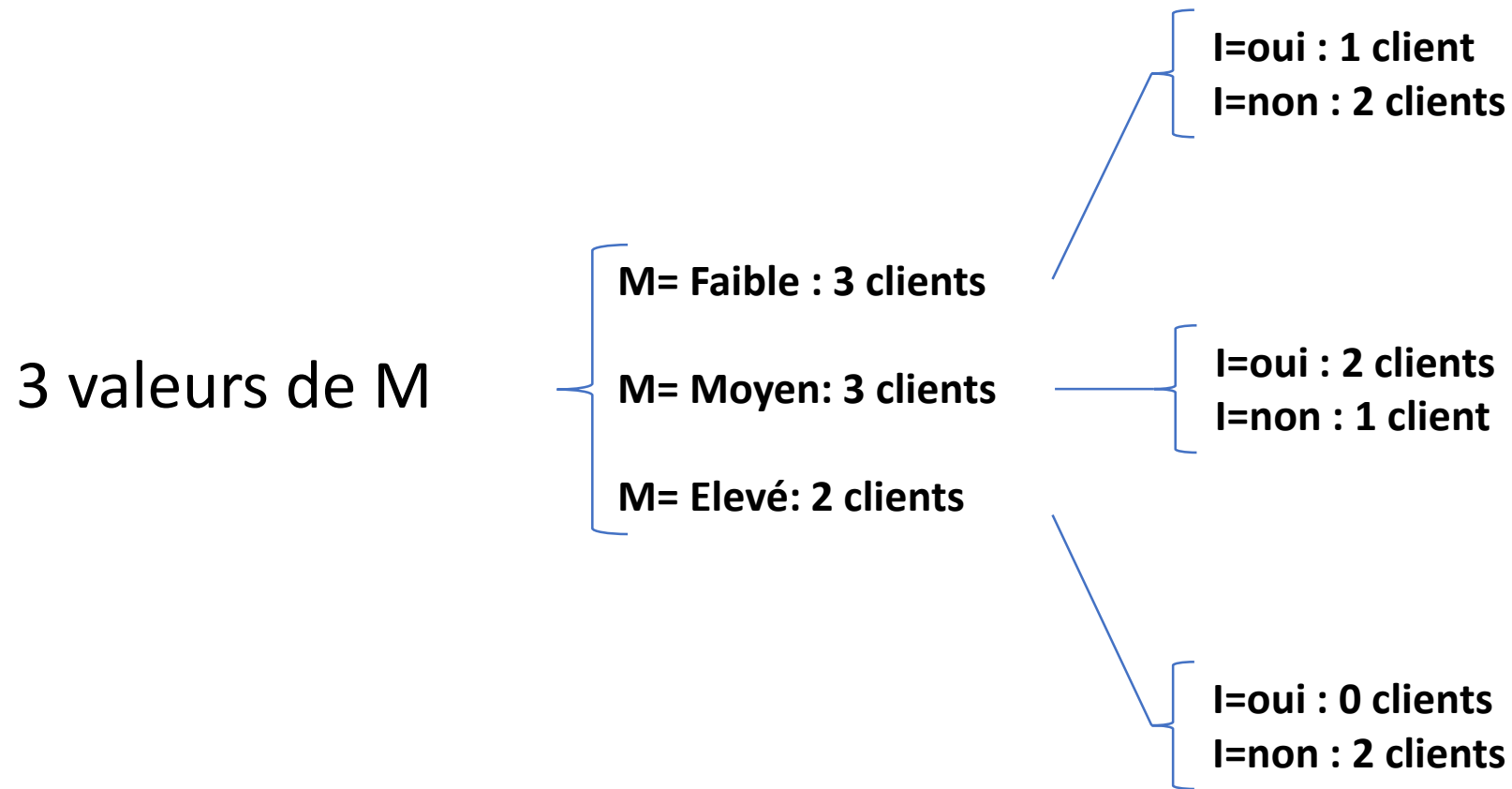
$$IG(\text{avant séparation}) = 1 - ((3/8)^2 + (5/8)^2) = 0.46875$$

↙
Fréquence des I
= oui

↘
Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de la variable M (Moyenne des montants sur le compte client):



Calcul de l'indice de gini

Indice de Gini de fils **M = Faible** :

3 clients $\left\{ \begin{array}{l} I=\text{oui} : 1 \text{ client} \\ I=\text{non} : 2 \text{ clients} \end{array} \right.$

$$IG(M=\text{Faible}) = 1 - ((1/3)^2 + (2/3)^2) = \mathbf{0.4444444}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **M = Moyen** :

3 clients { I=oui : 2 clients
I=non : 1 client

$$IG(M=Moyen) = 1 - ((2/3)^2 + (1/3)^2) = \mathbf{0.4444444}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **M = Elevé** :

2 clients {
I=oui : 0 clients
I=non : 2 clients

$$IG(M=Elevé) = 1 - ((0/2)^2 + (2/2)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de M:

$$\text{IG}(\text{avant s\u00e9paration}) - [\text{IG}(\text{M=Faible}) + \text{IG}(\text{M=Moyen}) + \text{IG}(\text{M=Elev\u00e9})]$$

=

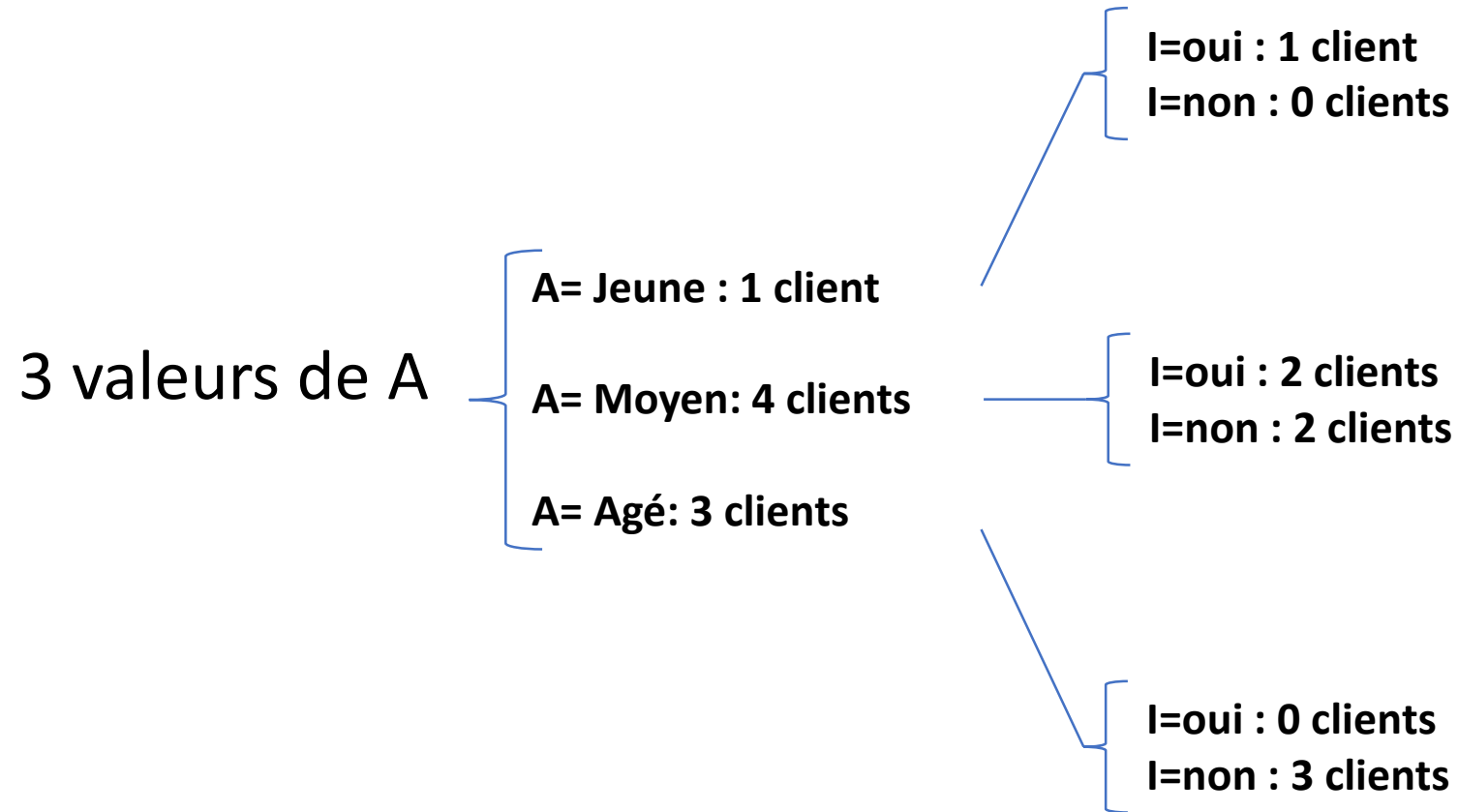
$$0.46875 - [0.4444444 + 0.4444444 + 0]$$

=

-0.4201388

Calcul de l'indice de gini

Indice de Gini de la variable A (Tranche d'âge du client):



Calcul de l'indice de gini

Indice de Gini de fils **A = Jeune** :

1 client {
I=ooui : 1 client
I=non : 0 clients

$$IG(A=Jeune) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **A = Moyen** :

4 clients {
I=oui : 2 clients
I=non : 2 clients

$$IG(A=Moyen) = 1 - ((2/4)^2 + (2/4)^2) = \mathbf{0.5}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **A = Agé** :

3 clients {
I=oui : 0 clients
I=non : 3 clients

$$IG(A=Agé) = 1 - ((0/3)^2 + (3/3)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de A:

$$\text{IG}(\text{avant s\'eparation}) - [\text{IG}(\text{A=Jeune}) + \text{IG}(\text{A=Moyen}) + \text{IG}(\text{A=Ag\'e})]$$

=

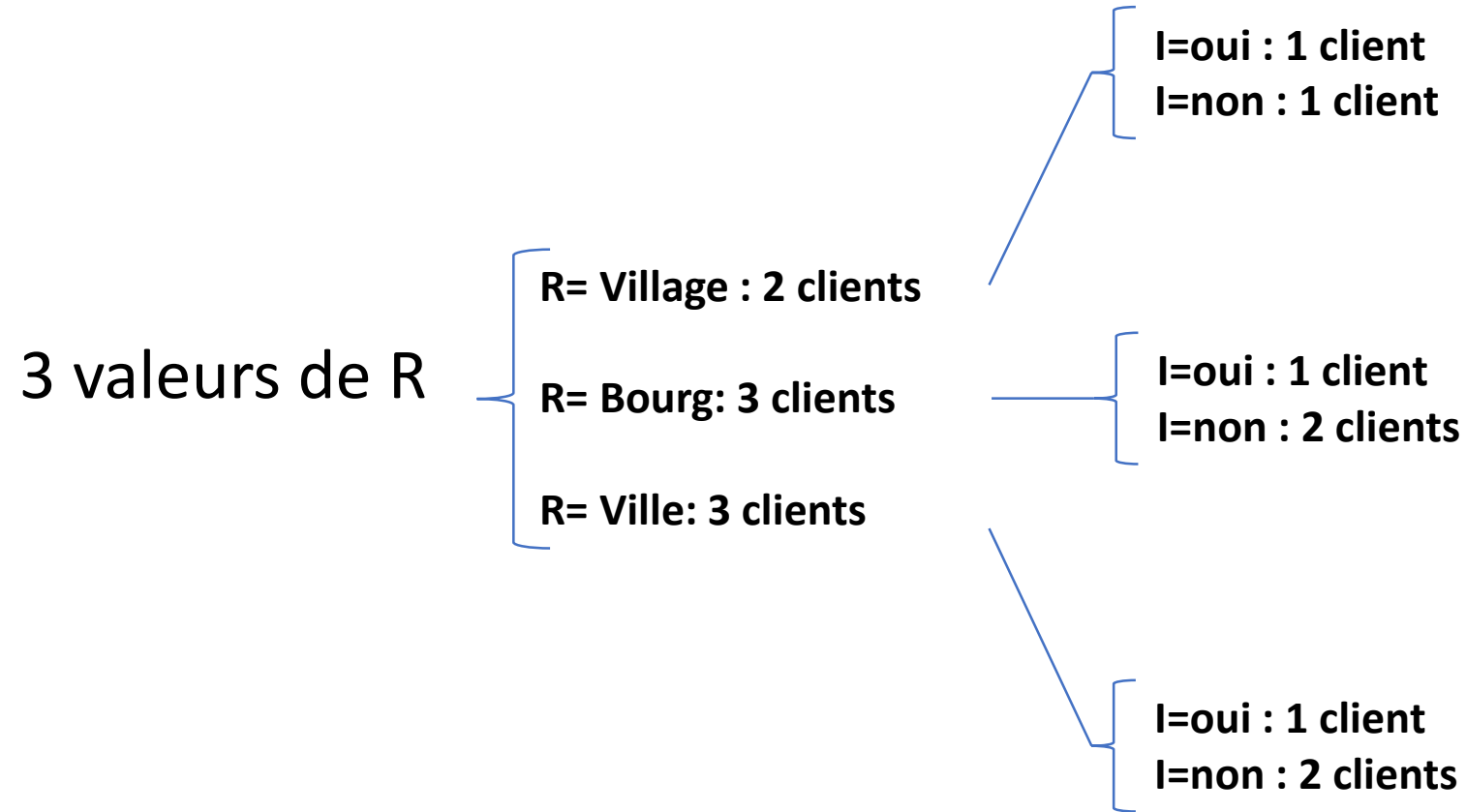
$$0.46875 - [0 + 0.5 + 0]$$

=

-0.03125

Calcul de l'indice de gini

Indice de Gini de la variable R(Localité de résidence du client):



Calcul de l'indice de gini

Indice de Gini de fils **R= Village** :

2 clients {
I=oui : 1 client
I=non : 1 client

$$IG(R= Village) = 1 - ((1/2)^2 + (1/2)^2) = \mathbf{0.5}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **R= Bourg** :

3 clients {
I=oui : 1 client
I=non : 2 clients

$$IG(R= Bourg) = 1 - ((1/3)^2 + (2/3)^2) = \mathbf{0.4444444}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **R= Ville**:

3 clients {
I=oui : 1 client
I=non : 2 clients

$$IG(R=Ville) = 1 - ((1/3)^2 + (2/3)^2) = \mathbf{0.4444444}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de R:

$$\text{IG}(\text{avant s\u00e9paration}) - [\text{IG}(\text{R=Village}) + \text{IG}(\text{R=Bourg}) + \text{IG}(\text{R=Ville})]$$

=

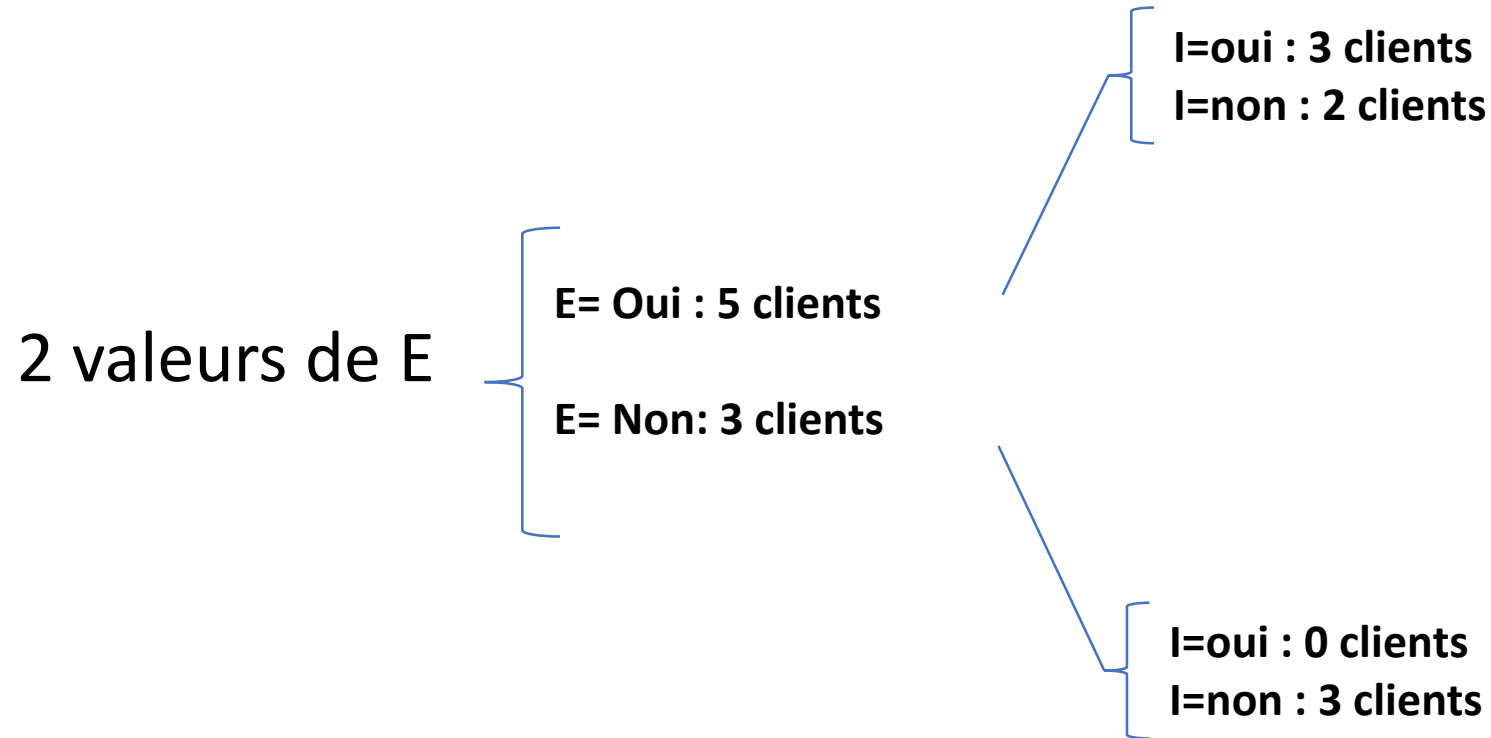
$$0.46875 - [0.4444444 + 0.5 + 0.4444444]$$

=

-0.9201388

Calcul de l'indice de gini

Indice de Gini de la variable E(Niveau d'études du client):



Calcul de l'indice de gini

Indice de Gini de fils **E= Oui** :

5 clients {
I=oui : 3 clients
I=non : 2 clients

$$IG(E=Oui) = 1 - ((3/5)^2 + (2/5)^2) = \mathbf{0.48}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **E= Non** :

3 clients {
I=oui : 0 clients
I=non : 3 clients

$$IG(E=Non) = 1 - ((0/3)^2 + (3/3)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de E:

$$\text{IG}(\text{avant s\u00e9paration}) - [\text{IG}(E=\text{Oui}) + \text{IG}(E=\text{Non})]$$

=

$$0.46875 - [0.48 + 0]$$

=

$$\boxed{-0.01125388}$$

PREMIER RESULTAT DE l'indice de gini

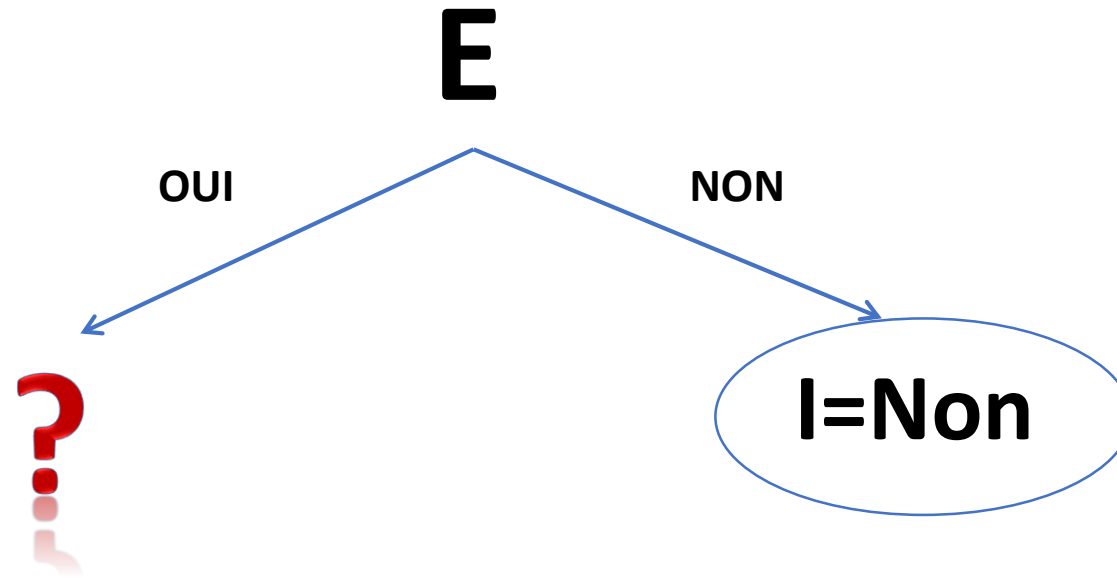
La variable la plus séparatrice est celle qui maximise :

$$IG(\text{avant séparation}) - [IG(\text{fils}_1) + IG(\text{fils}_2) + \dots + IG(\text{fils}_n)]$$



E

Construction de l'arbre



CALCUL DE L'INDICE DE GINI : E=OUI

Indice de Gini avant séparation avec E = Oui :

5 clients {
I=ooui : 3 clients
I=non : 2 clients

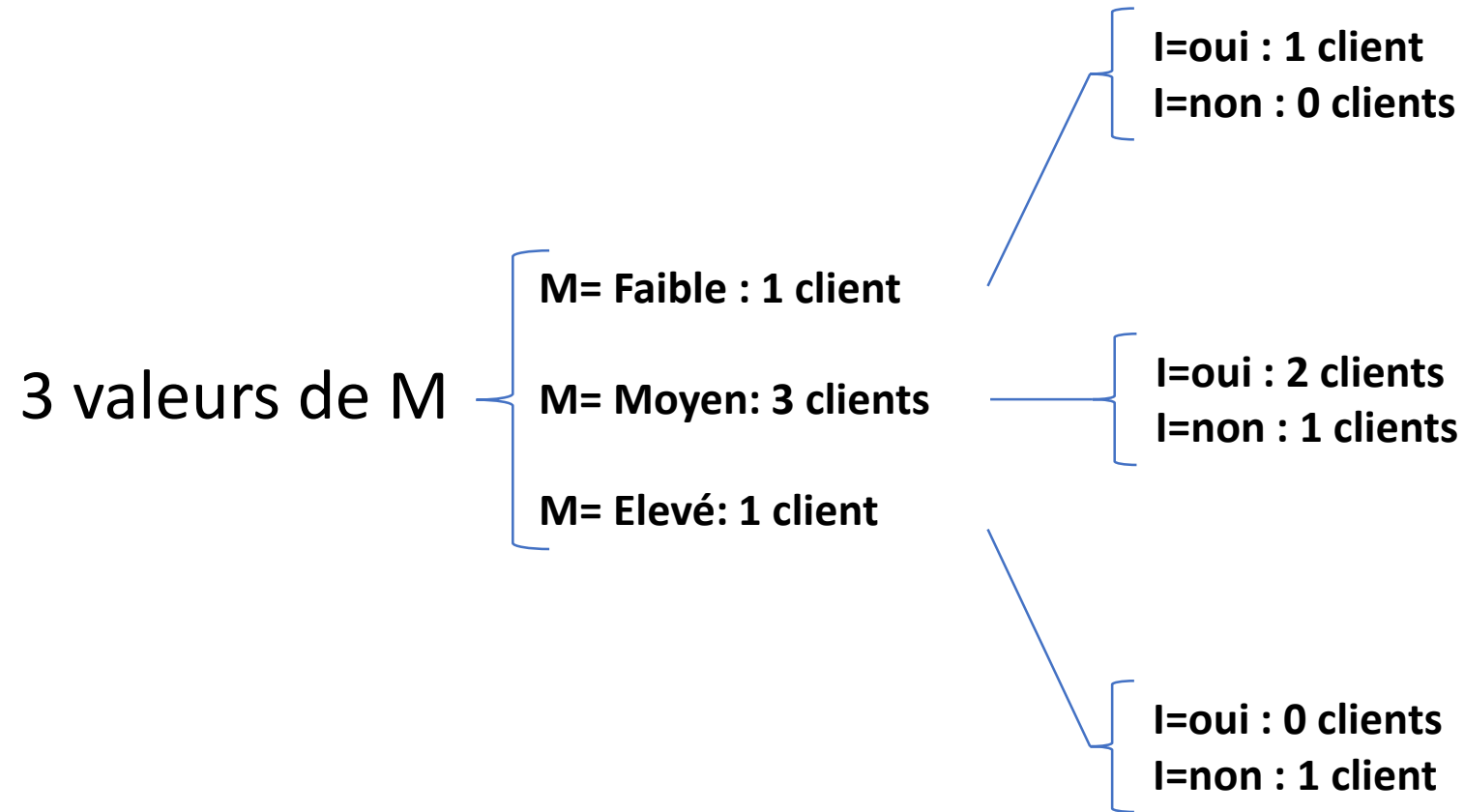
$$IG(\text{avant séparation}_1) = 1 - ((3/5)^2 + (2/5)^2) = 0.48$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de la variable M (Moyenne des montants sur le compte client)
avec **E=Oui**:



Calcul de l'indice de gini

Indice de Gini de fils **M = Faible & E = Oui** :

1 client {
I=oui : 1 client
I=non : 0 clients

$$IG(M=Faible \& E=Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **M = Moyen & E = Oui** :

3 clients {
I=oui : 2 clients
I=non : 1 client

$$IG(M=Moyen \& E=Oui) = 1 - ((2/3)^2 + (1/3)^2) = \mathbf{0.4444444}$$

↙
Fréquence des I
= oui

↘
Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **M = Elevé & E = Oui:**

1 client {
I=oui : 0 clients
I=non : 1 client

$$IG(M=Elevé \& E=Oui) = 1 - ((0/1)^2 + (1/1)^2) = \mathbf{0}$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de M avec E=Oui :

$$\text{IG}(\text{avant s\'eparation}_1) - [\text{IG}(M=\text{Faible}) + \text{IG}(M=\text{Moyen}) + \text{IG}(M=\text{Elev\'e})]$$

=

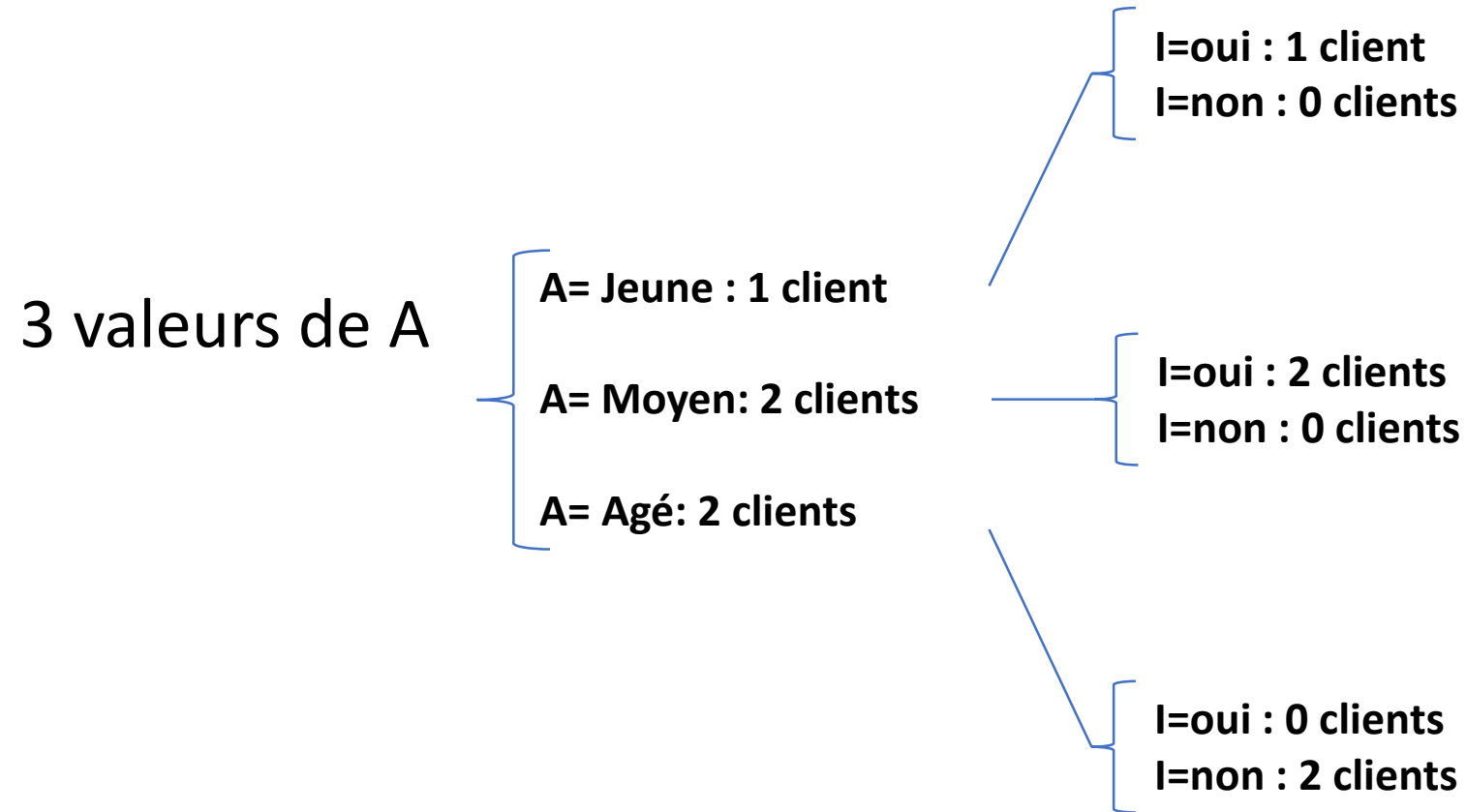
$$0.48 - [0 + 0.4444444 + 0]$$

=

0.0355556

Calcul de l'indice de gini

Indice de Gini de la variable A (Tranche d'âge du client) avec **E=Oui** :



Calcul de l'indice de gini

Indice de Gini de fils **A = Jeune & E = Oui :**

1 client {
I=oui : 1 client
I=non : 0 clients

$$IG(A=Jeune \& E = Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **A = Moyen & E = Oui** :

2 clients {
I=oui : 2 clients
I=non : 0 clients

$$IG(A=Moyen \& E = Oui) = 1 - ((2/2)^2 + (0/2)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **A = Agé & E = Oui** :

2 clients {
I=oui : 0 clients
I=non : 2 clients

$$IG(A=Agé \& E = Oui) = 1 - ((0/2)^2 + (2/2)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de A avec E=Oui:

$$IG(\text{avant s\u00e9paration}_1) - [IG(A=\text{Jeune}) + IG(A=\text{Moyen}) + IG(A=\text{Ag\u00e9})]$$

=

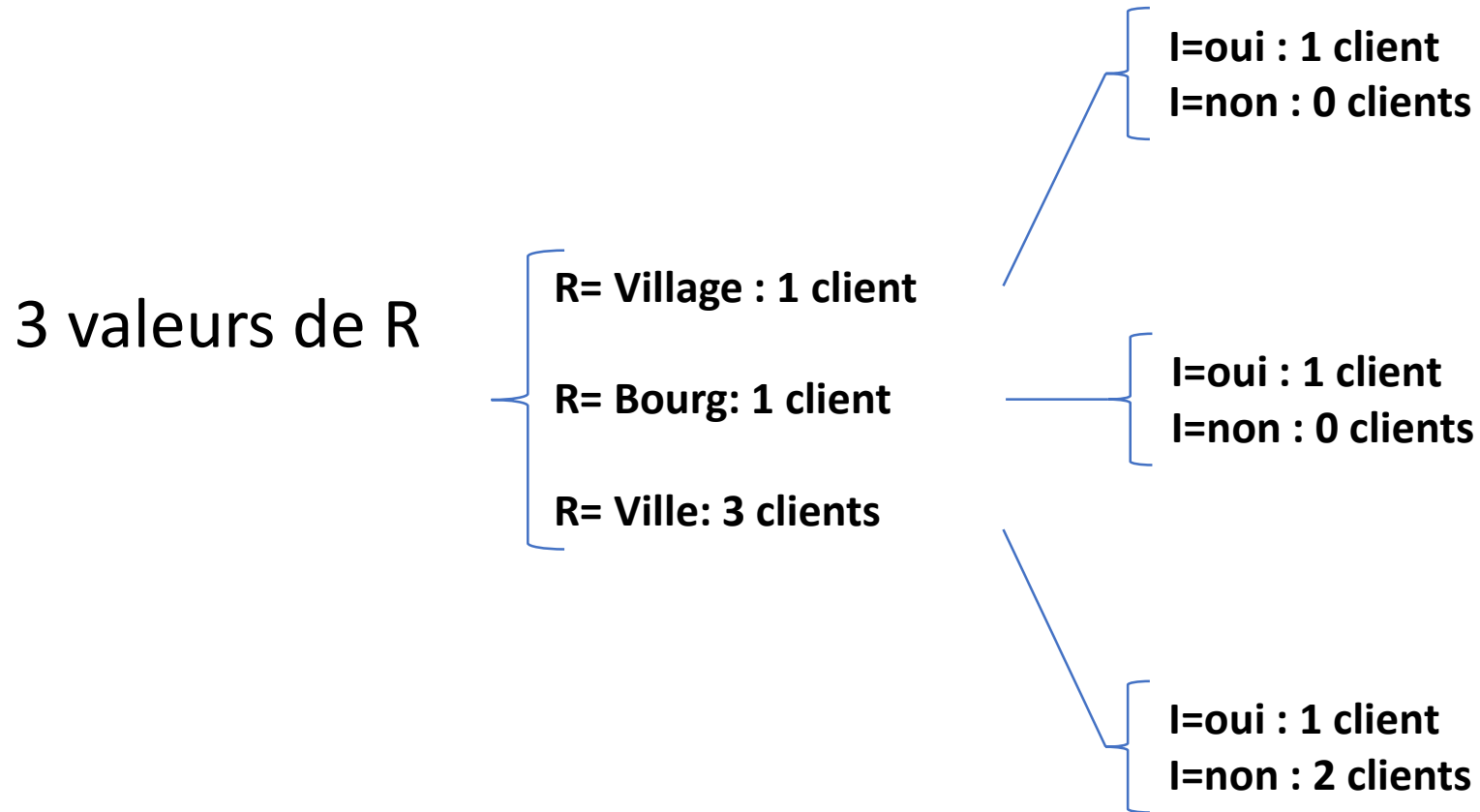
$$0.48 - [0 + 0 + 0]$$

=

0.48

Calcul de l'indice de gini

Indice de Gini de la variable R(Localité de résidence du client) avec **E=Oui** :



Calcul de l'indice de gini

Indice de Gini de fils **R= Village & E = Oui :**

1 clients {
I=oui : 1 client
I=non : 0 clients

$$IG(R= Village \& E = Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **R= Bourg & E = Oui :**

1 client {
I=oui : 1 client
I=non : 0 clients

$$IG(R= Bourg \& E = Oui) = 1 - ((1/1)^2 + (0/1)^2) = 0$$

Fréquence des I
= oui

Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de fils **R= Ville & E = Oui** :

3 clients {
I=oui : 1 client
I=non : 2 clients

$$IG(R=Ville \& E = Oui) = 1 - ((1/3)^2 + (2/3)^2) = \mathbf{0.4444444}$$

↙
Fréquence des I
= oui

↘
Fréquence des I
= non

Calcul de l'indice de gini

Indice de Gini de R avec E=Oui: :

$$IG(\text{avant s\u00e9paration}_1) - [IG(R=\text{Village}) + IG(R=\text{Bourg}) + IG(R=\text{Ville})]$$

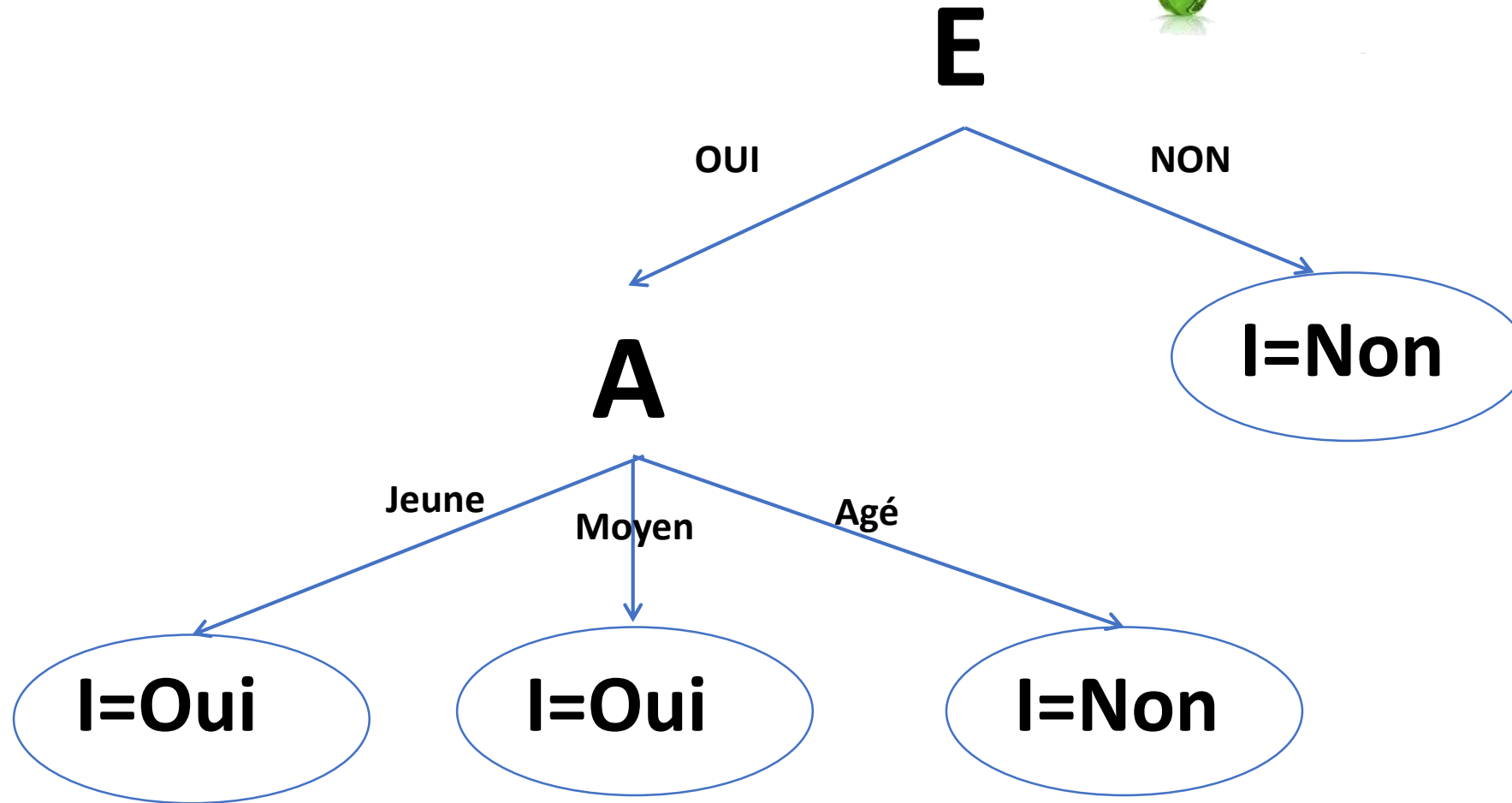
=

$$0.48 - [0 + 0 + 0.4444444]$$

=

0.0355556

l'arbre de décision



Exercices

Une banque souhaite promouvoir une offre commerciale via les adresses mails de ses clients.

Pour cela elle fait appel à vous et à vos connaissances en fouille de donnée pour sélectionner ceux qui sont potentiellement intéressés.

Trois attributs descriptifs sont à votre disposition :

- L'âge en deux tranches : [18; 35] et [36 et plus]
- Le sexe H : Homme ou F : Femme
- Propriétaire O : oui ou N : non
- L'attribut cible qui prend deux valeurs : O (intéressé) et N (pas intéressé).

Le résultat d'une enquête préliminaire sur un échantillon représentatif de clients donne :

Age	Sexe	Propriétaire	Intéressé
20	H	N	N
25	F	N	N
32	H	O	O
34	H	O	O
37	H	N	O
41	F	O	N
45	H	O	O
45	F	O	N
52	H	O	N
60	F	O	N

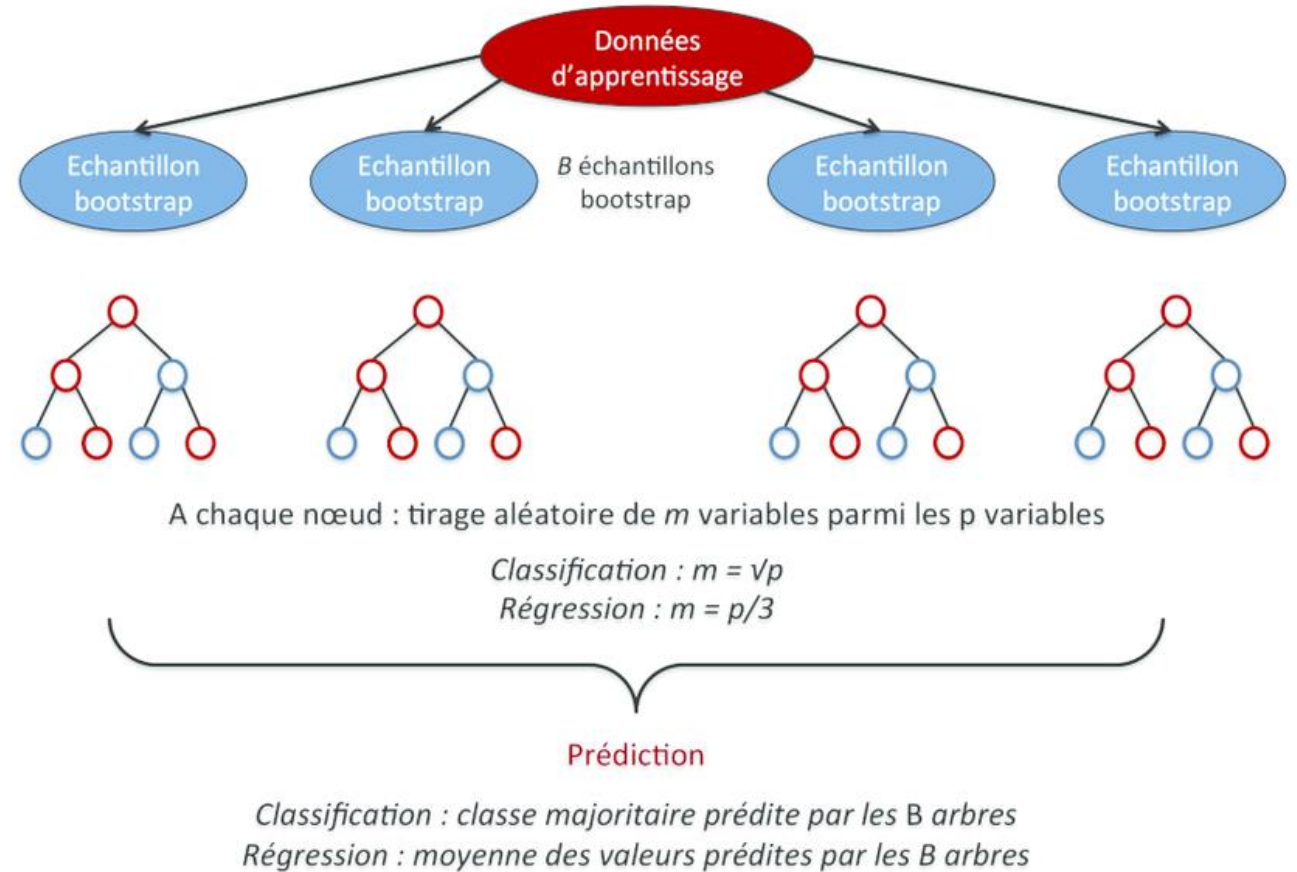
déduire la variable la plus décisive par rapport à l'appartenance d'un individu à l'origine orientale.

	Yeux	Cheveux	Taille	Oriental
1	Noir	Noir	Petit	Oui
2	Noir	Blanc	Grand	Oui
3	Noir	Blanc	Petit	Oui
4	Noir	Noir	Grand	Oui
5	Brun	Noir	Grand	Oui
6	Brun	Blanc	Petit	Oui
7	Bleu	Blond	Grand	Non
8	Bleu	Blond	Petit	Non
9	Bleu	Blanc	Grand	Non
10	Bleu	Noir	Petit	Non
11	Brun	Blond	Petit	Non

Forêt Aléatoire

Principe des RF

- Principe : l'union fait la force
- Une « forêt » = un ensemble d'arbres
- apprendre grand nombre T (\sim qlq 10aines ou 100aines) d'arbres simples
- utilisation par vote des arbres (classe majoritaire, voire probabilités des classes par % des votes) si classification, ou moyenne des arbres si régression
- Algo proposé en 2001 par Breiman & Cutter

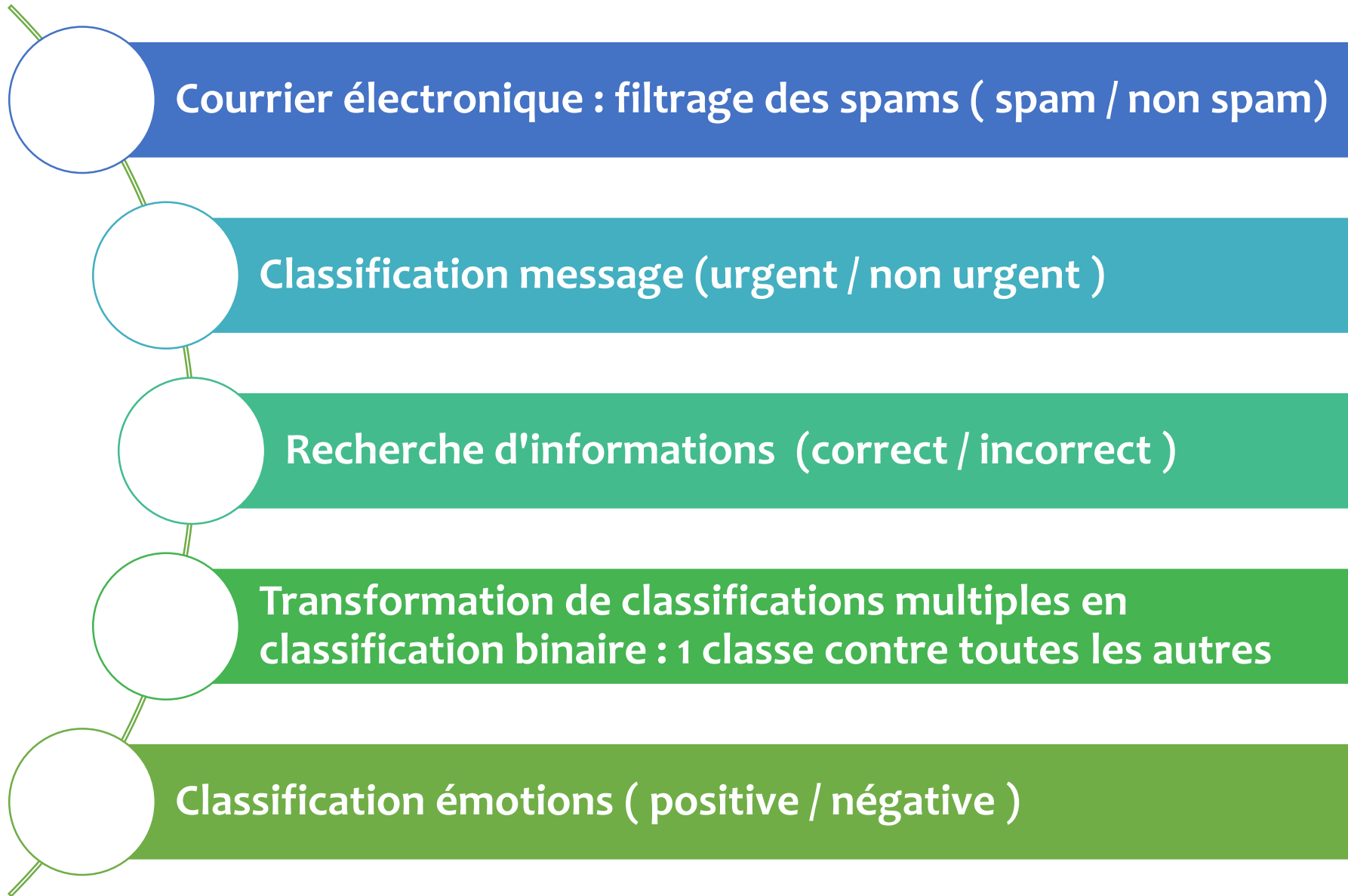


Avantages et inconvénients des RF

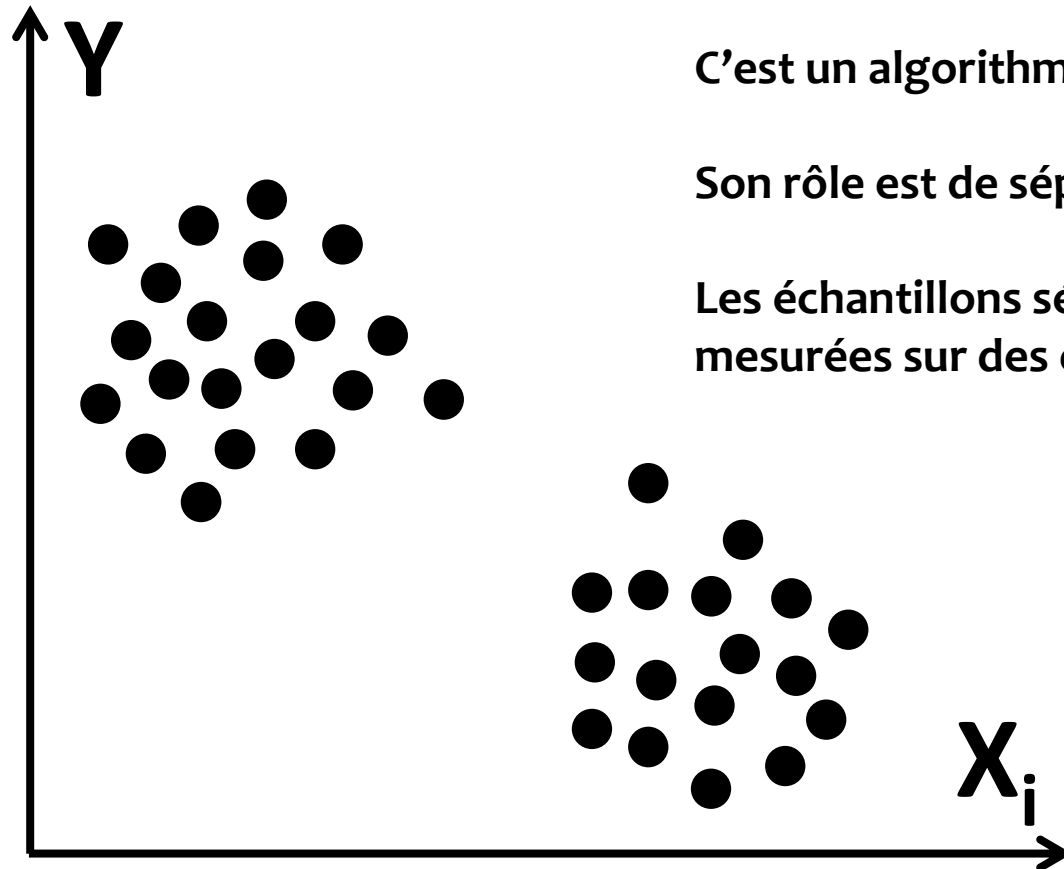
- Avantages
 - Reconnaissance TRES RAPIDE
 - Multi-classes par nature
 - Efficace sur inputs de grande dimension
 - Robustesse aux outliers
- Inconvénients
 - Apprentissage souvent long
 - Valeurs extrêmes souvent mal estimées dans cas de régression

SVM

EXEMPLE D'UTILISATION DU CLASSIFIEUR À VASTE MARGE



Qu'est-ce qu'un classifieur

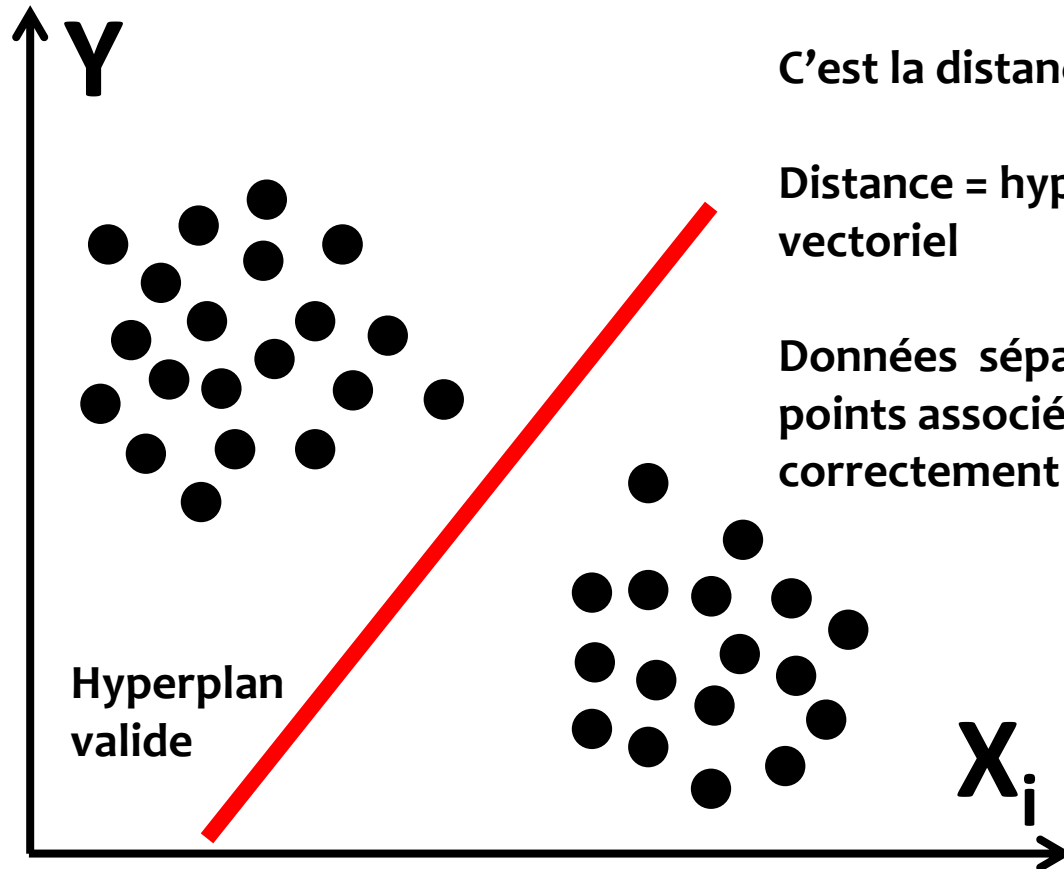


C'est un algorithmes de classement statistique.

Son rôle est de séparer des échantillons

Les échantillons séparés ont des propriétés similaires,
mesurées sur des observations

Qu'est-ce qu'une marge ?

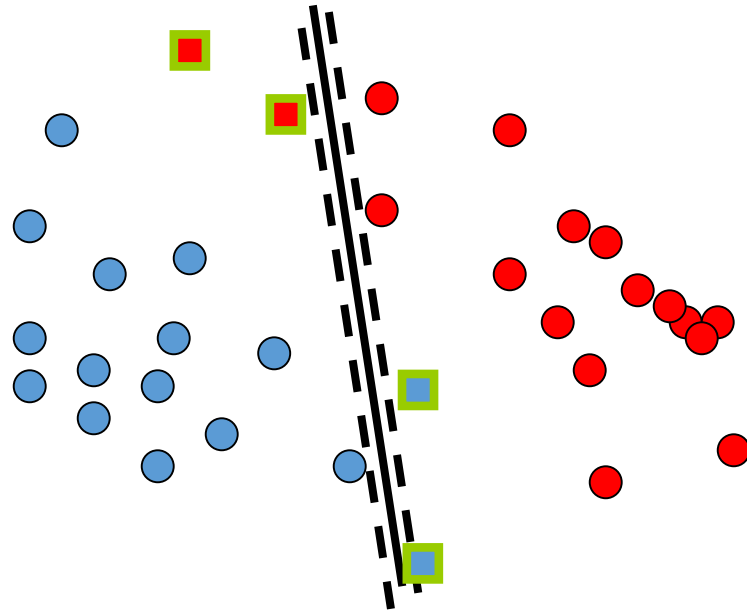


C'est la distance entre les échantillons séparés

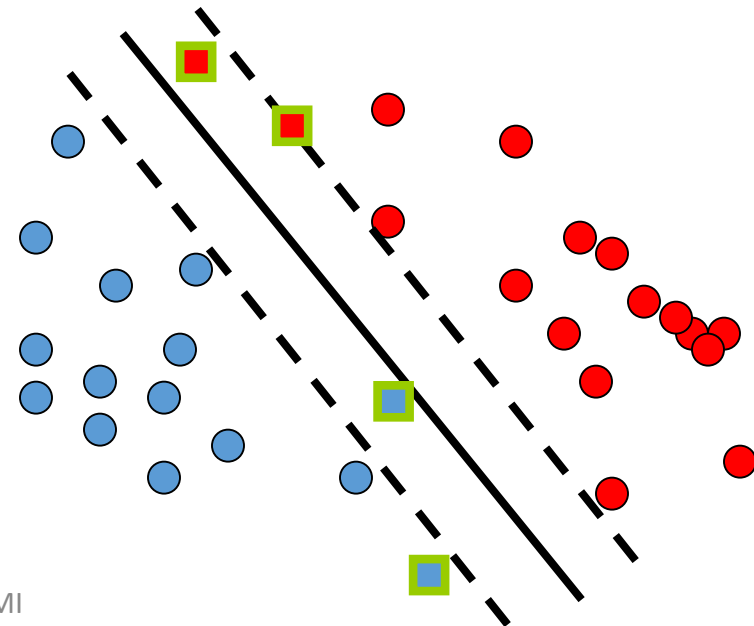
Distance = hyperplan : Plan de séparation - Espace vectoriel

Données séparables linéairement : si tous les points associés aux données peuvent être séparés correctement par une frontière linéaire

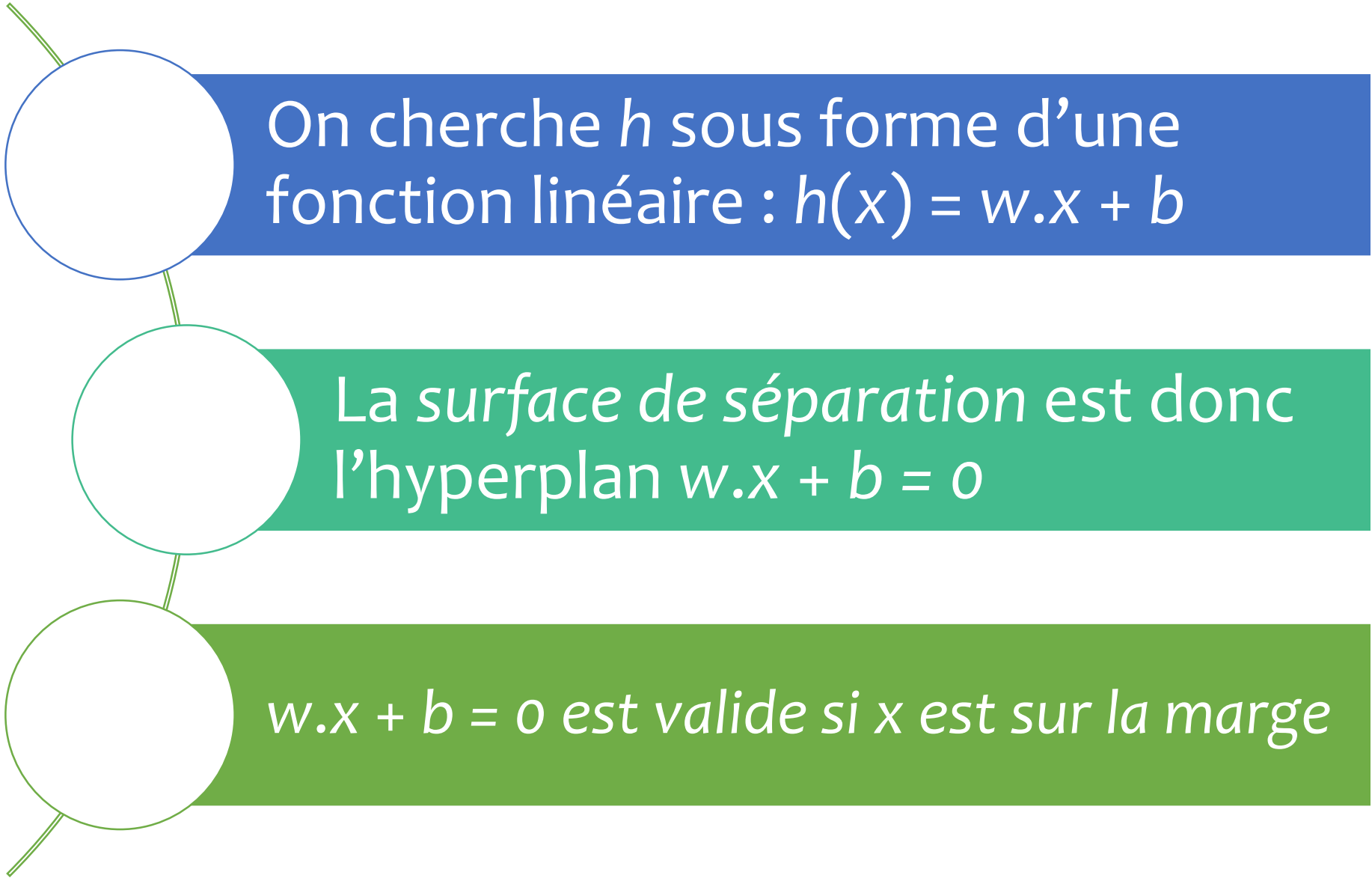
CHOIX D'UN CLASSIFIEUR À VASTE MARGE



MAUVAIS



BON

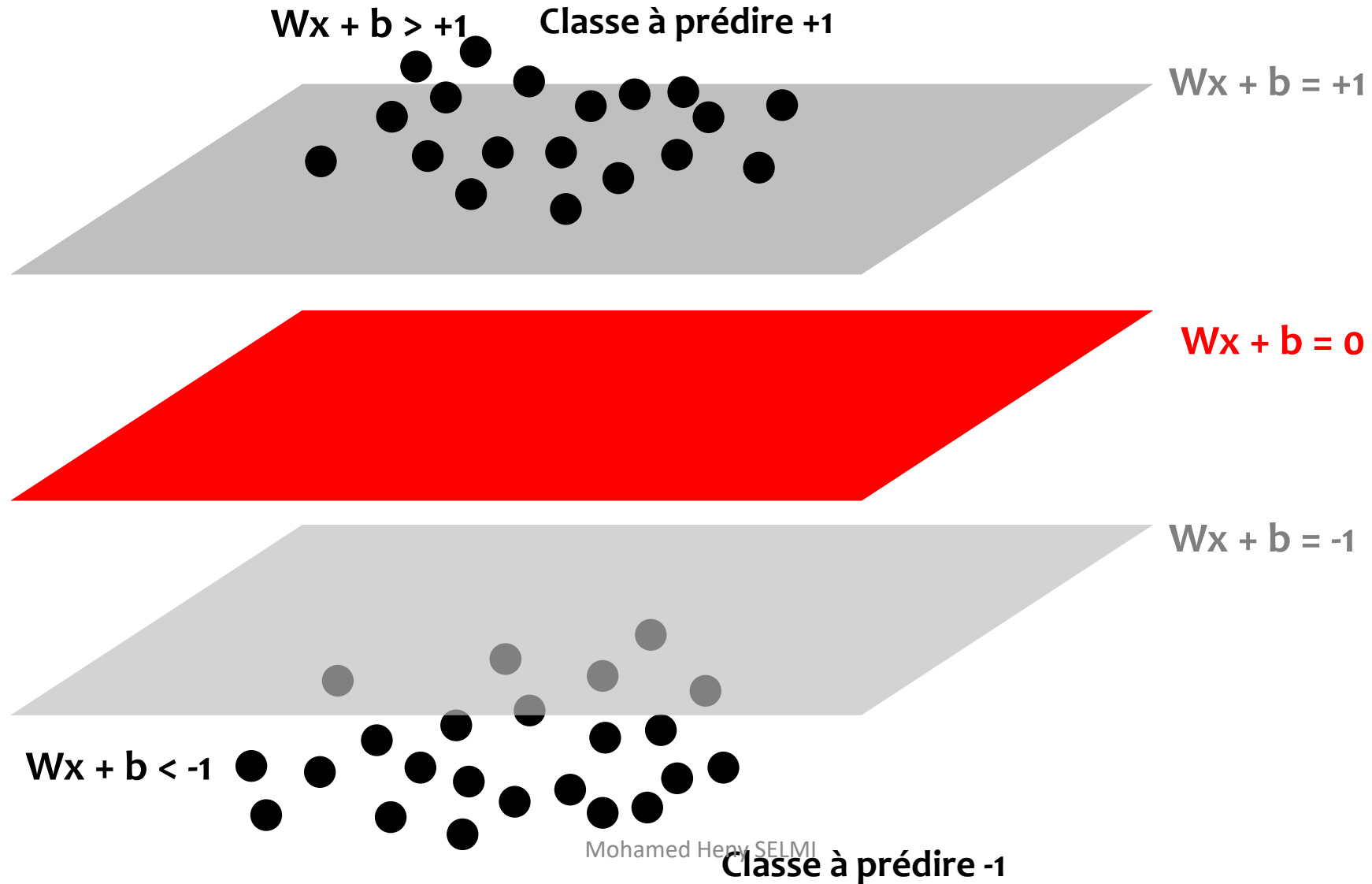


On cherche h sous forme d'une fonction linéaire : $h(x) = w.x + b$

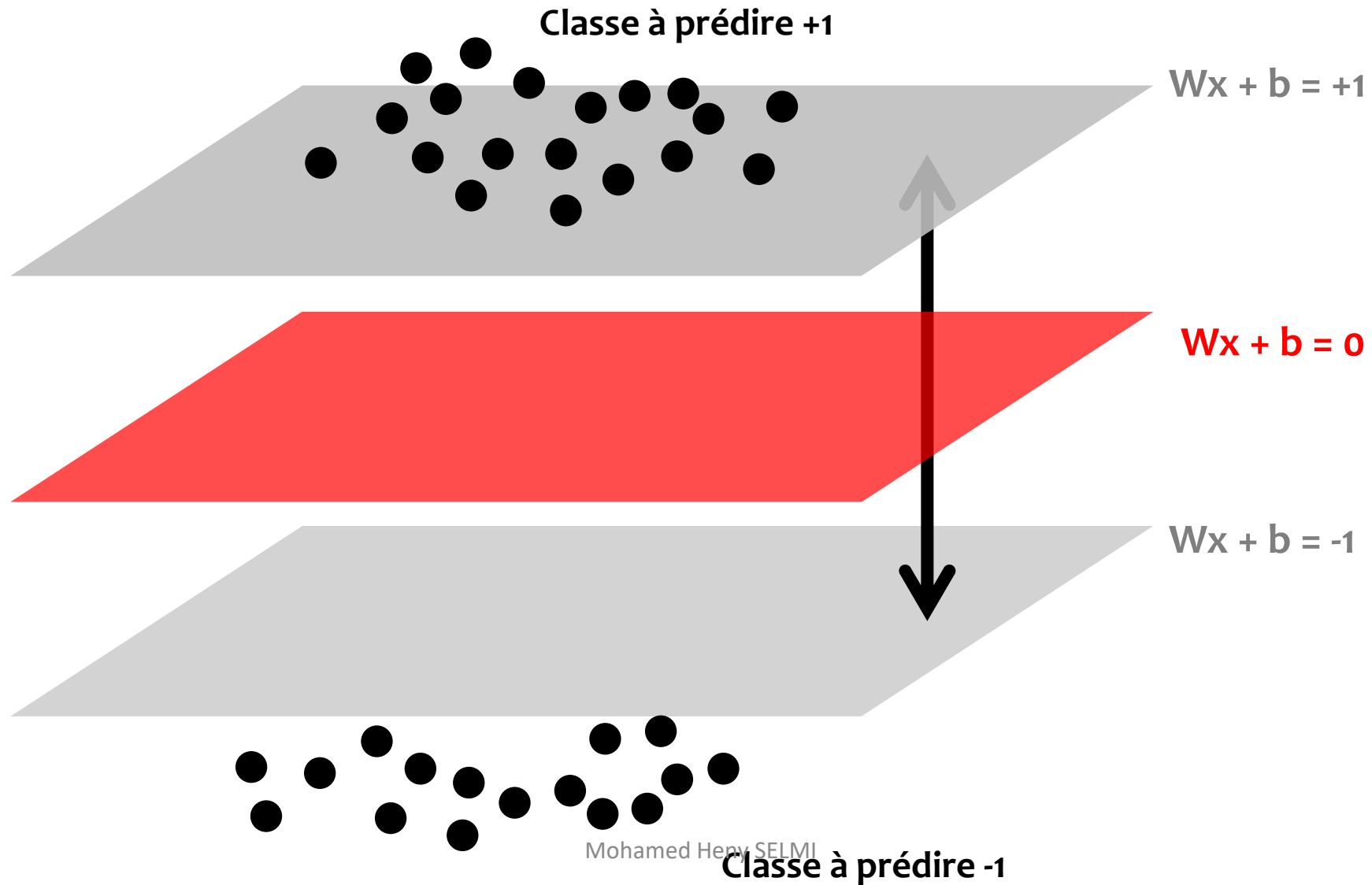
La *surface de séparation* est donc l'hyperplan $w.x + b = 0$

$w.x + b = 0$ est valide si x est sur la marge

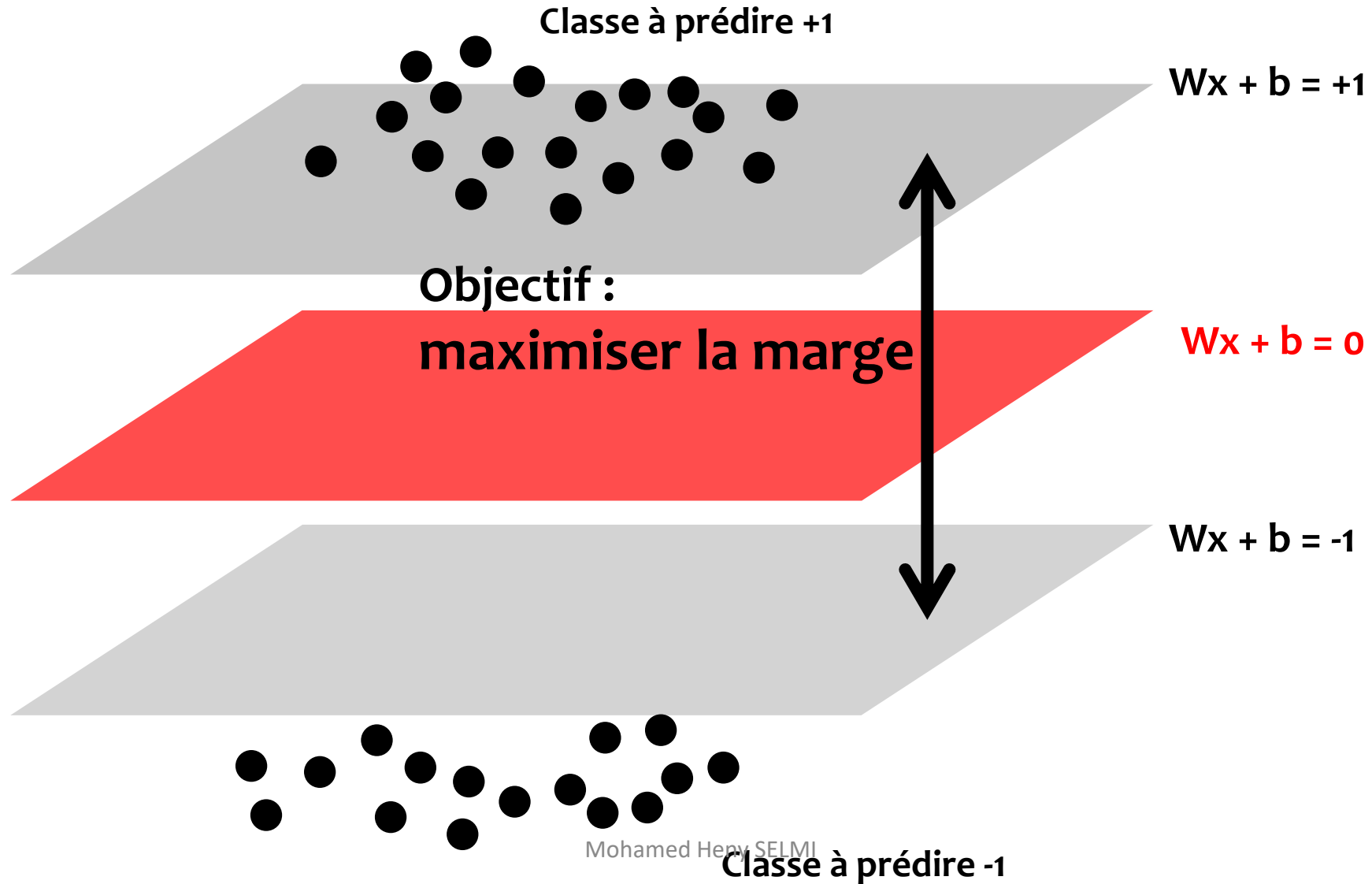
HYPERPLANS SÉPARATEURS



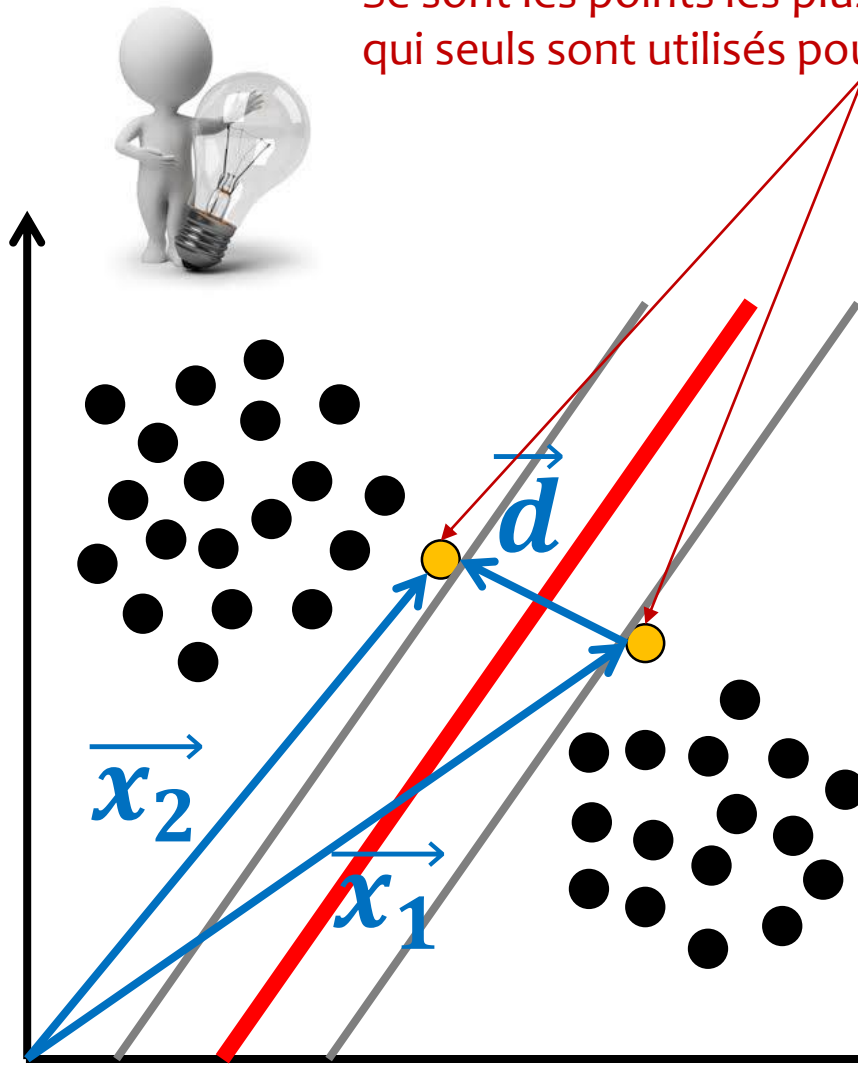
MARGE MAXIMALE ENTRE HYPERPLANS SÉPARATEURS



MARGE MAXIMALE ENTRE HYPERPLANS SÉPARATEURS



Vecteurs supports (Support vector machine):
Se sont les points les plus proches,
qui seuls sont utilisés pour la détermination de l'hyperplan



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

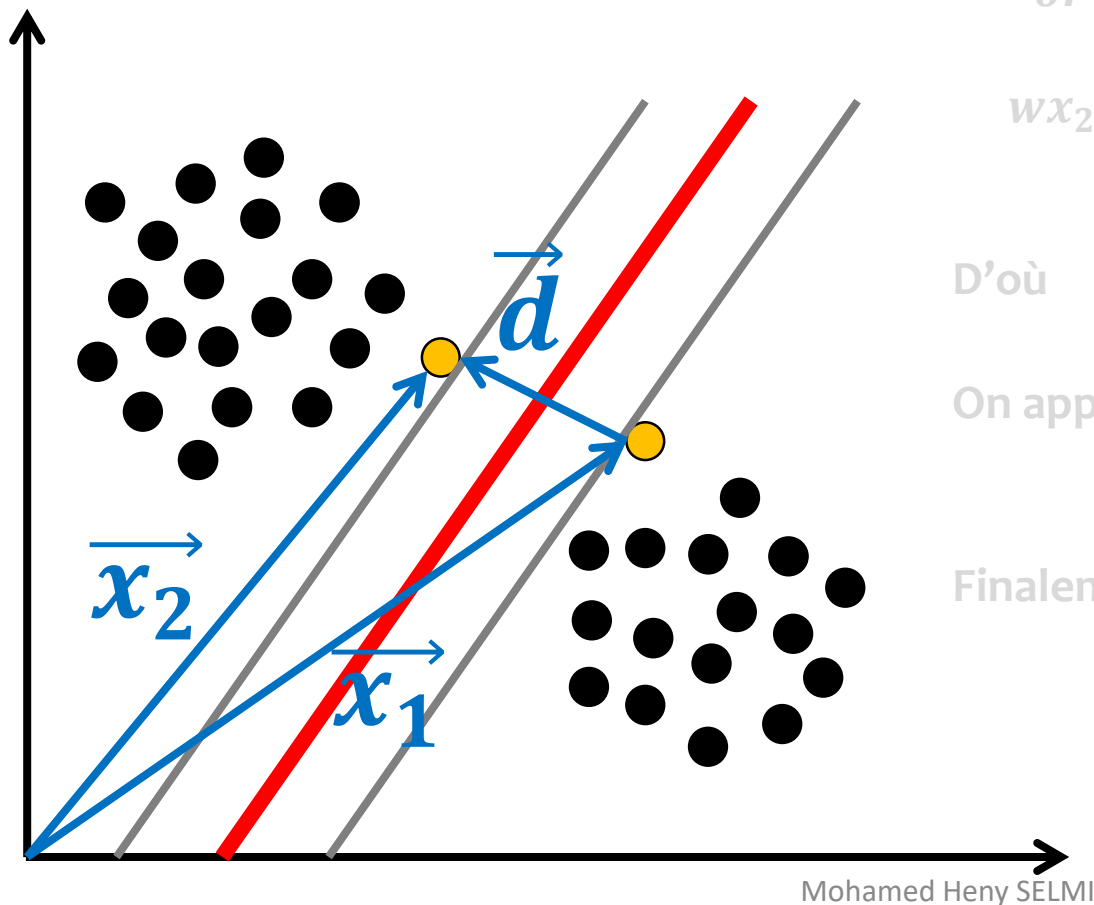
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\vec{x}_2 = \vec{x}_1 + \vec{d}$$

$$\vec{x}_2 - \vec{x}_1 = \vec{d}$$

$$w \cdot (\vec{x}_2 - \vec{x}_1) = w \cdot \vec{d}$$

$$\text{or } wx_2 + b = +1 \text{ et } wx_1 + b = -1$$

donc

$$wx_2 + b - wx_1 - b = +1 - (-1) = 2$$

$$wx_2 - wx_1 = 2$$

$$w(x_2 - x_1) = 2$$

D'où

$$2 = w \cdot \vec{d}$$

On applique la norme :

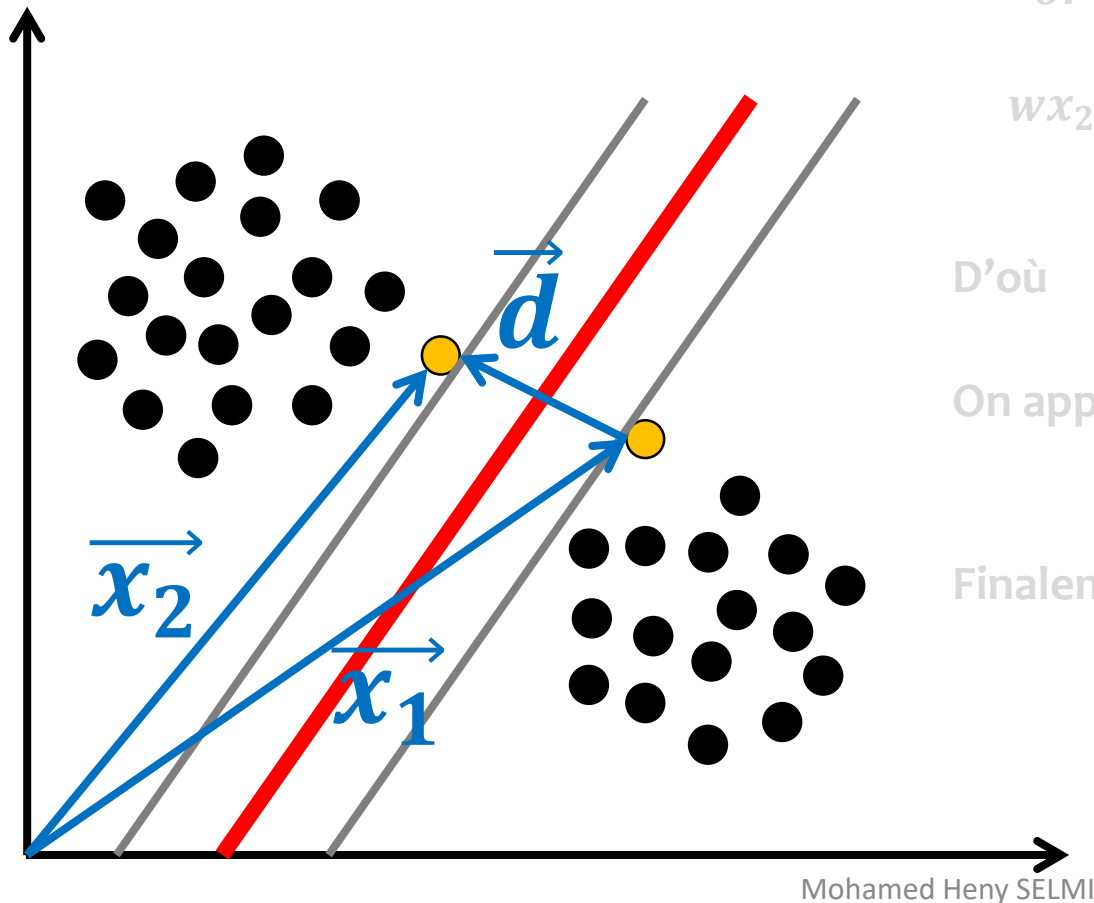
$$\|2\| = \|w\| \cdot \|\vec{d}\|$$

$$2 = \|w\| \cdot d$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\vec{x}_2 = \vec{x}_1 + \vec{d}$$

$$\vec{x}_2 - \vec{x}_1 = \vec{d}$$

$$w \cdot (\vec{x}_2 - \vec{x}_1) = w \cdot \vec{d}$$

$$\text{or } wx_2 + b = +1 \text{ et } wx_1 + b = -1$$

donc

$$wx_2 + b - wx_1 - b = +1 - (-1) = 2$$

$$wx_2 - wx_1 = 2$$

$$w(x_2 - x_1) = 2$$

D'où

$$2 = w \cdot \vec{d}$$

On applique la norme :

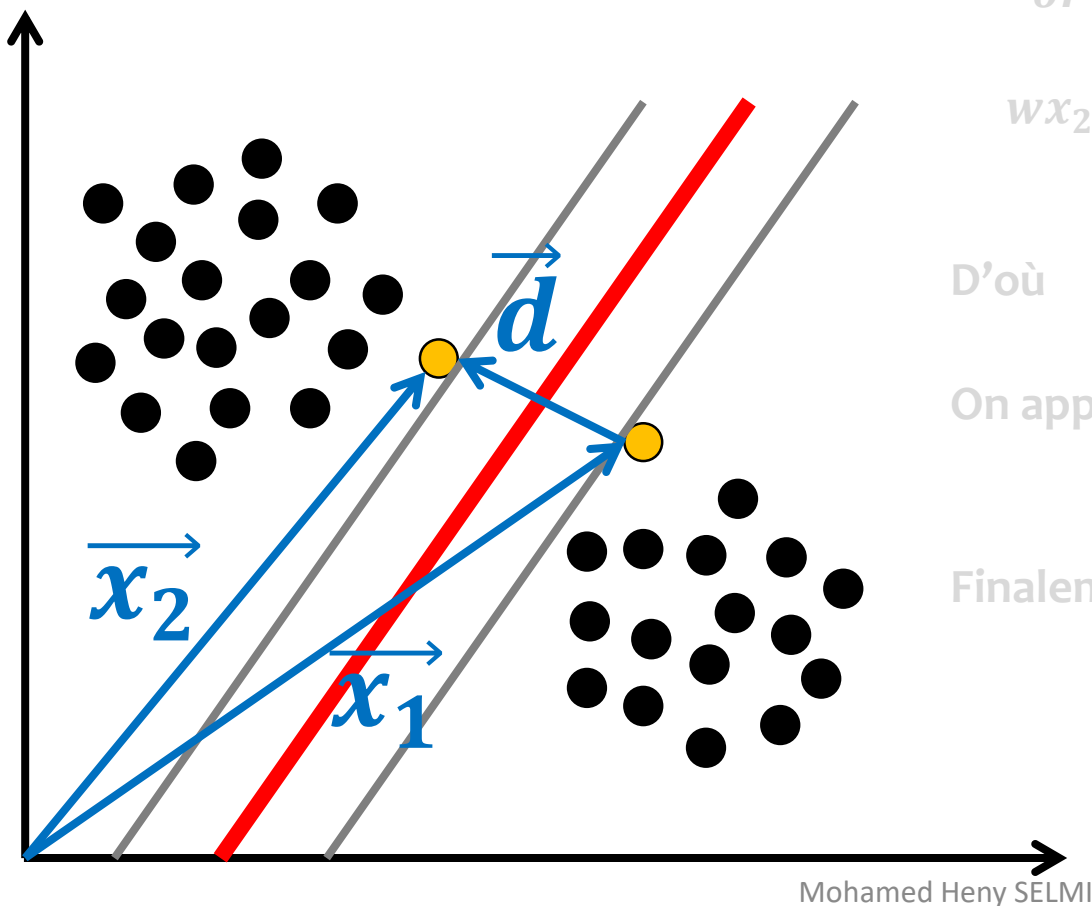
$$\|2\| = \|w\| \cdot \|\vec{d}\|$$

$$2 = \|w\| \cdot d$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\vec{x}_2 = \vec{x}_1 + \vec{d}$$

$$\vec{x}_2 - \vec{x}_1 = \vec{d}$$

$$w \cdot (\vec{x}_2 - \vec{x}_1) = w \cdot \vec{d}$$

$$\text{or } wx_2 + b = +1 \text{ et } wx_1 + b = -1$$

donc

$$wx_2 + b - wx_1 - b = +1 - (-1) = 2$$

$$wx_2 - wx_1 = 2$$

$$w(x_2 - x_1) = 2$$

D'où

$$2 = w \cdot \vec{d}$$

On applique la norme :

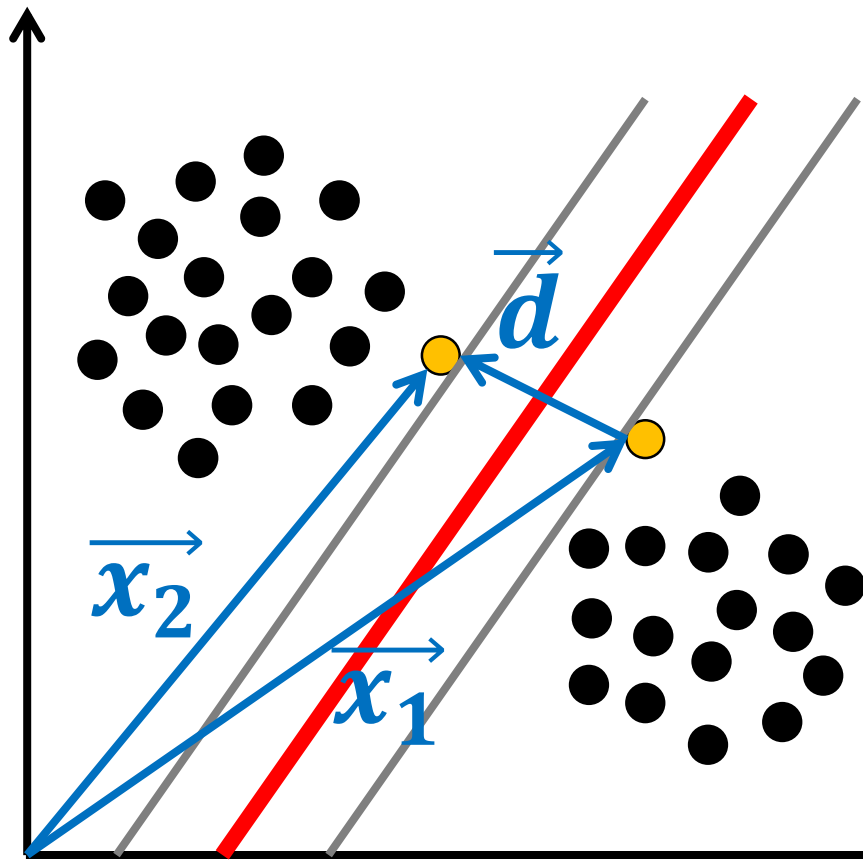
$$\|2\| = \|w\| \cdot \|\vec{d}\|$$

$$2 = \|w\| \cdot d$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc}\end{aligned}$$

$$\begin{aligned}wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

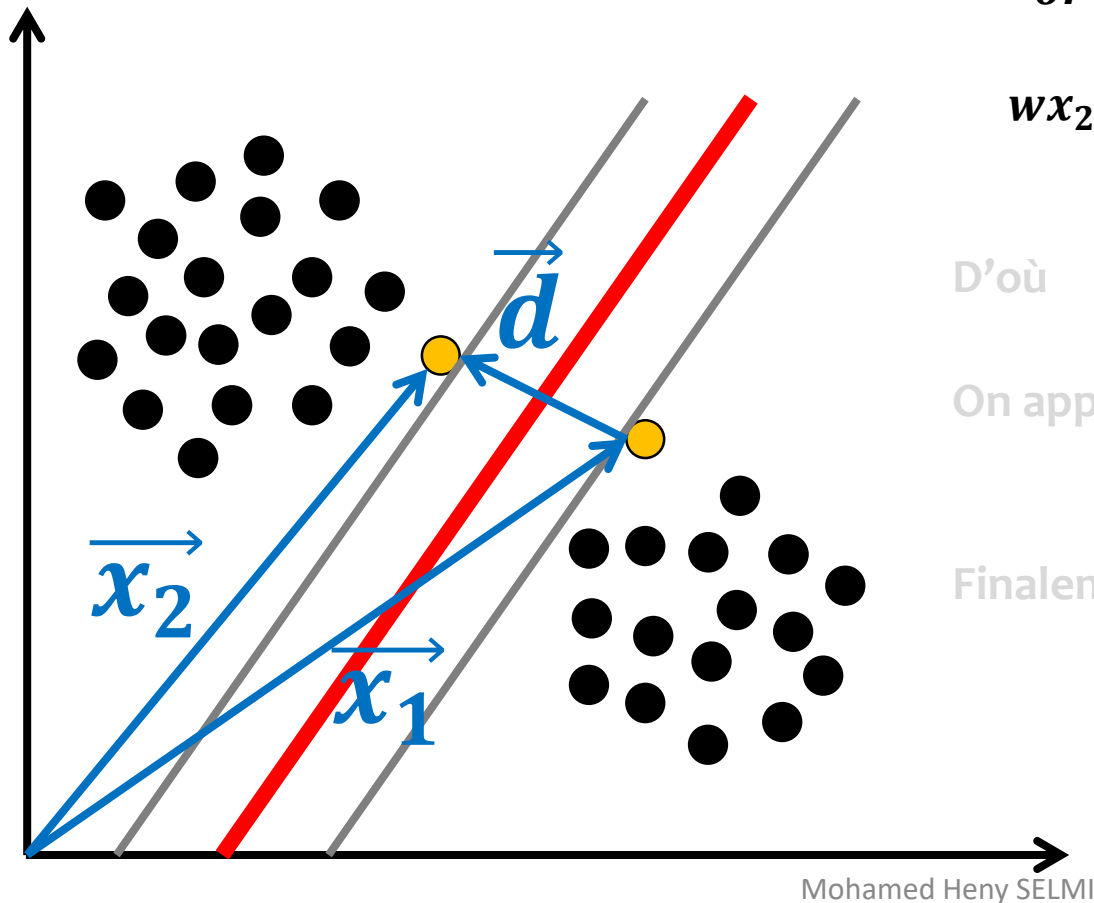
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + \cancel{b} - wx_1 - \cancel{b} &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

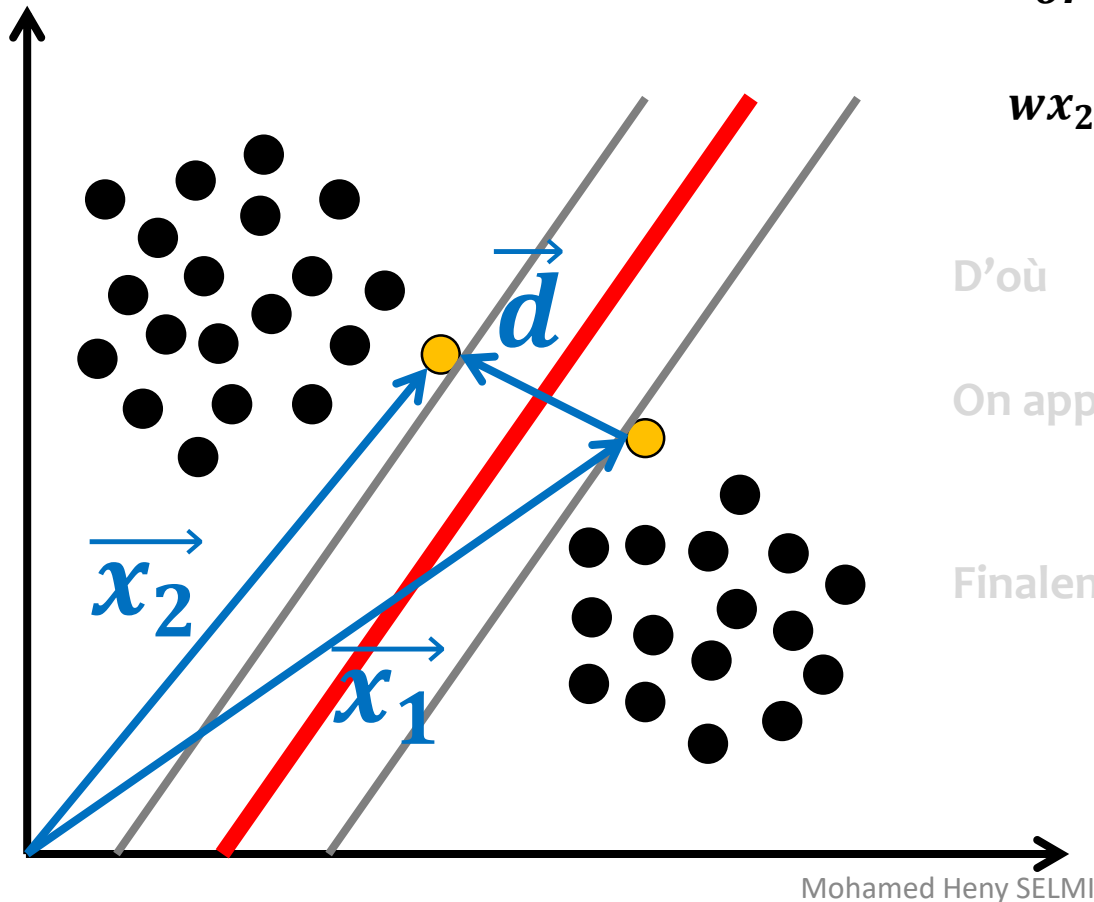
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

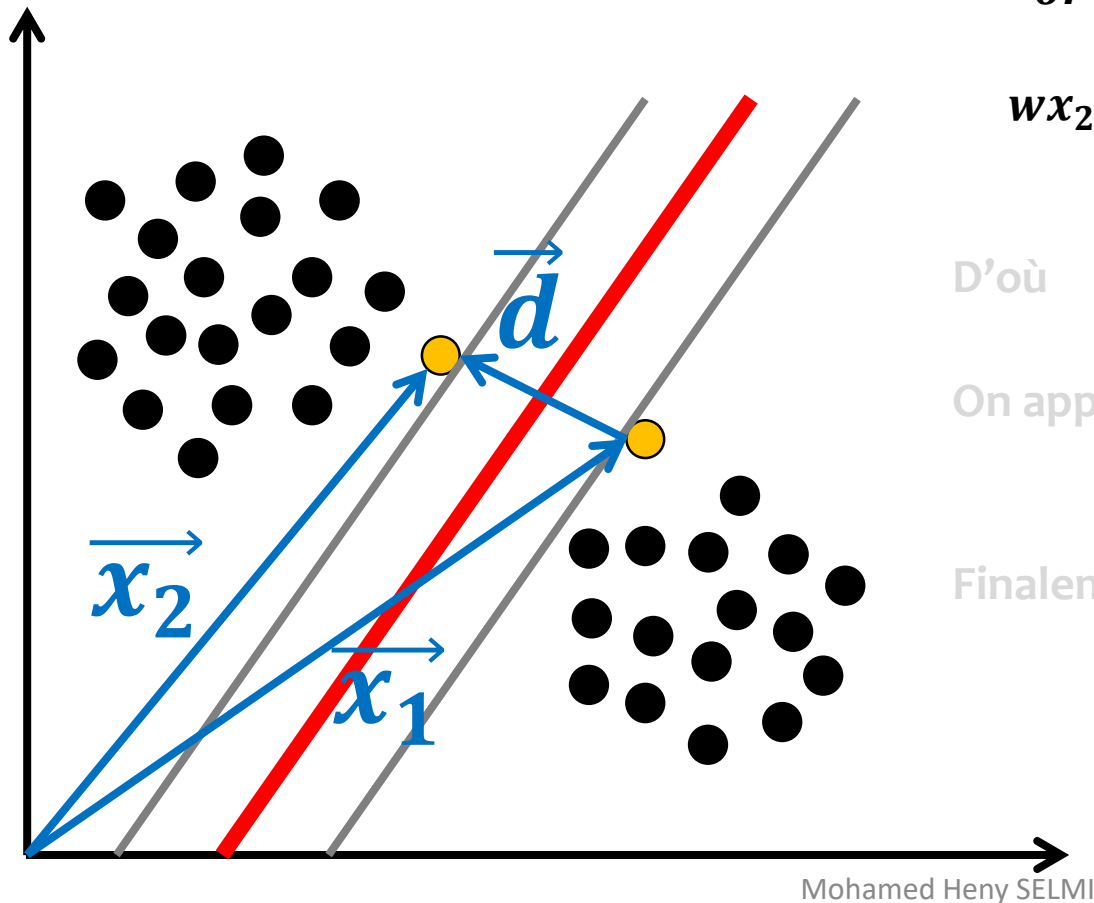
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

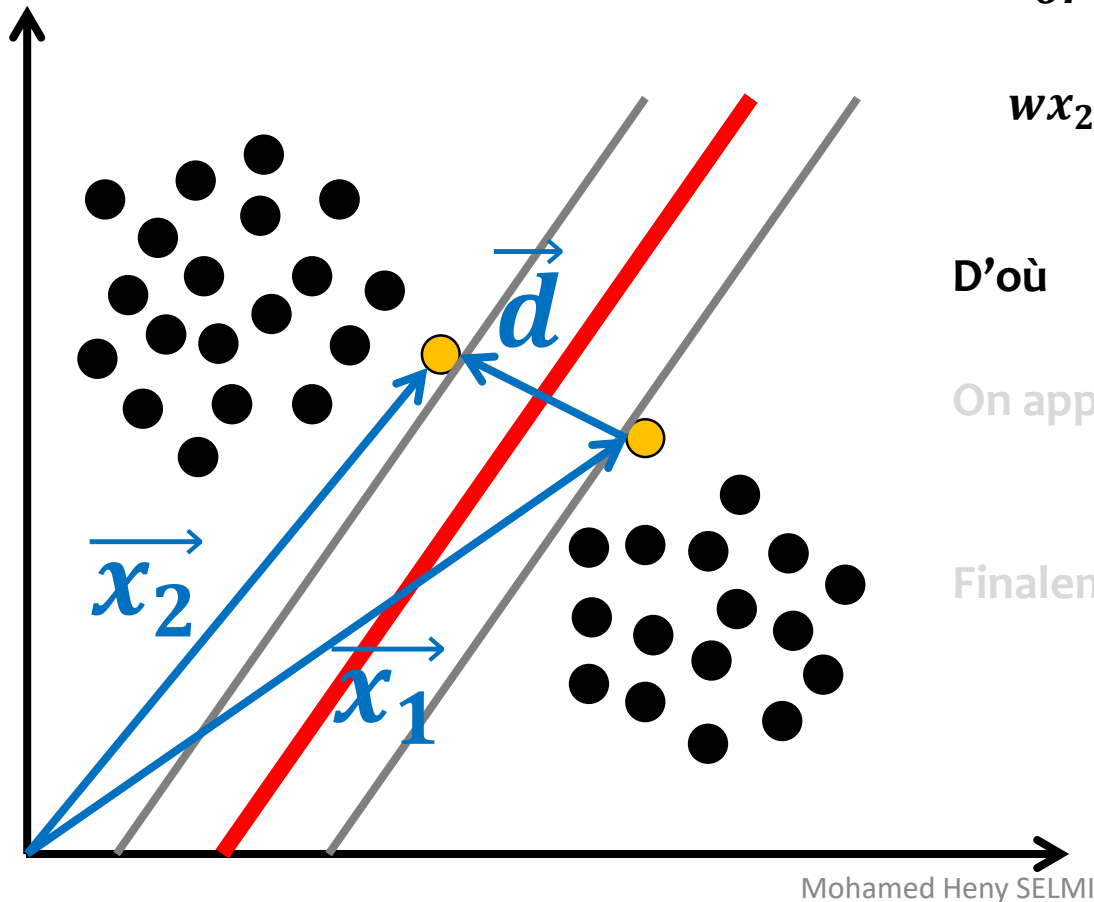
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

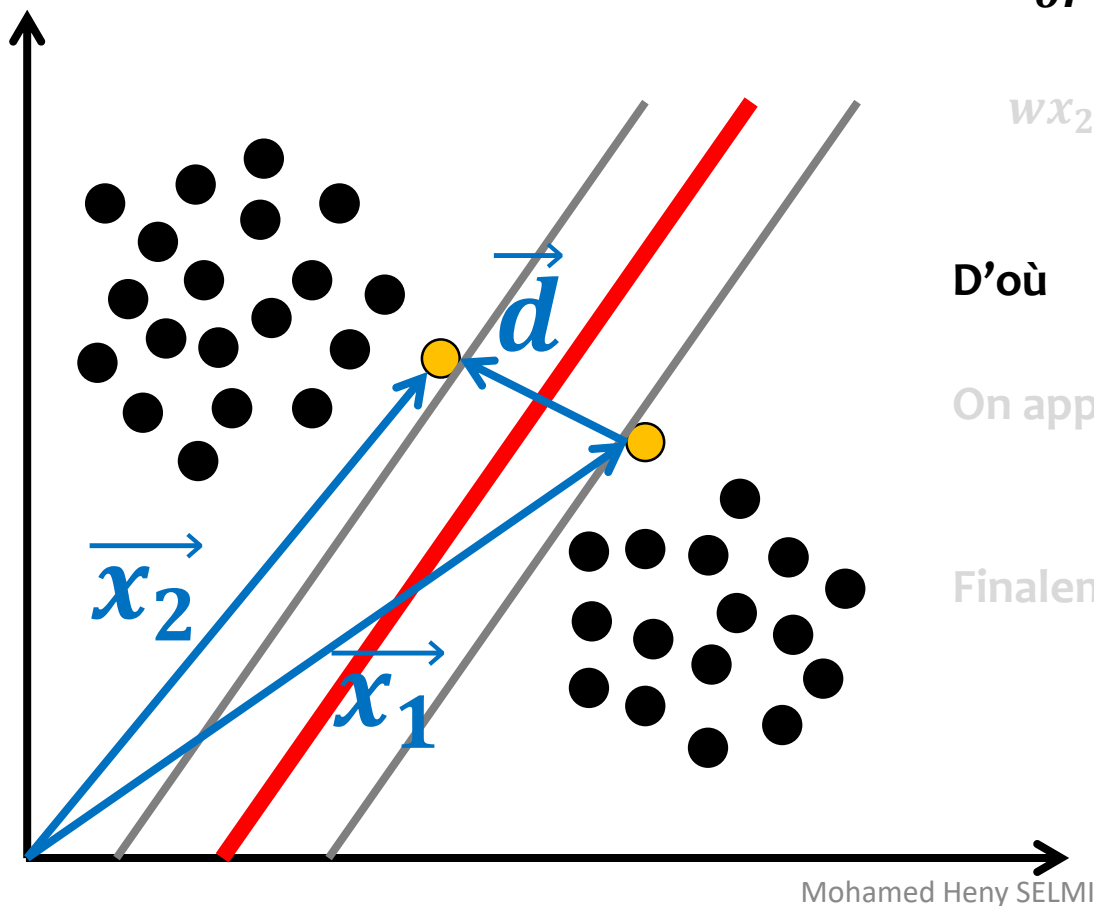
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(\vec{x}_2 - \vec{x}_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

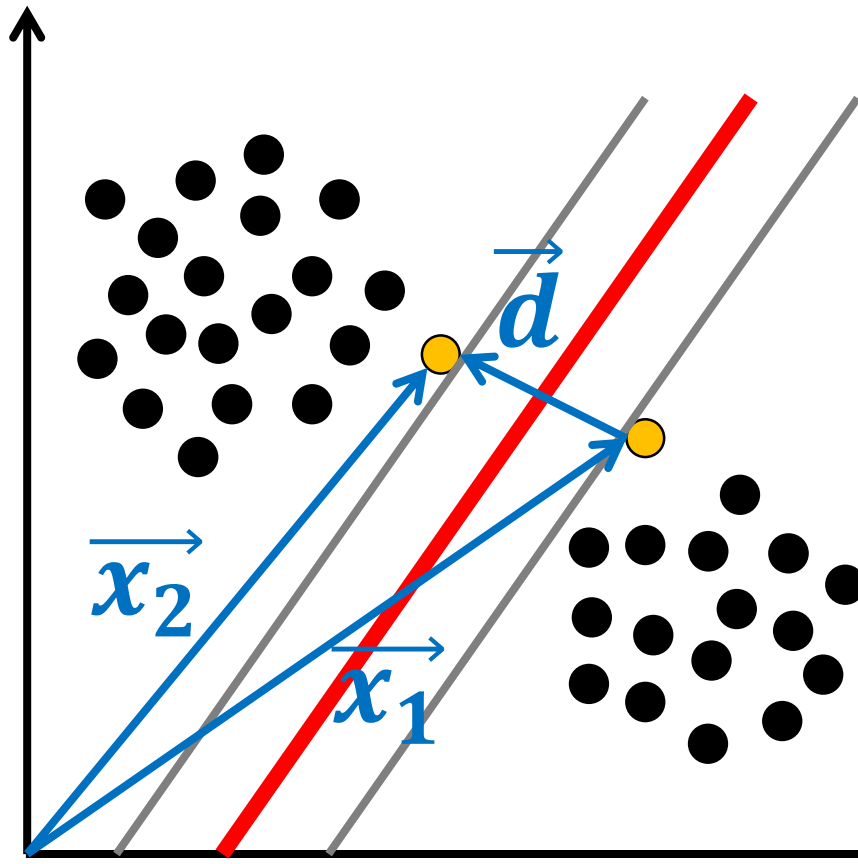
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

On applique la norme :

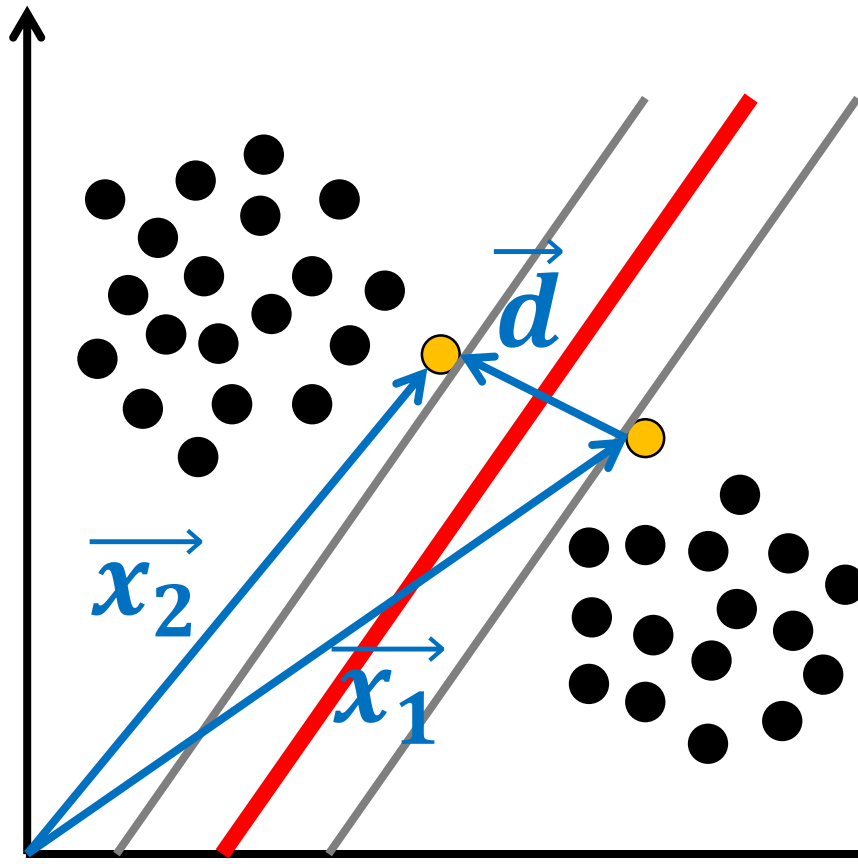
$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$



Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



Mohamed Heny SELMI

$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

On applique la norme :

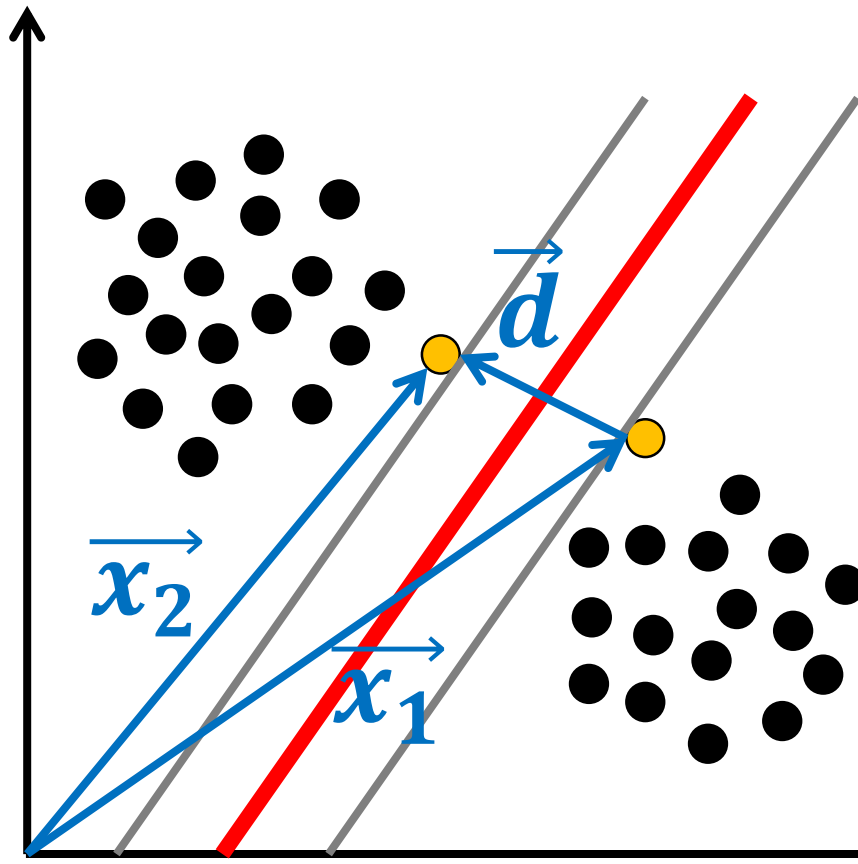
$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$



Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



Mohamed Heny SELMI

$$\vec{x_2} = \vec{x_1} + \vec{d}$$

$$\vec{x_2} - \vec{x_1} = \vec{d}$$

$$w \cdot (\vec{x_2} - \vec{x_1}) = w \cdot \vec{d}$$

$$\text{or } wx_2 + b = +1 \text{ et } wx_1 + b = -1$$

donc

$$wx_2 + b - wx_1 - b = +1 - (-1) = 2$$

$$wx_2 - wx_1 = 2$$

$$w(x_2 - x_1) = 2$$

D'où

$$2 = w \cdot \vec{d}$$

On applique la norme :

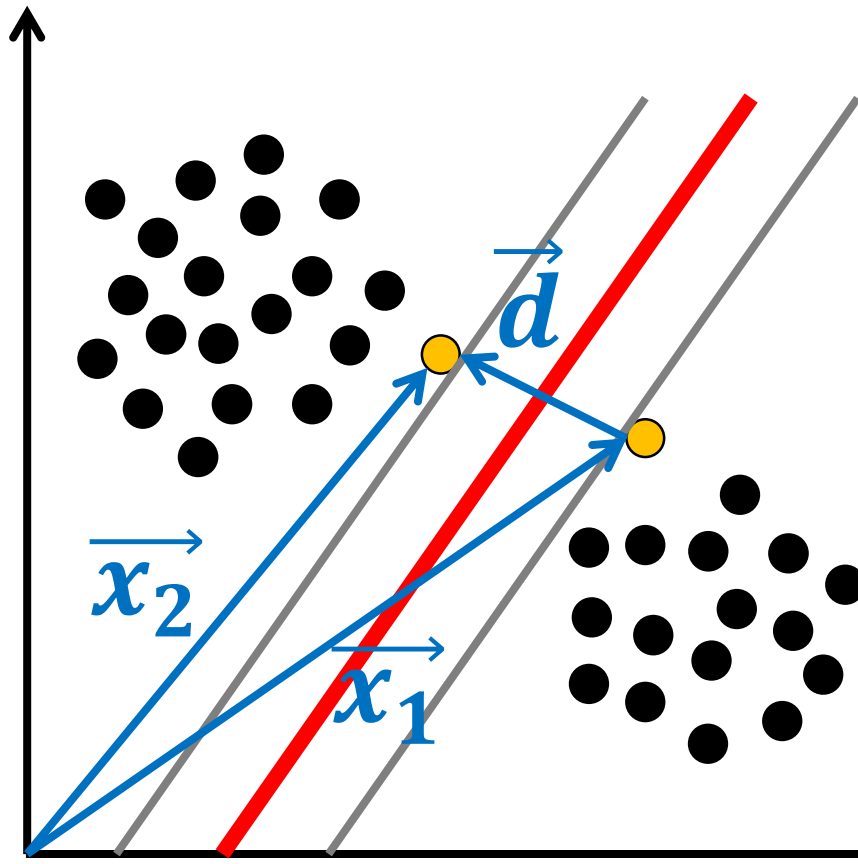
$$\|2\| = \|w\| \cdot \|\vec{d}\|$$

$$2 = \|w\| \cdot d$$

Finalement :

$$d = \frac{2}{\|w\|}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

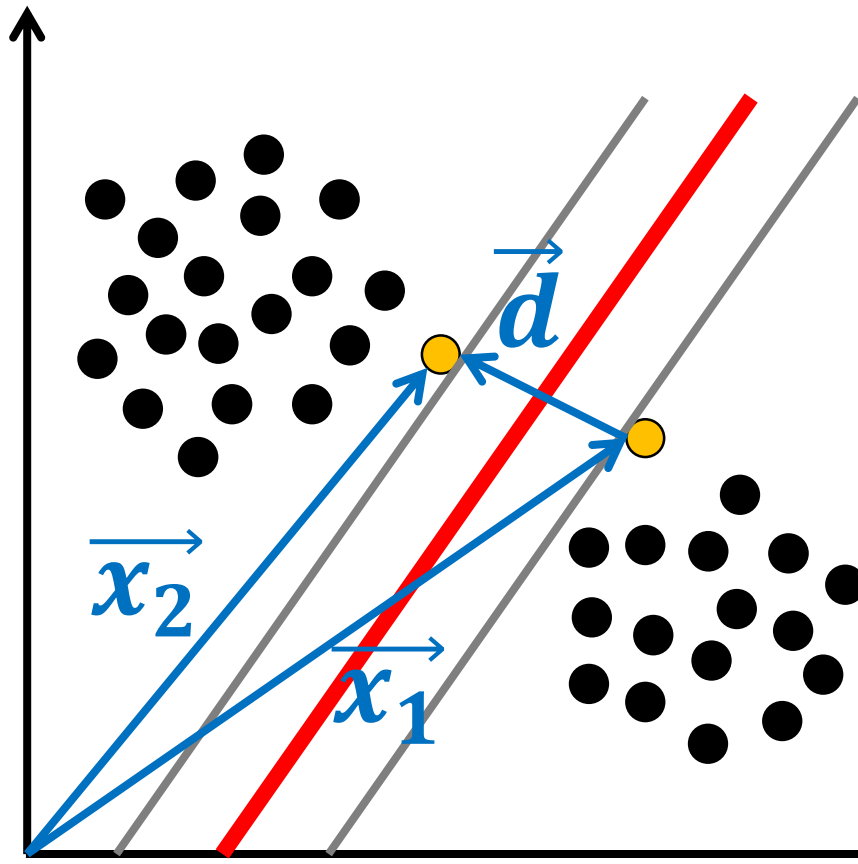
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$d = \frac{2}{\|w\|} \text{ à maximiser}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

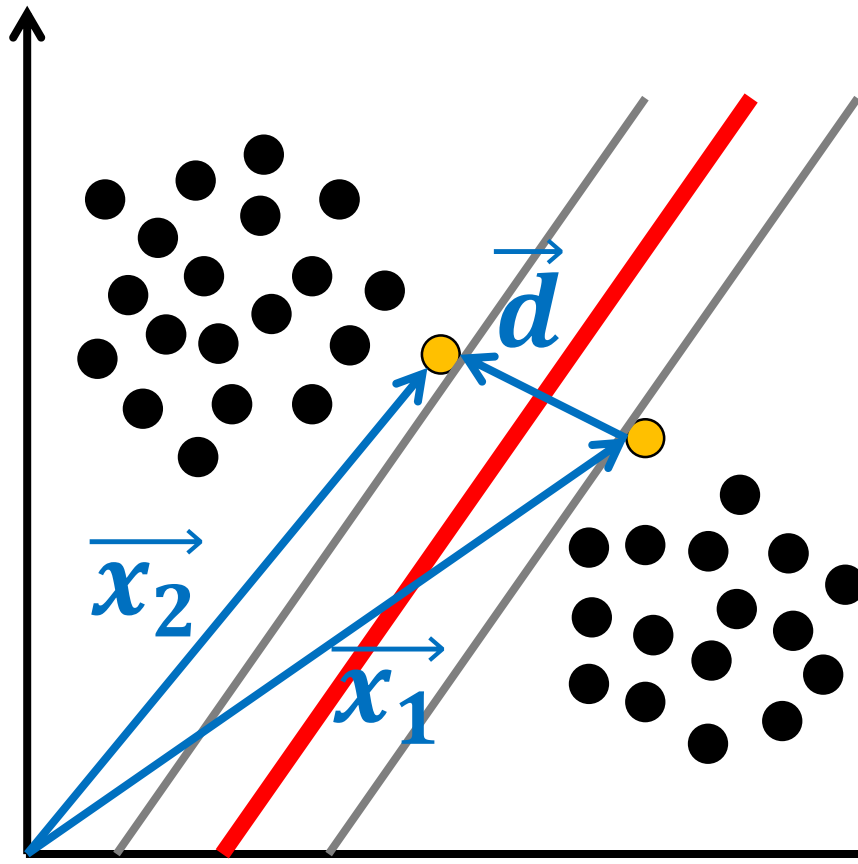
On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$\begin{aligned}d &= \frac{2}{\|w\|} \text{ à maximiser} \\ &\equiv \text{minimiser } \|w\|\end{aligned}$$

CALCUL DE LA MARGE ENTRE HYPERPLANS SÉPARATEURS



$$\begin{aligned}\vec{x}_2 &= \vec{x}_1 + \vec{d} \\ \vec{x}_2 - \vec{x}_1 &= \vec{d} \\ w \cdot (\vec{x}_2 - \vec{x}_1) &= w \cdot \vec{d} \\ \text{or } wx_2 + b &= +1 \text{ et } wx_1 + b = -1 \\ \text{donc} \\ wx_2 + b - wx_1 - b &= +1 - (-1) = 2 \\ wx_2 - wx_1 &= 2 \\ w(x_2 - x_1) &= 2\end{aligned}$$

D'où

$$2 = w \cdot \vec{d}$$

On applique la norme :

$$\begin{aligned}\|2\| &= \|w\| \cdot \|\vec{d}\| \\ 2 &= \|w\| \cdot d\end{aligned}$$

Finalement :

$$\begin{aligned}d &= \frac{2}{\|w\|} \text{ à maximiser} \\ &\equiv \text{minimiser } \|w\| \text{ tout en préservant le} \\ &\text{pouvoir de classification}\end{aligned}$$



Afin d'optimiser la marge, il faut réaliser les deux objectifs suivants :

1

Classer correctement les individus

$$\begin{aligned}\forall y_i = +1, wx_i + b &\geq +1 \\ \forall y_i = -1, wx_i + b &\leq -1\end{aligned}$$



$$y_i (wx_i + b) \geq +1 \quad \forall i$$

2

Maximiser la marge

$$d = \frac{2}{\|w\|} \text{ à maximiser}$$



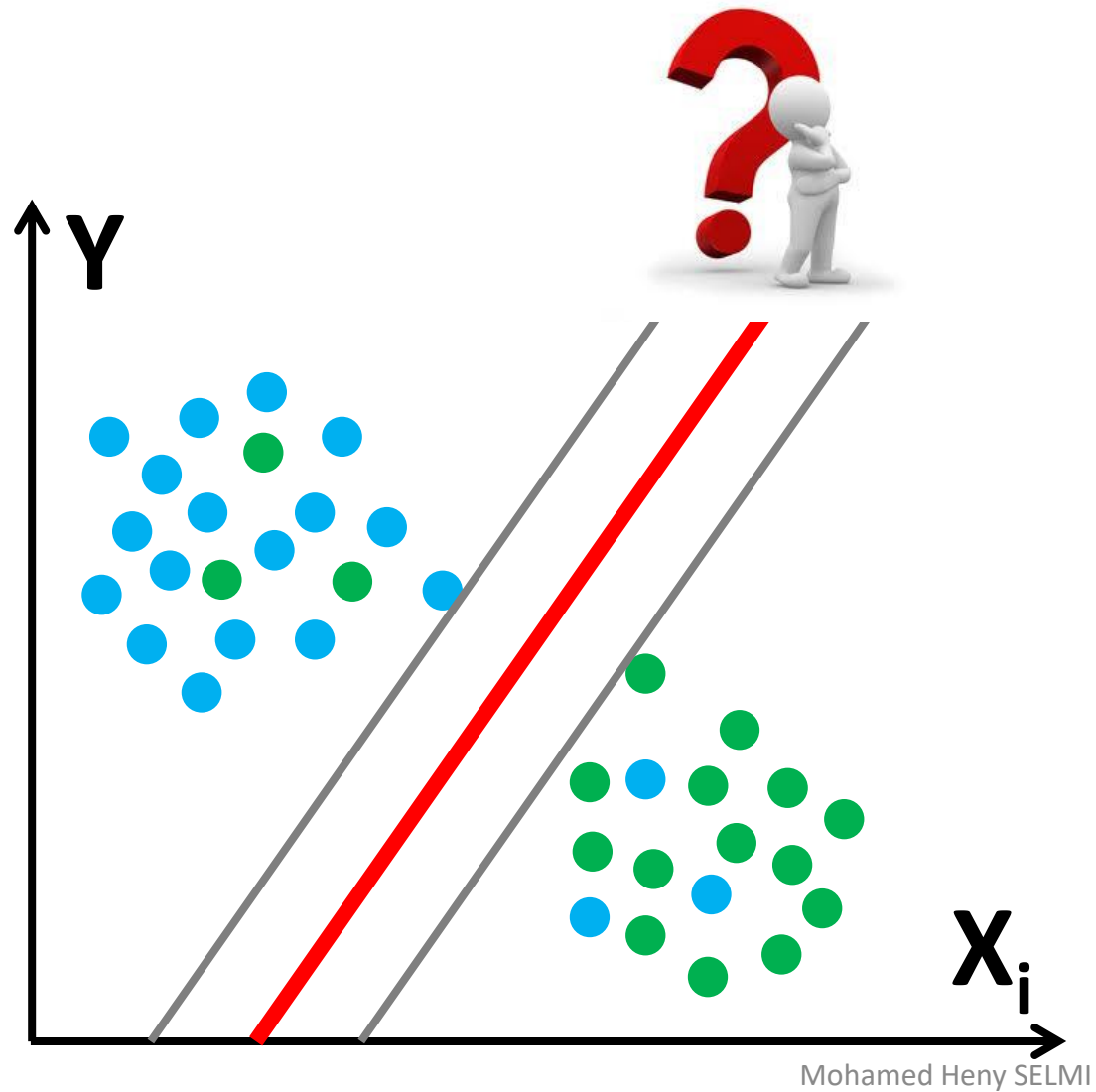
$$\text{minimiser } \|w\| \equiv \text{minimiser } \frac{\|w\|^2}{2}$$

$\|w\|^2$ pour éviter la racine

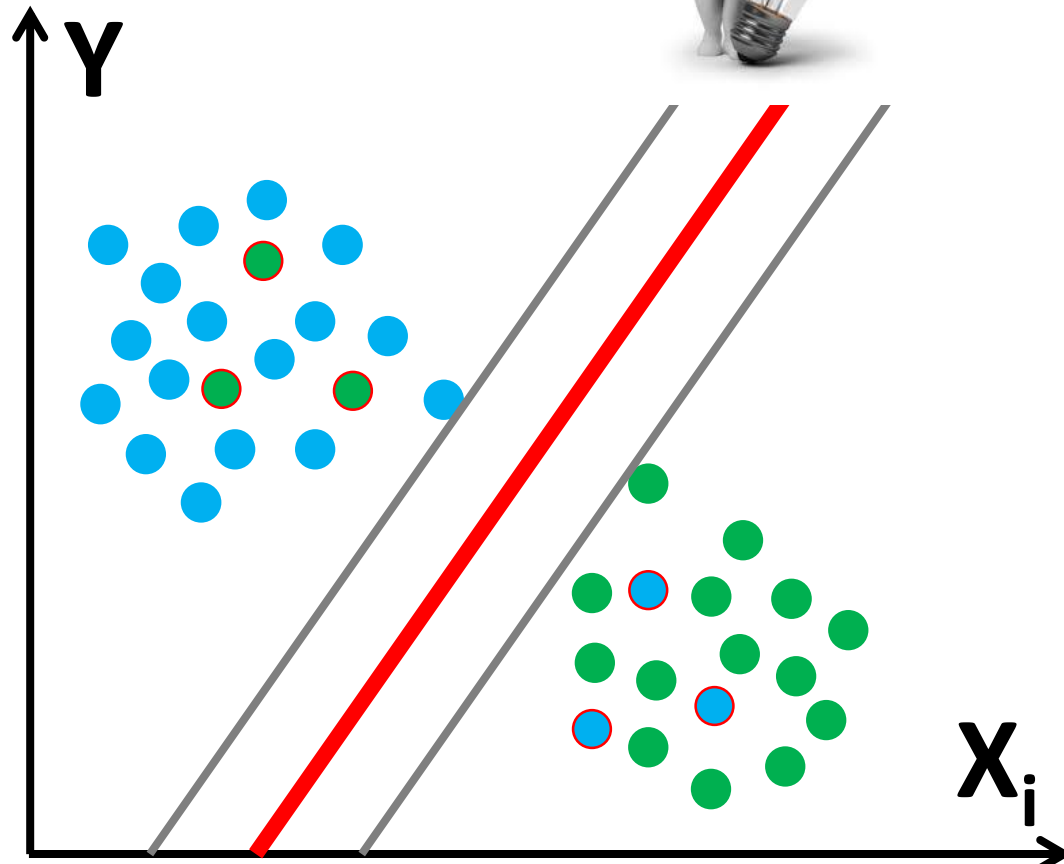
$\frac{1}{2}$ reste utile en cas de dérivation

$$\frac{1}{2} \|w\|^2 = \frac{1}{2} \left(\sqrt{\sum_i w_i^2} \right)^2 = \frac{1}{2} \sum_i w_i^2$$

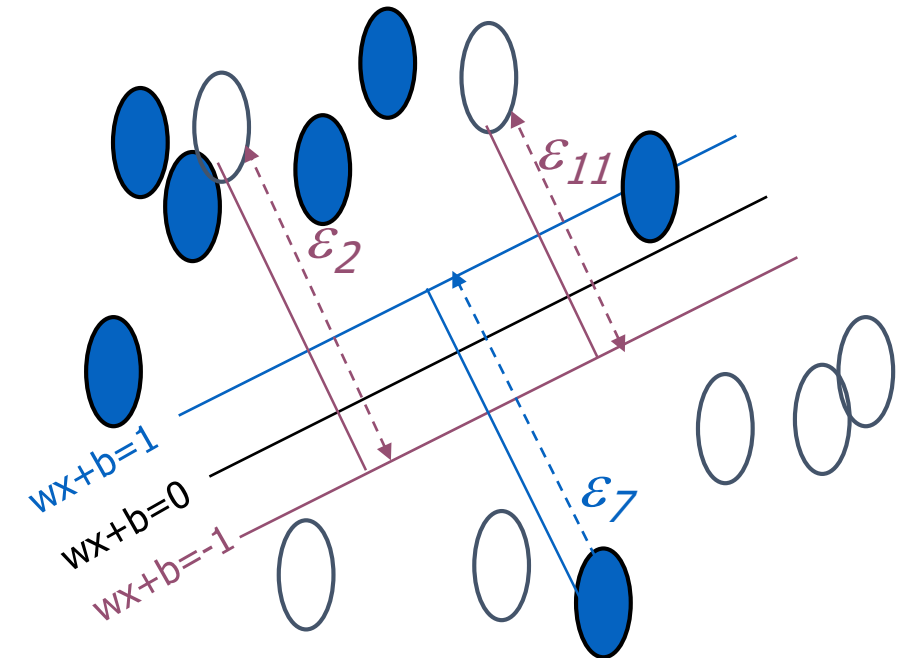
PROBLÉMATIQUE DES INDIVIDUS MAL PLACÉS



PROBLÉMATIQUE DES INDIVIDUS MAL PLACÉS



minimiser $\frac{1}{2} \|w\|^2 + C \cdot \sum_1^N \varepsilon_i$
(taux d'erreur de classification)
 N : nombre d'individus mal classés



Afin d'optimiser la marge, il faut réaliser les deux objectifs suivants :

1

Classer correctement les individus

$$\begin{aligned}\forall y_i = +1, wx_i + b &\geq +1 \\ \forall y_i = -1, wx_i + b &\leq -1\end{aligned}$$



$$y_i (wx_i + b) \geq +1 - \epsilon_i \quad \forall i, \epsilon_i$$



2

Maximiser la marge

$$d = \frac{2}{\|w\|} \text{ à maximiser}$$



$$\text{minimiser } \|w\| \equiv \text{minimiser } \frac{\|w\|^2}{2}$$

$\|w\|^2$ pour éviter la racine

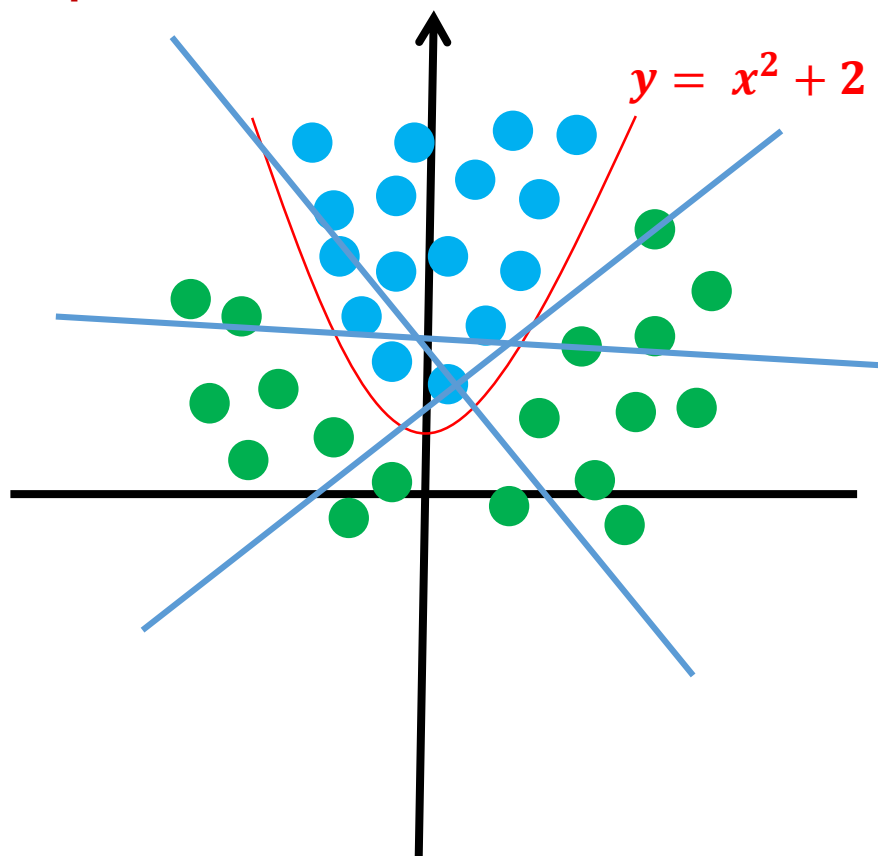
$\frac{1}{2}$ reste utile en cas de dérivation

$$\frac{1}{2} \|w\|^2 + C \cdot \sum_1^N \epsilon_i$$

C paramètre de control du sur apprentissage

PROBLÈME LINÉAIREMENT NON SÉPARABLE

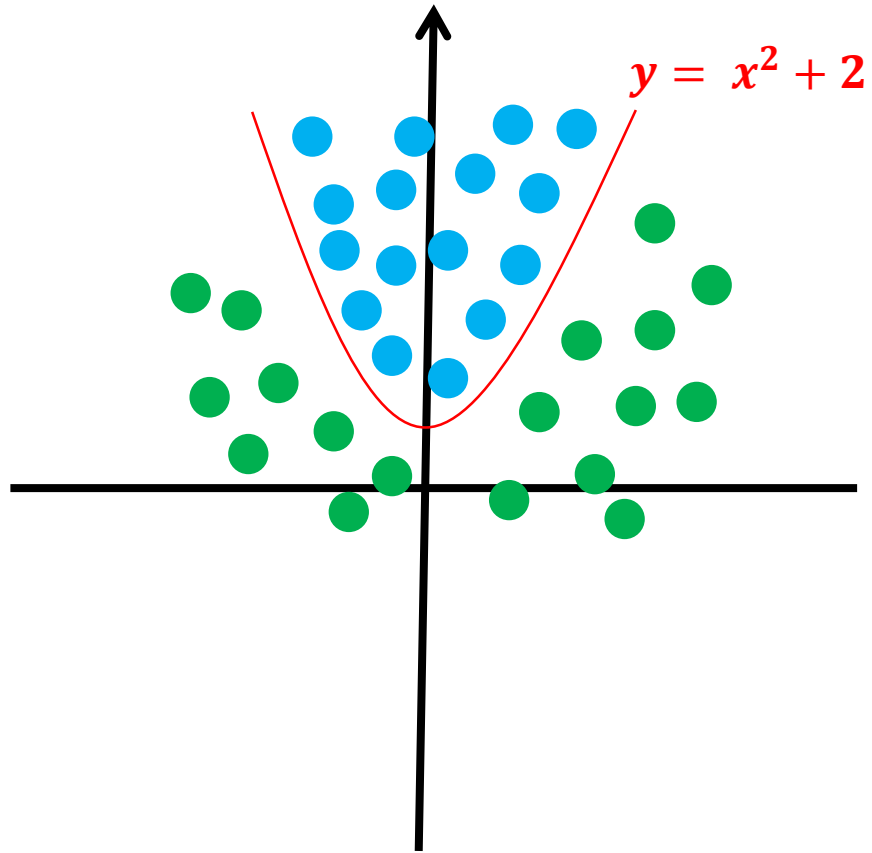
Exemple



la séparation linéaire est impossible :
Recours à d'autres type de classifieurs

PROBLÈME LINÉAIREMENT NON SÉPARABLE

Exemple

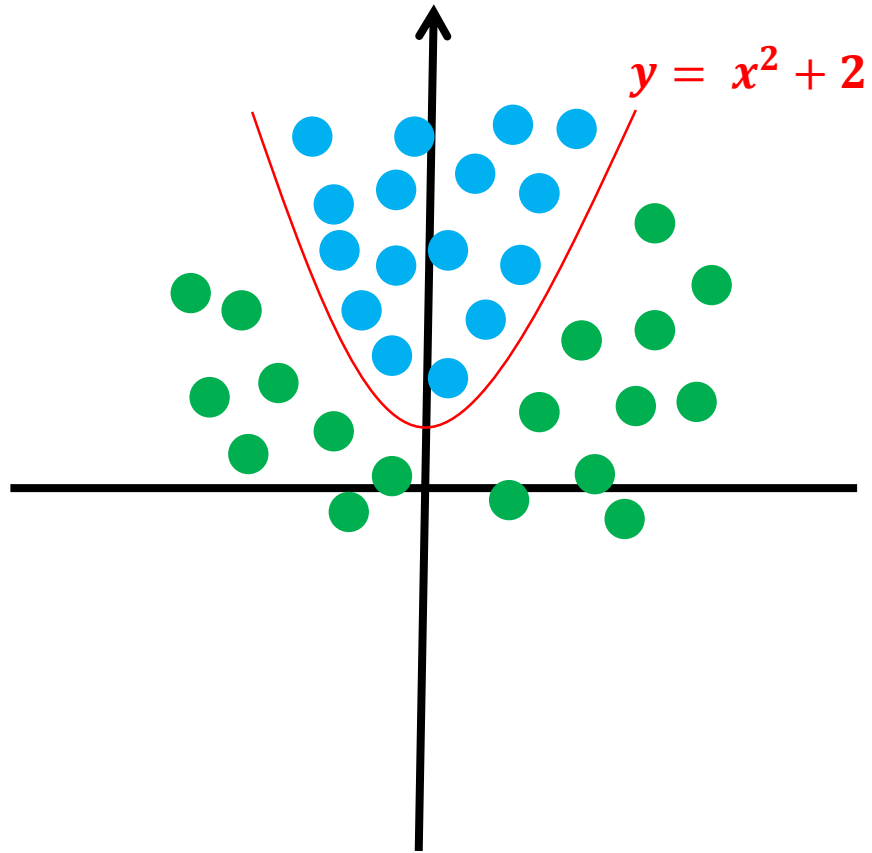


la séparation linéaire est impossible :
Recours à d'autres type de classifieurs

Classifieur non linéaire

PROBLÈME LINÉAIREMENT NON SÉPARABLE

Exemple



la séparation linéaire est impossible :
Recours à d'autres type de classifieurs

Classifieur non linéaire

Fonction Noyau

- Linéaire : $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- Polynomiale : $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^p$
- Gaussienne :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

k Plus Proches voisins

k NN

- Apprendre par analogie

Recherchant d'un ou des cas similaires déjà résolus

- Classifier ou estimer

“Dis moi qui sont tes amis, et je te dirais qui tu es”

- Pas de construction de modèle

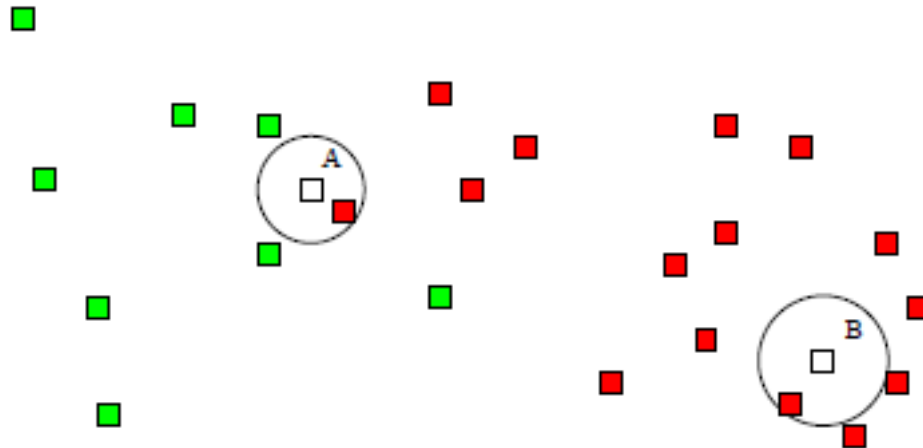
C'est l'échantillon d'apprentissage, associé à une fonction de distance et d'une fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constitue le modèle

Principe :

- ✓ Regarder la classe des k exemples les plus proches ($k = 1, 3, \dots$)
- ✓ Affecter la classe majoritaire au nouvel exemple

Exemple :

- ✓ Deux classes : verte et rouge
- ✓ $k = 1$

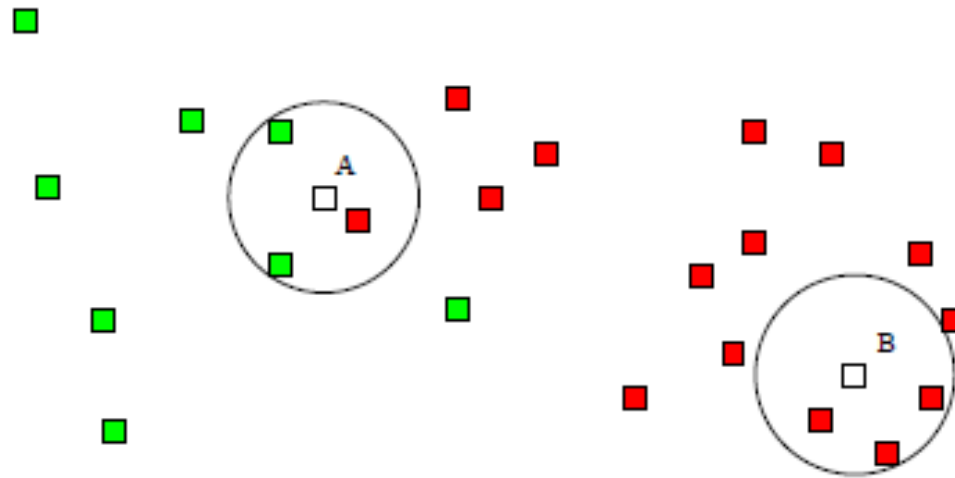


Principe :

- ✓ Regarder la classe des k exemples les plus proches ($k = 1, 3, \dots$)
- ✓ Affecter la classe majoritaire au nouvel exemple

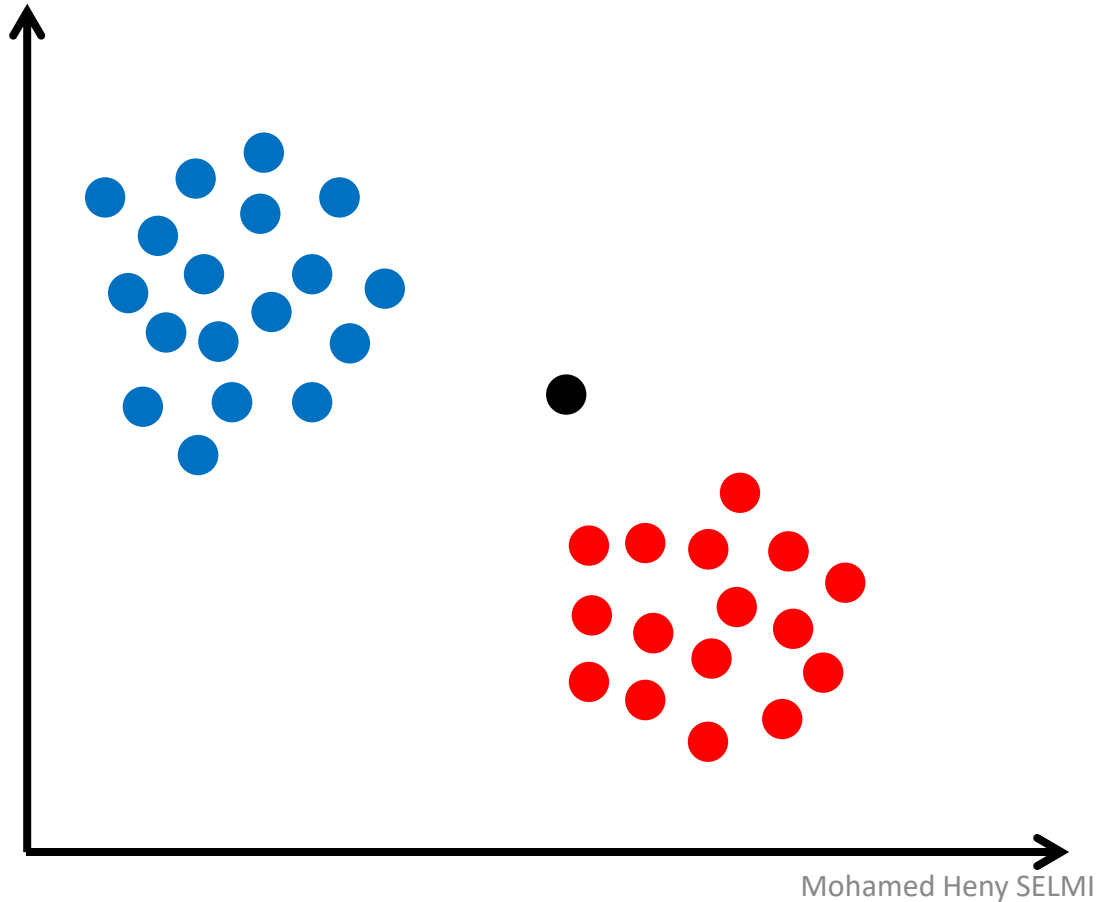
Exemple :

- ✓ Deux classes : verte et rouge
- ✓ $k = 3$



objectifs :

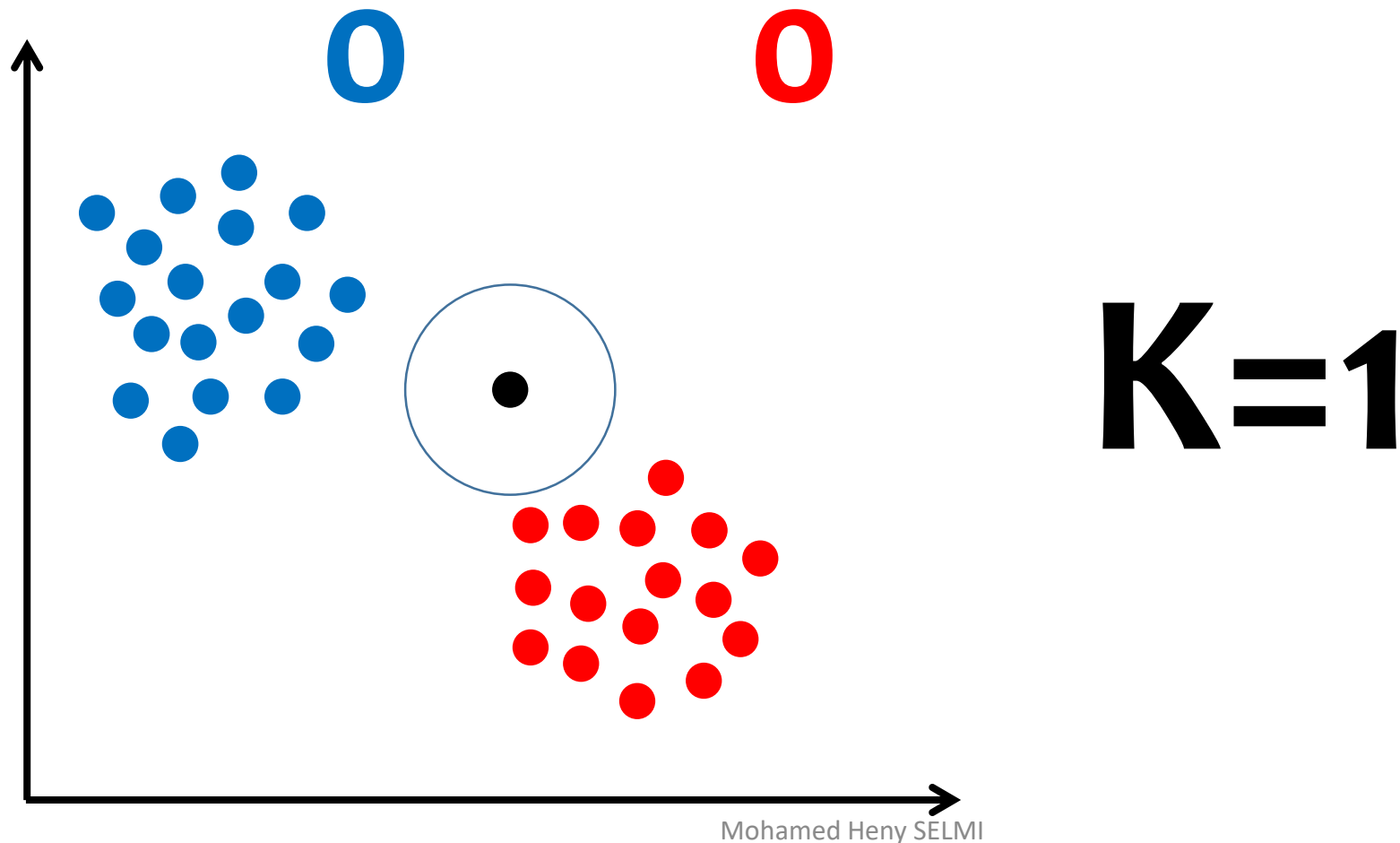
✓ Prédire la classe d'appartenance du point noir ?



PRINCIPE DES K PLUS PROCHES VOISINS

objectifs :

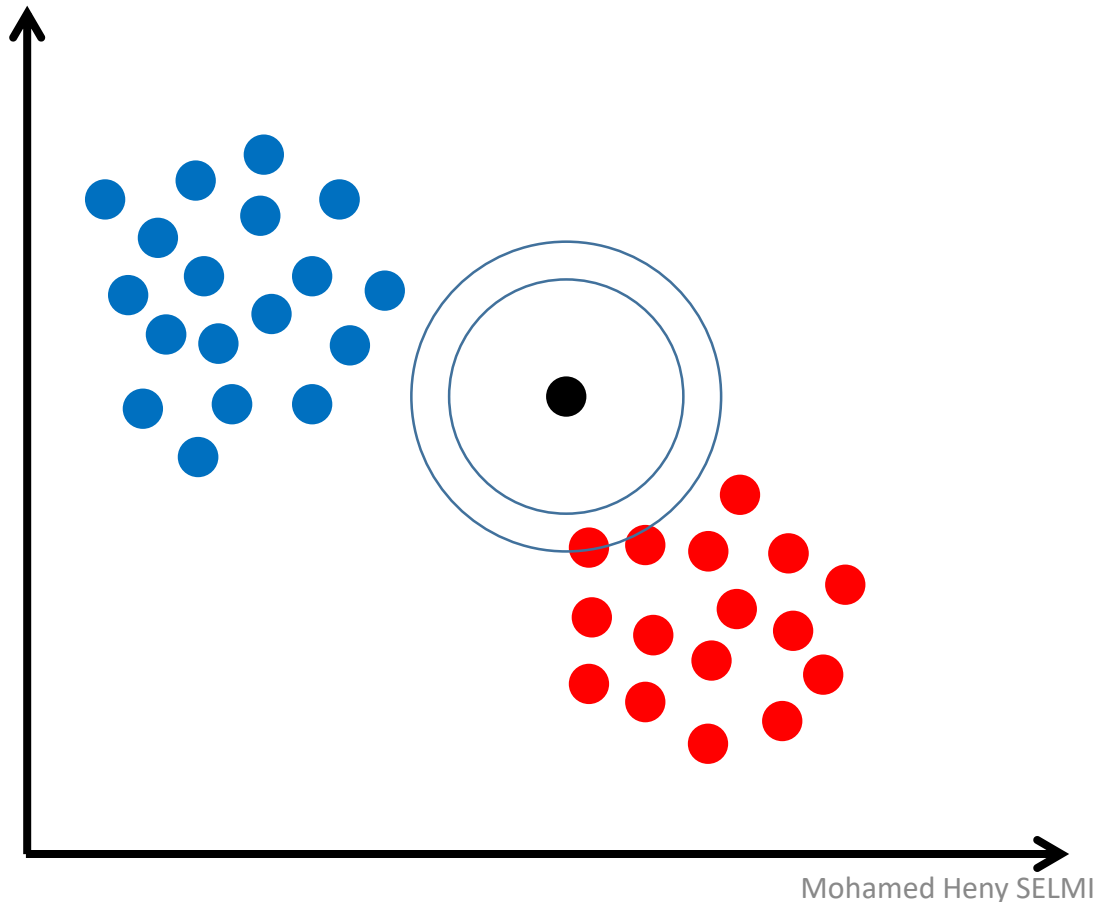
✓ Prédire la classe d'appartenance du point noir



PRINCIPE DES K PLUS PROCHES VOISINS

objectifs :

✓ Prédire la classe d'appartenance du point noir

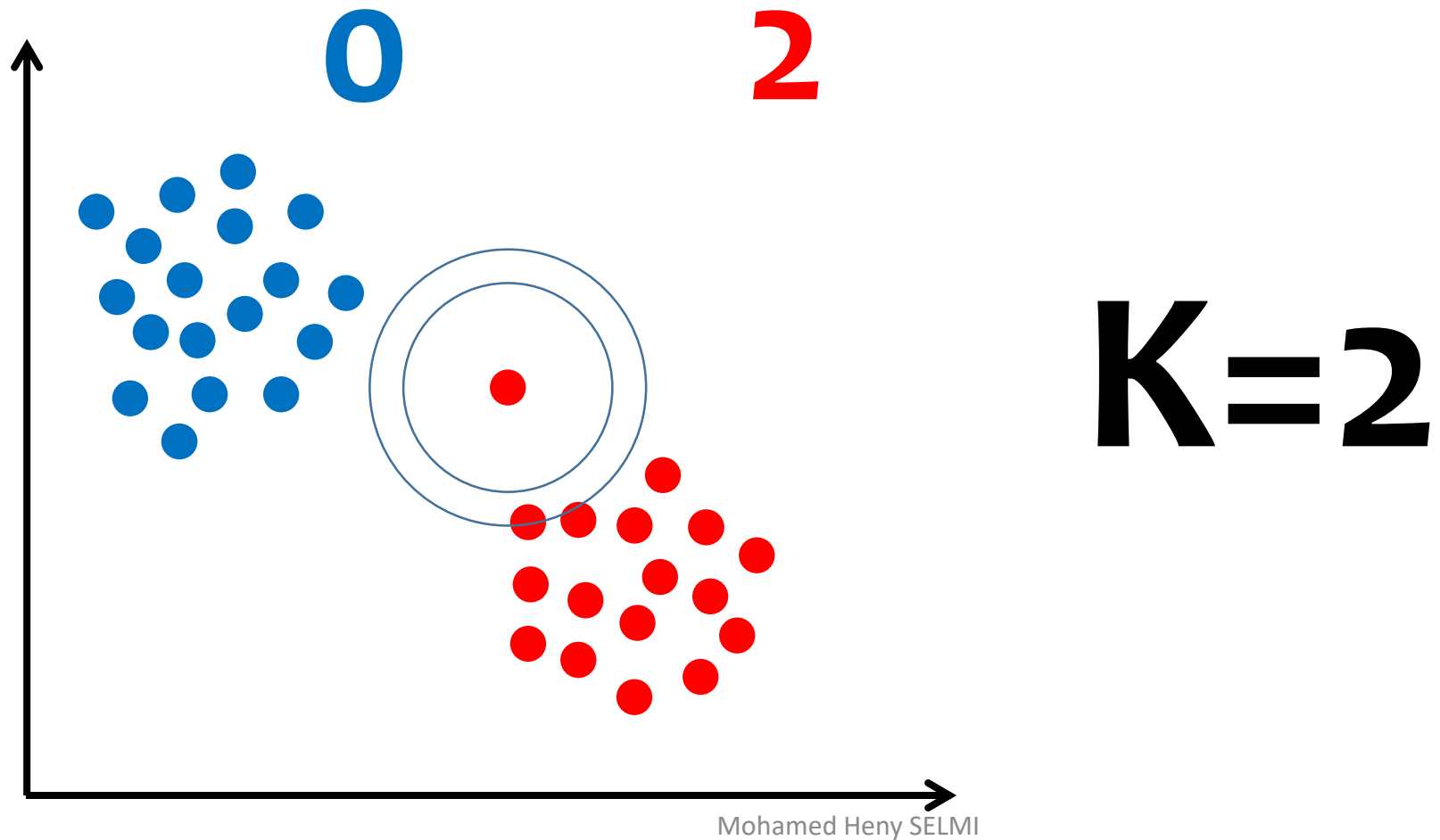


K=2

PRINCIPE DES K PLUS PROCHES VOISINS

objectifs :

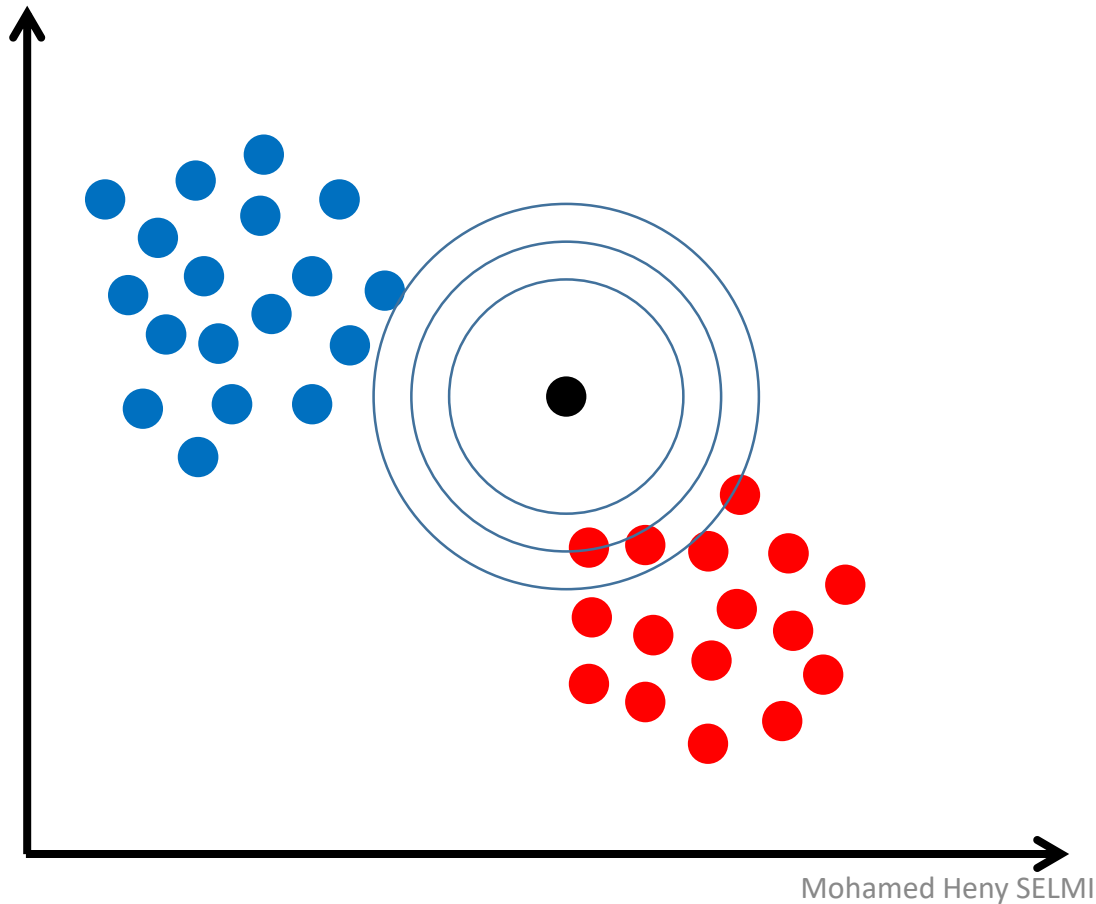
✓ Prédire la classe d'appartenance du point noir



PRINCIPE DES K PLUS PROCHES VOISINS

objectifs :

✓ Prédire la classe d'appartenance du point noir

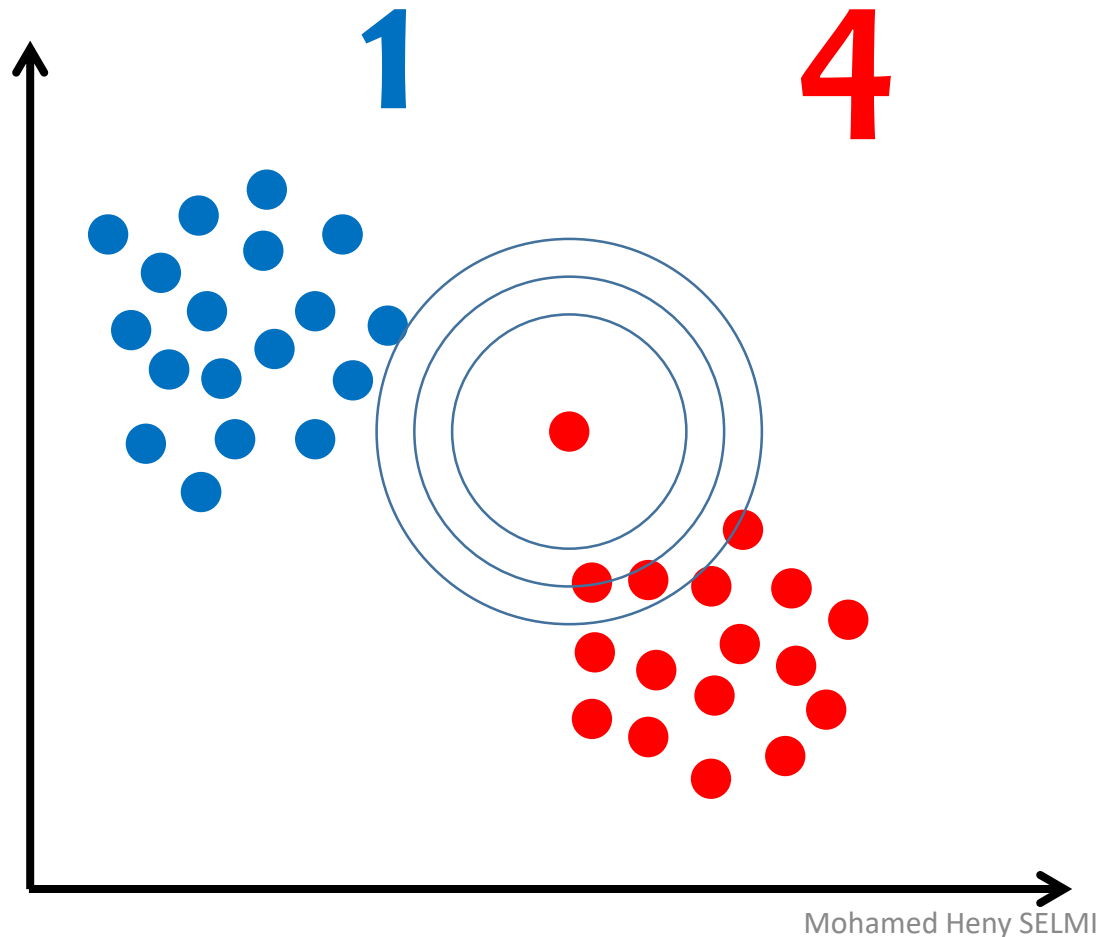


K=3

PRINCIPE DES K PLUS PROCHES VOISINS

objectifs :

✓ Prédire la classe d'appartenance du point noir

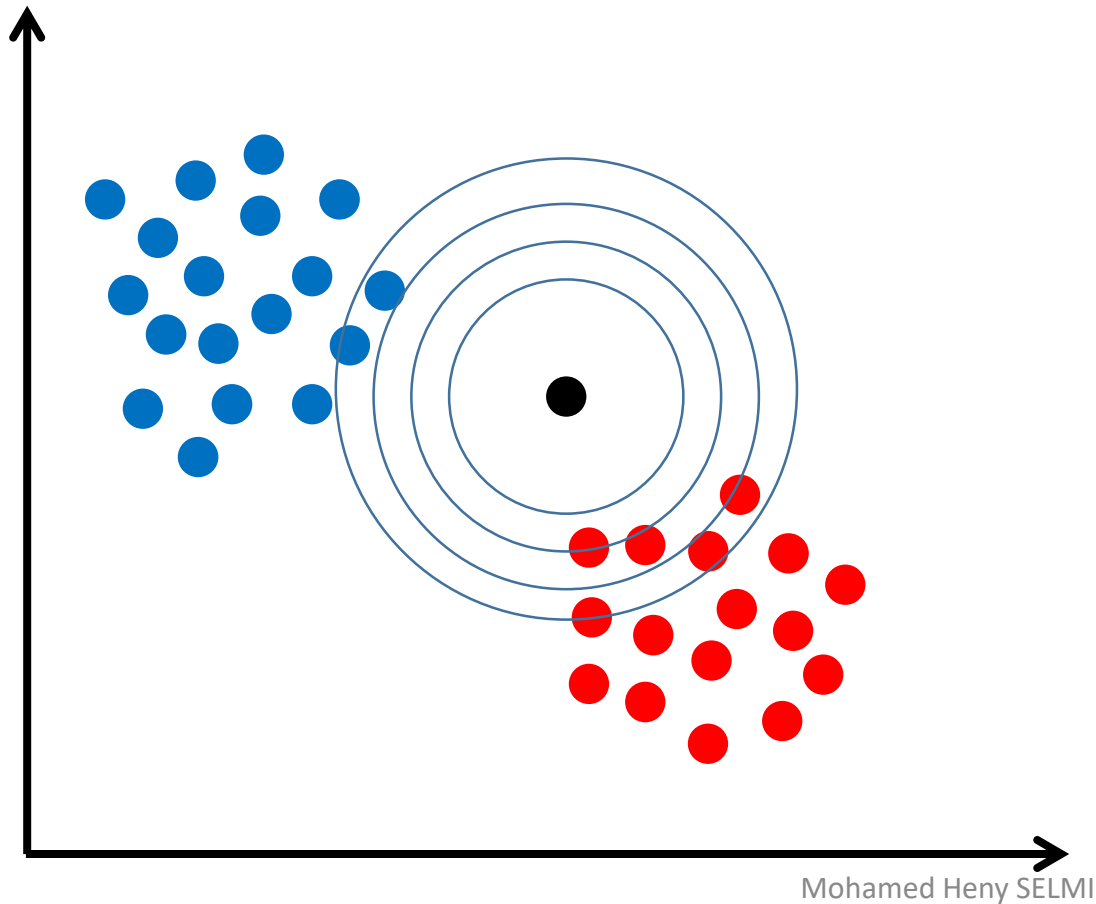


K=3

PRINCIPE DES K PLUS PROCHES VOISINS

objectifs :

✓ Prédire la classe d'appartenance du point noir

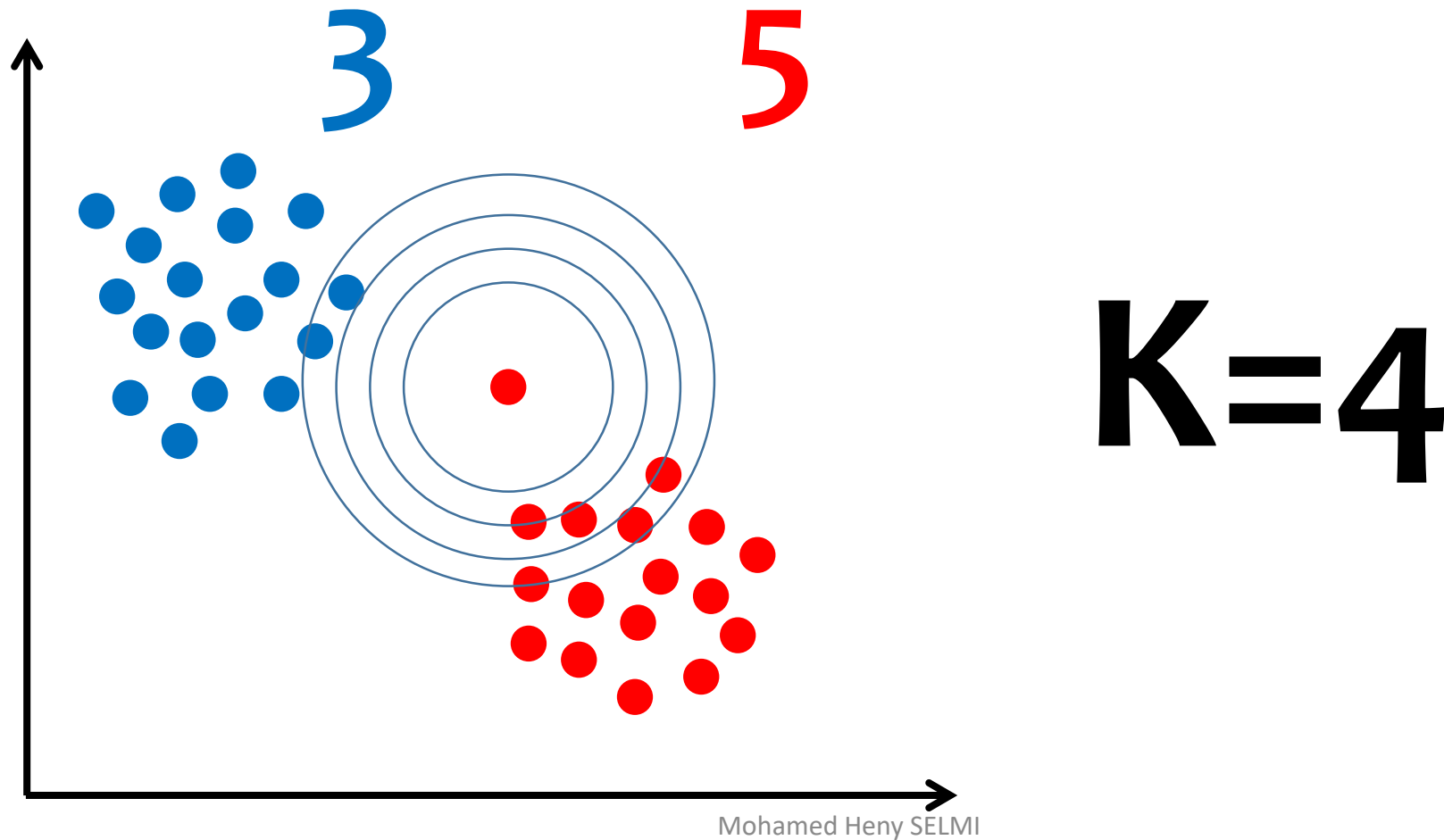


K=4

PRINCIPE DES K PLUS PROCHES VOISINS

objectifs :

✓ Prédire la classe d'appartenance du point noir



ALGORITHME

notations :

- Soit $L = \{(x', c) \mid x' \in R^d, c \in C\}$ l'ensemble d'apprentissage
- Soit x l'exemple qu'on désire déterminer la classe d'appartenance

algorithme kNN

DEBUT kNN

POUR chaque exemple de $(x', c) \in L$ **FAIRE**
calculer la distance $D(x, x')$
FIN POUR

POUR chaque $\{x' \in kppv(x)\}$ **FAIRE**
compter le nombre d'occurrences de chaque classe
FIN POUR

attribuer à x la classe la plus fréquente

FIN kNN

Les attributs ont le même poids

centrer et réduire pour éviter les biais

certain peuvent être moins classant que d'autres

Apprentissage paresseux

rien n'est préparé avant le classement

tous les calculs sont fait lors du classement

nécessité de technique d'indexation pour large BD

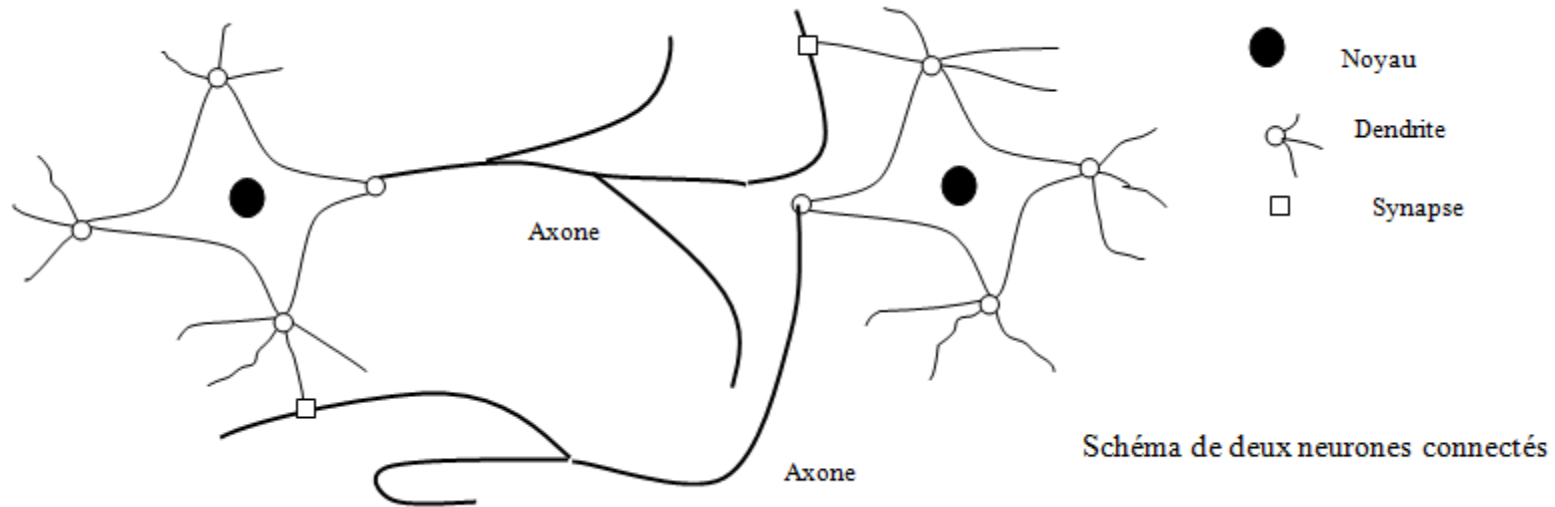
Calcul du score d'une classe

peut changer les résultats; variantes possibles

Réseaux de Neurones

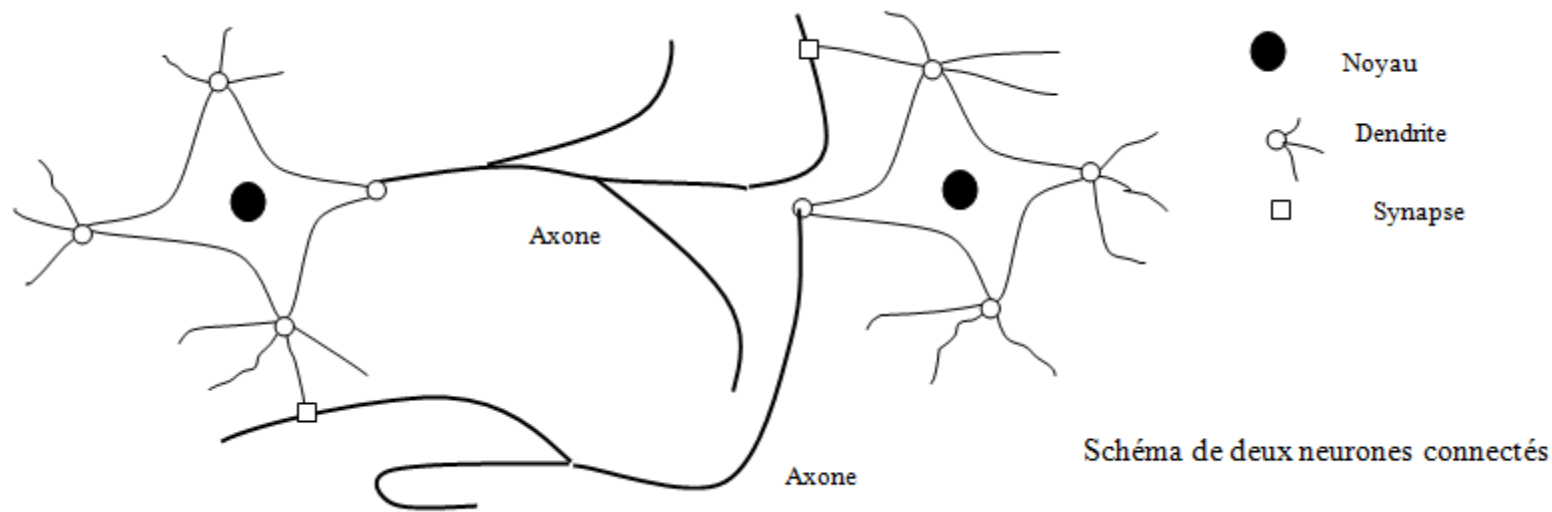
Fondement biologique

- ✓ L'élément fondamental du système nerveux est le neurone ou cellule nerveuse.
- ✓ Le neurone comprend une masse protoplasmique qui entoure le noyau, de nombreuses arborisations protoplasmiques ou dendrites et un long prolongement cylindrique ou axone.
- ✓ Le système nerveux peut être vu comme un ensemble de neurones interconnectés.



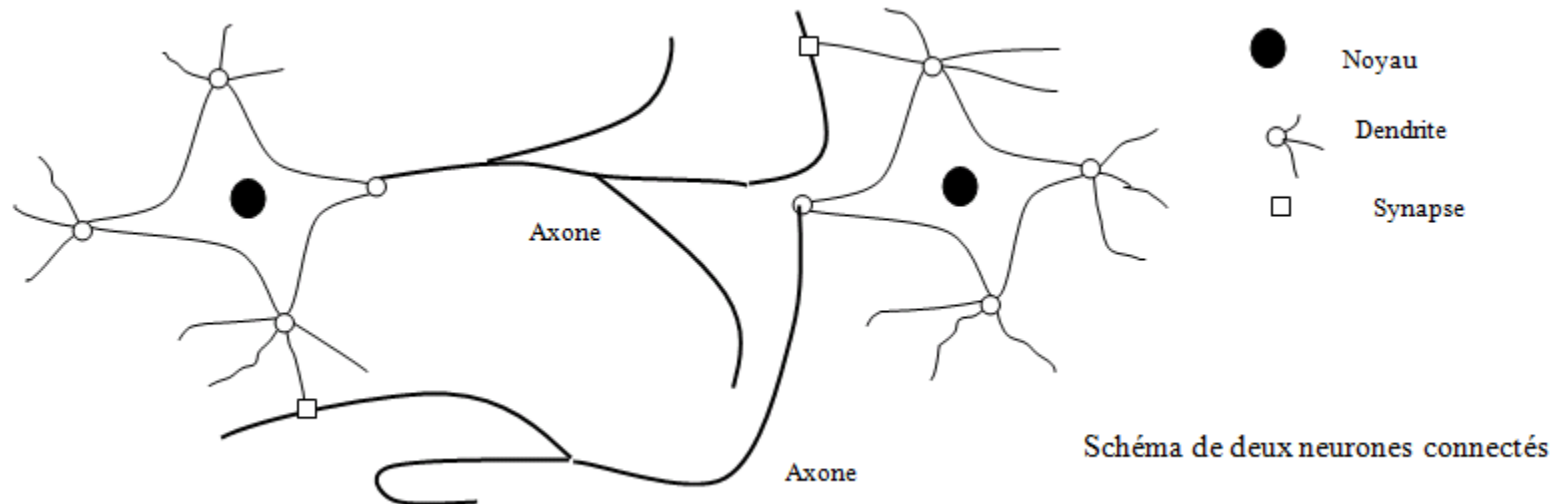
Fondement biologique

- L'axone se divise et ses ramifications sont reliées aux dendrites d'autres neurones ainsi qu'éventuellement aux siennes : les points de jonction sont les **synapses**; elles sont caractérisées par une **efficacité synaptique** qui peut être vue comme un amplificateur de l'impulsion qui traverse la synapse.



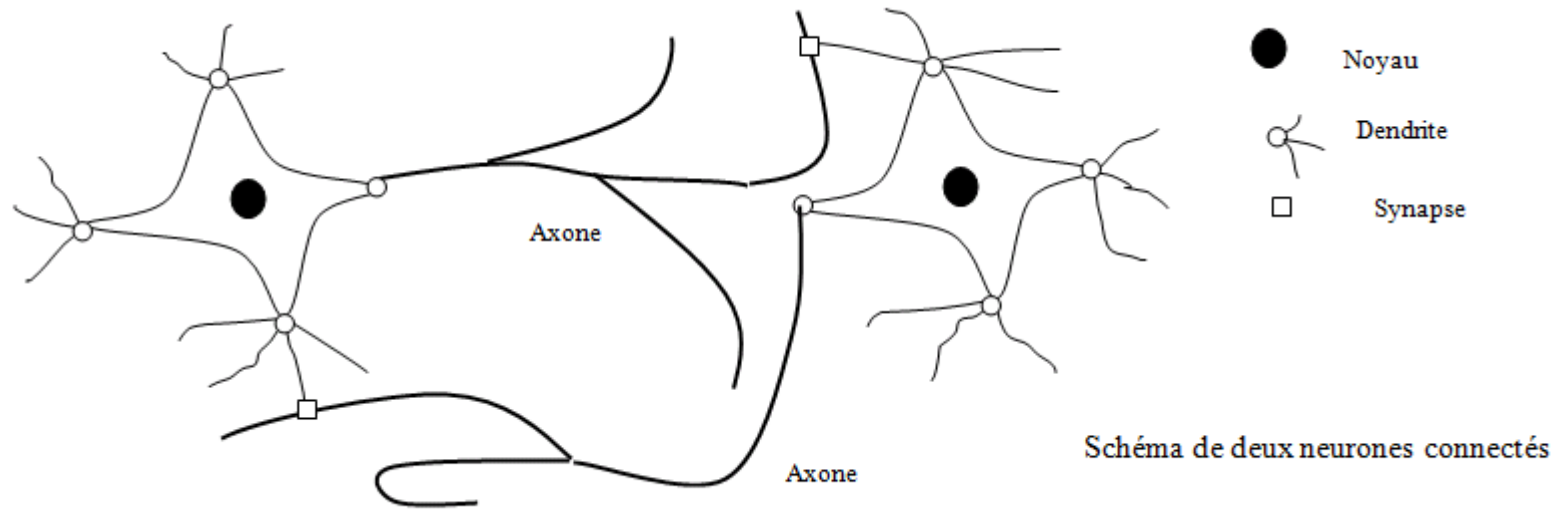
FONDEMENT BIOLOGIQUE

- Chaque neurone possède en son noyau un additionneur qui somme les impulsions électriques amplifiées par les synapses à l'entrée dans le neurone et un seuil de stimulation.
- Si l'excitation totale du neurone excède le seuil de stimulation, le noyau initie une impulsion.



Fondement biologique

- Les dendrites sont donc les organes d'entrées du neurone et l'axone son unité de sortie.
- L'impulsion peut prendre la forme d'une excitation ou d'une inhibition : l'activité d'un groupe de neurones peut renforcer ou prévenir l'activité concurrente d'un autre groupe.

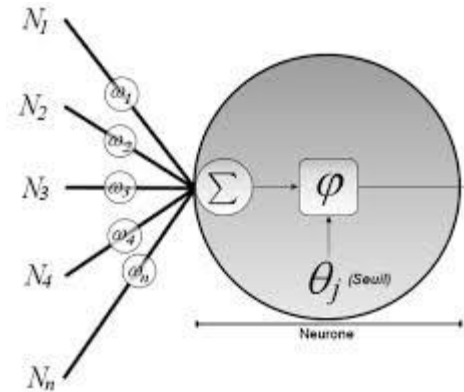
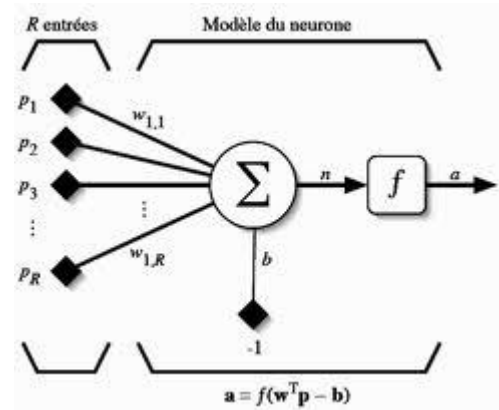


Quelques chiffres

- i. La durée d'une impulsion est de l'ordre de la milliseconde et l'amplitude d'environ 0,1 Volt.
- ii. La vitesse de propagation de l'influx nerveux est de 100 m/s environ donc bien inférieure à la vitesse de transmission de l'information dans un réseau électronique.
- iii. Chaque neurone intègre en permanence jusqu'à 1000 signaux synaptiques mais le nombre de contacts synaptiques par neurones peut atteindre plusieurs dizaine de milliers.
- iv. Le cerveau contient environ 100 milliards de neurones donc, par analogie avec la théorie des graphes, le nombre de connexions (arcs) est de l'ordre de 10^4 (degré) * 10^{11} (nombre de sommets) soit 10^{15} environ.

Métaphore biologique

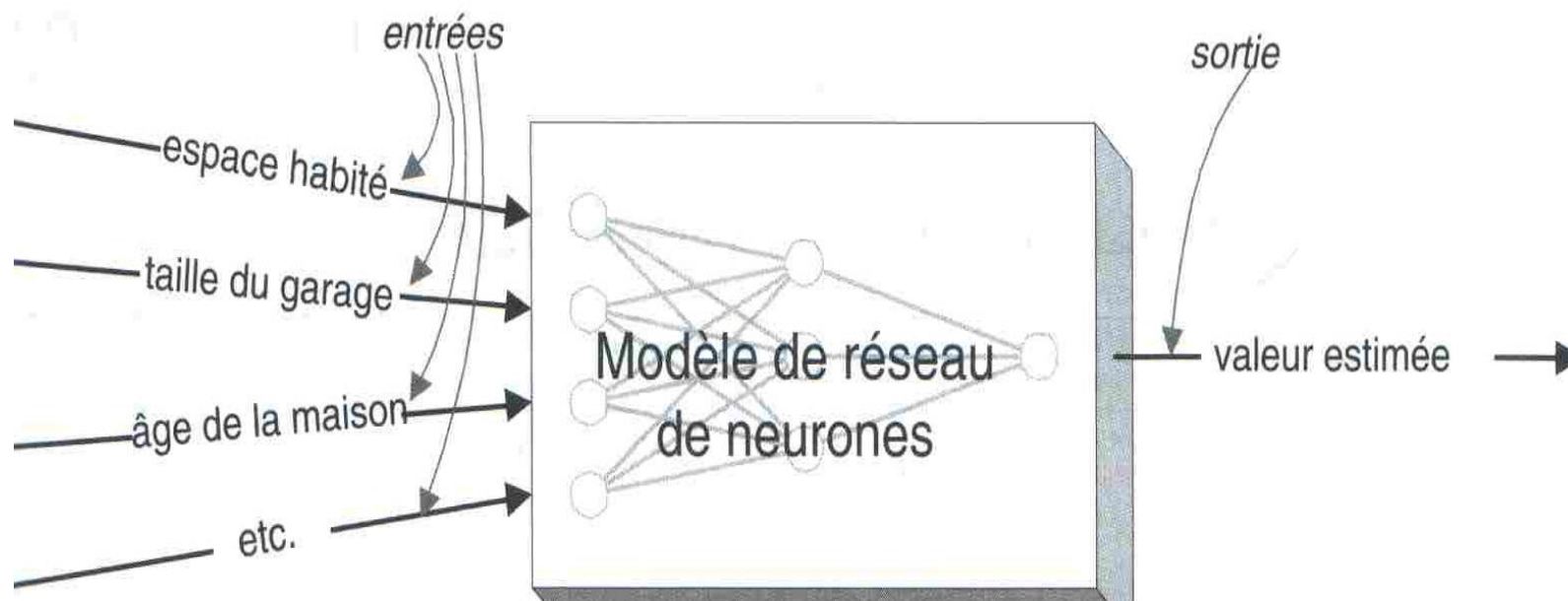
- Réception d'une information (signal)
- • Activation + Traitement (simple) par un neurone
- • Transmission aux autres neurones (si seuil franchi)
- • A la longue : renforcement de certains liens APPRENTISSAGE



L'ensemble des neurones se présente donc comme un graphe pondéré sur lequel va circuler un signal généré par des stimulus extérieurs (entrées).

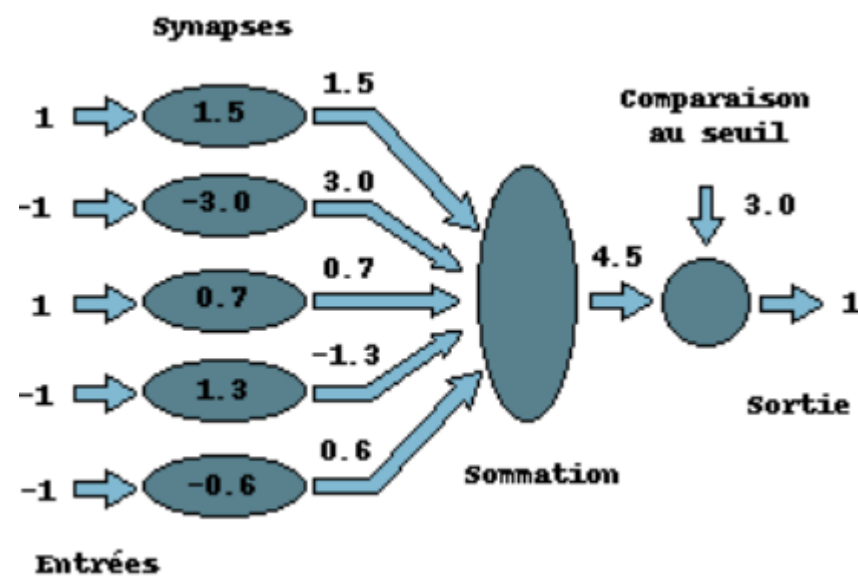
Un réseau de neurones : ensemble de nœuds connectés entre eux, chaque variable correspondant à un nœud

Illustration du réseau de neurones



Neurone formel

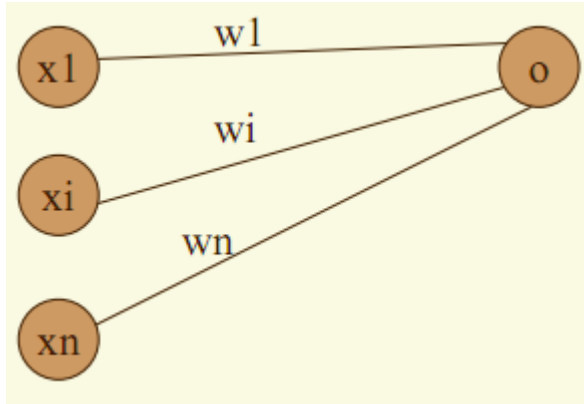
- Principes :
 - coefficient synaptique : coefficient réel
 - sommation des signaux arrivant au neurone
 - sortie obtenue après application d'une fonction de transfert



Neurone formel

Modélisation :

- Le neurone reçoit les entrées $x_1, \dots, x_i, \dots, x_n$.
- Le potentiel d'activation du neurone p est défini comme la somme pondérée : les poids sont les coefficients synaptiques w_i , des entrées.
- La sortie o est alors calculée en fonction du seuil θ



Soit : $p = x.w = x_1.w_1 + \dots + x_i.w_i + \dots + x_n.w_n$

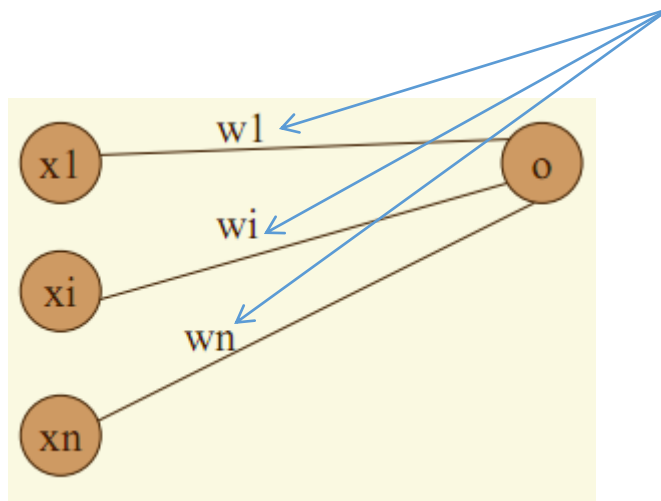
Alors : $o = 1$ si $p > \theta$

$o = 0$ si $p \leq \theta$

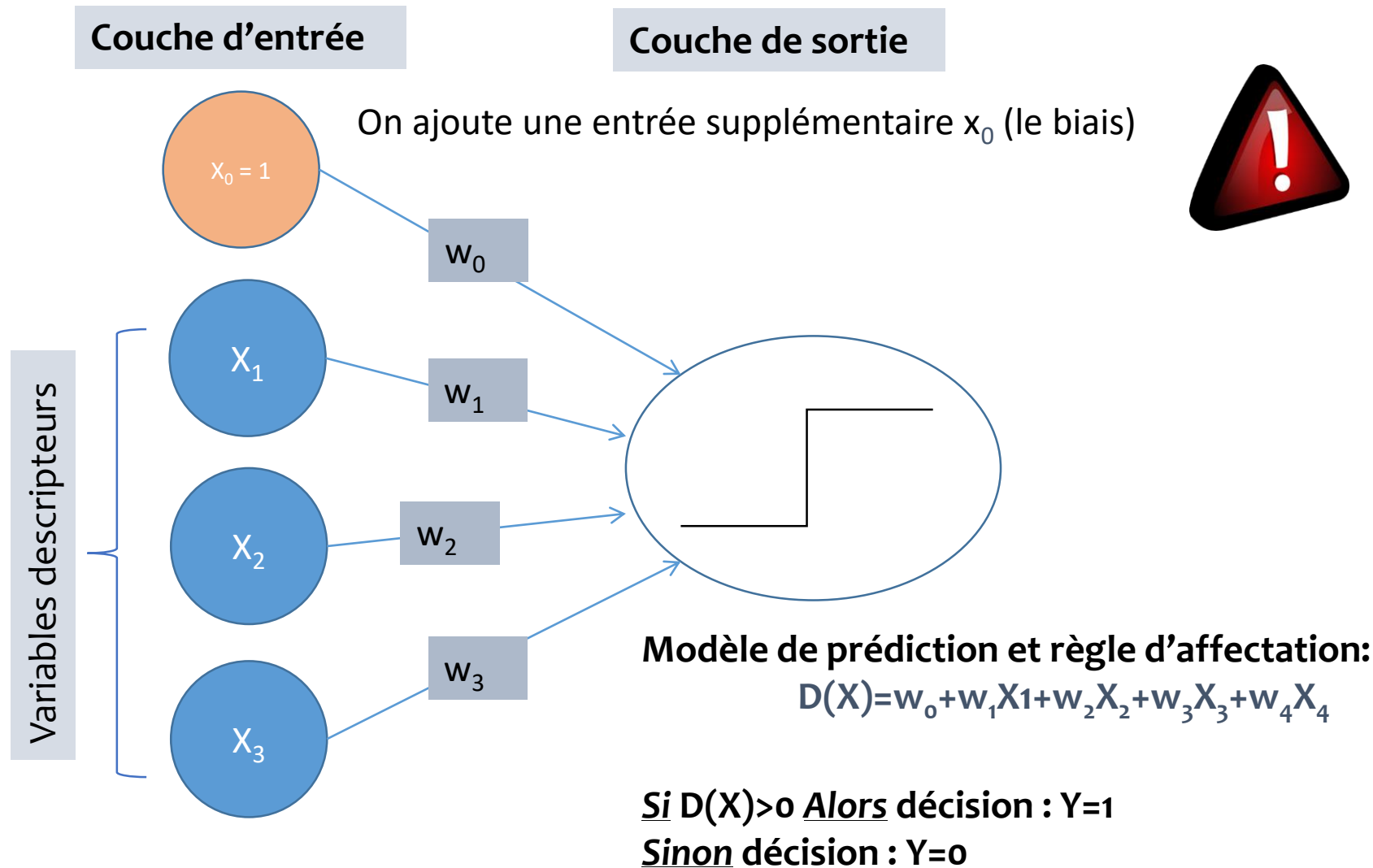
Définitions

- Déterminer un réseau de neurones = Trouver les coefficients synaptiques.
- On parle de phase d'apprentissage : les caractéristiques du réseau sont modifiées jusqu'à ce que le comportement désiré soit obtenu.

coefficients synaptiques

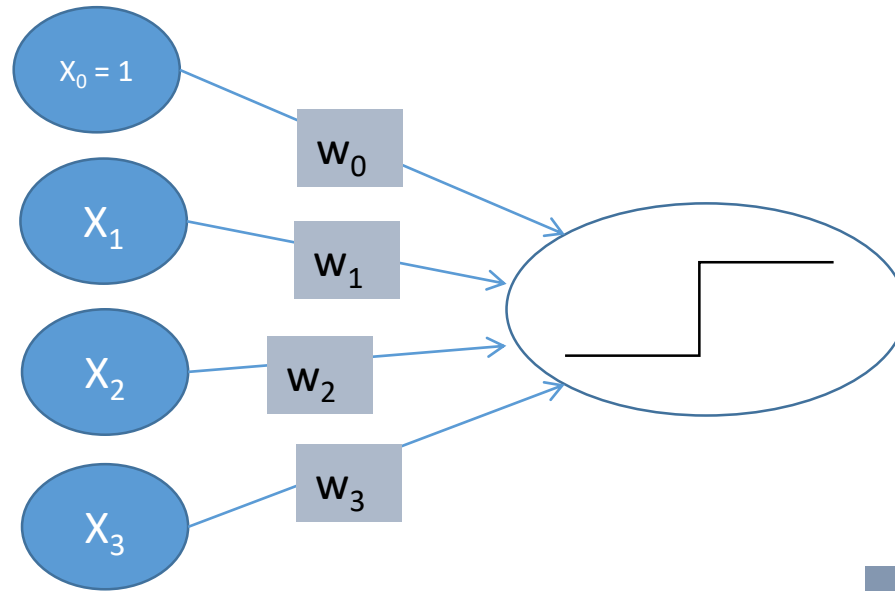


Simulation du perceptron simple



Le perceptron simple est un modèle de prédiction linéaire

Apprentissage du perceptron simple



Comment calculer les poids synaptiques
à partir d'un fichier de données

$(Y ; x_1, x_2, x_3)$?

Apprentissage par du perceptron

l'algorithme

- i. On note S la base d'apprentissage.
- ii. S est composée de couples (x, c) où :
 - x est le vecteur associé à l'entrée (x_0, x_1, \dots, x_n)
 - c la sortie correspondante souhaitée
- // On cherche à déterminer les coefficients (w_0, w_1, \dots, w_n) .
- iii. Initialiser aléatoirement les coefficients w_i .

Répéter

Prendre un exemple (x, c) dans S

Calculer la sortie o du réseau pour l'entrée x

Mettre à jour les poids :

Pour i de 0 à n

$$w_i = w_i + \epsilon * (c - o) * x_i$$

Fin Pour

Fin Répéter

Apprentissage par du perceptron

l'algorithme

- i. On note S la base d'apprentissage.
- ii. S est composée de couples (x, c) où :
 - x est le vecteur associé à l'entrée (x_0, x_1, \dots, x_n)
 - c la sortie correspondante souhaitée
- // On cherche à déterminer les coefficients (w_0, w_1, \dots, w_n) .
- iii. Initialiser aléatoirement les coefficients w_i .

Répéter

Prendre un exemple (x, c) dans S

Calculer la sortie o du réseau pour l'entrée x

Mettre à jour les poids :

Pour i de 0 à n

$$w_i = w_i + \epsilon * (c - o) * x_i$$

Fin Pour

Fin Répéter

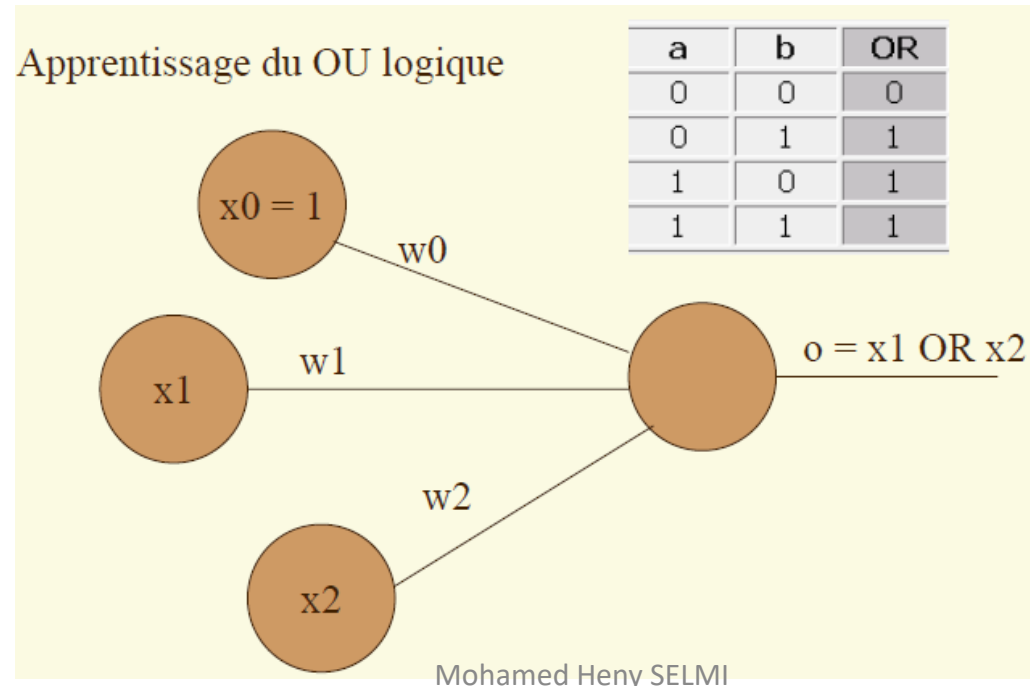
// nombre d'itérations connu d'avance ou dépassement d'un seuil défini à l'avance



Perceptron : exemple

Apprentissage par l'algorithme de perceptron du OU : les descriptions appartiennent à $\{0,1\}^2$, les entrées du perceptron appartiennent à $\{0,1\}^3$, la première composante correspond à l'entrée x_0 et vaut toujours 1, les deux composantes suivantes correspondent aux variables x_1 et x_2 .

On suppose qu'à l'initialisation, les poids suivants ont été choisis $w_0=0$; $w_1 = 1$ et $w_2 = -1$.



Perceptron : exemple

Apprentissage par l'algorithme de perceptron : exemple

$\varepsilon = 1$

x_0 vaut toujours 1

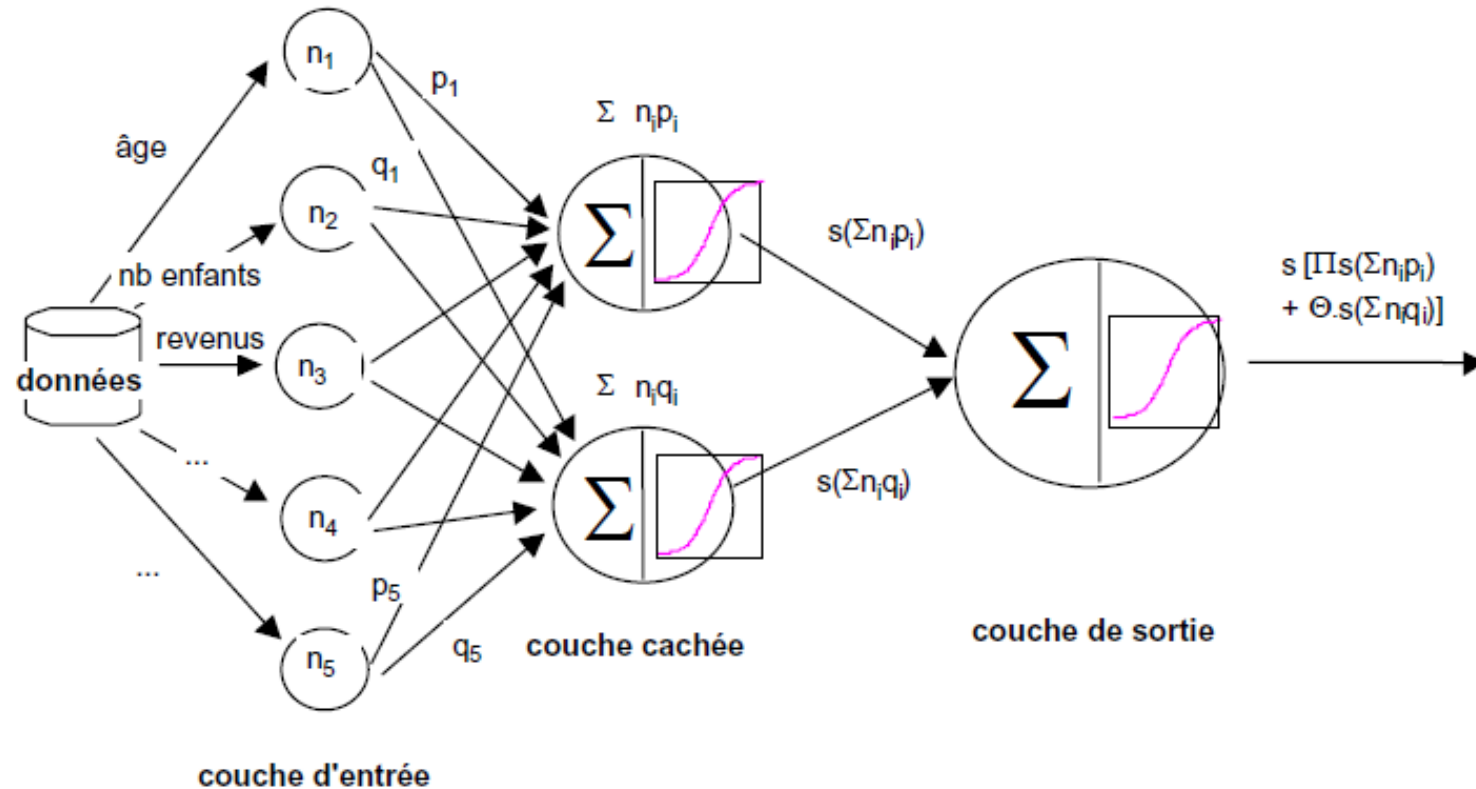
Initialisation : $w_0 = 0$; $w_1 = 1$; $w_2 = -1$

Étape	w_0	w_1	w_2	Entrée	$\sum_0^2 w_i x_i$	o	c	w_0	w_1	w_2
init								0	1	-1
1	0	1	-1	100	0	0	0	$0+0 \times 1$	$1+0 \times 0$	$-1+0 \times 0$
2	0	1	-1	101	-1	0	1	$0+1 \times 1$	$1+1 \times 0$	$-1+1 \times 1$
3	1	1	0	110	2	1	1	1	1	0
4	1	1	0	111	2	1	1	1	1	0
5	1	1	0	100	1	1	0	$1+(-1) \times 1$	$1+(-1) \times 0$	$0+(-1) \times 0$
6	0	1	0	101	0	0	1	$0+1 \times 1$	$1+1 \times 0$	$0+1 \times 1$
7	1	1	1	110	2	1	1	1	1	1
8	1	1	1	111	3	1	1	1	1	1
9	1	1	1	100	1	1	0	$1+(-1) \times 1$	$1+(-1) \times 0$	$1+(-1) \times 0$
10	0	1	1	101	1	1	1	0	1	1

Donc : $w_0 = 0$; $w_1 = 1$; $w_2 = 1$

Ce perceptron calcule le OU logique pour tout couple $(x_1 ; x_2)$

Les réseaux à couches cachées



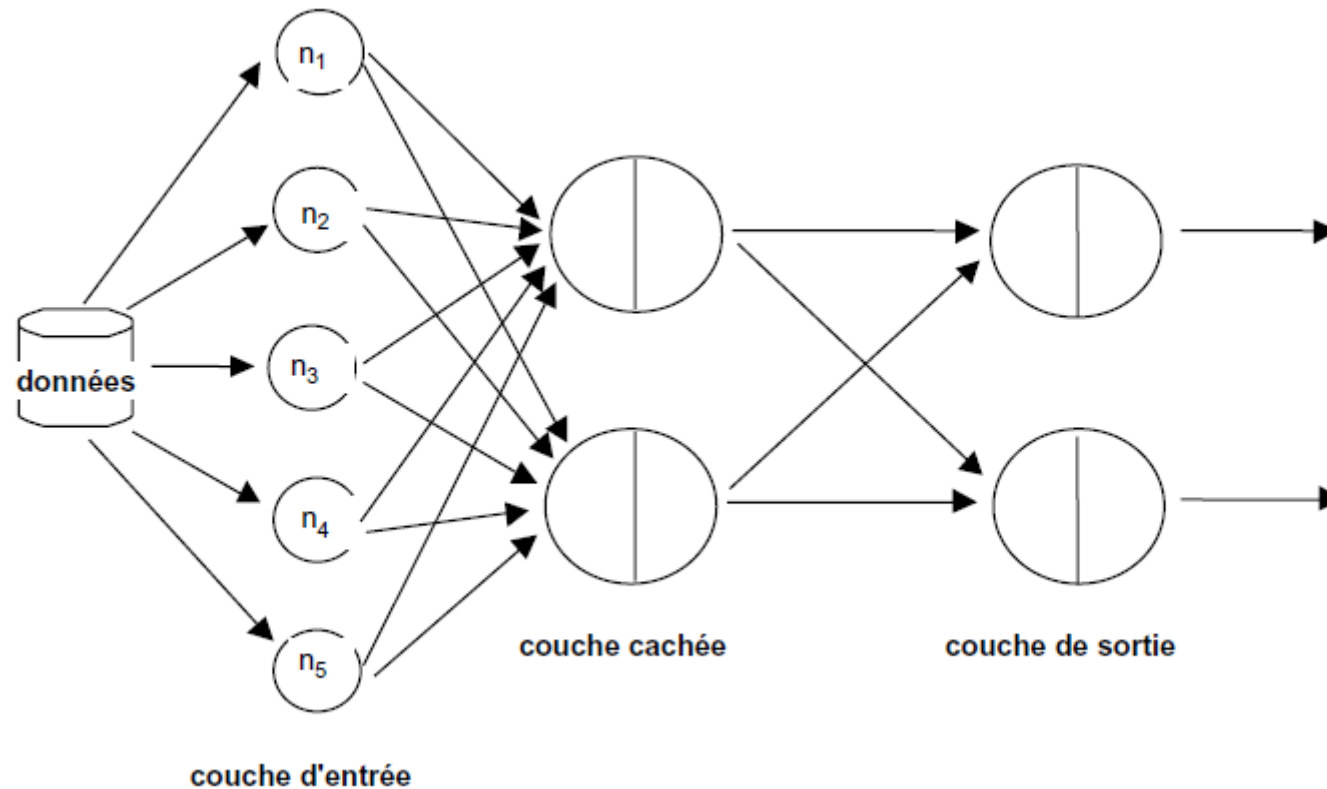
- On augmente le pouvoir de prédiction en ajoutant une ou plusieurs couches cachées entre les couches d'entrée et de sortie

Les réseaux à couches cachées

- Le pouvoir de prédiction augmente avec le nombre de nœuds des couches cachées
 - le nombre de couches cachées est très généralement 1 ou 2
 - lorsque ce nombre = 0, le réseau effectue une régression linéaire ou logistique (selon la fonction de transfert)
- Mais ce dernier doit néanmoins être limité pour que le réseau de neurones ne se contente pas de mémoriser l'ensemble d'apprentissage mais puisse le généraliser, sinon, il y a sur-apprentissage
- Le fait que toutes les valeurs soient comprises entre 0 et 1 permet de prendre en entrée d'un nœud la sortie d'un nœud précédent
- Autre but de la normalisation des valeurs : éviter que les données avec de grandes valeurs « écrasent » les autres

Les réseaux à plusieurs sorties

- La couche de sortie du réseau peut parfois avoir plusieurs nœuds, lorsqu'il y a plusieurs valeurs à prédire.



Avantages des réseaux de neurones

- Aptitude à modéliser des structures complexes et des données irrégulières
 - Prise en compte des relations non linéaires (interactions) entre les variables.
- Assez bonne robustesse aux données bruitées
- Aptitude à modéliser des problèmes très variés.

Inconvénients

- Résultats totalement non explicites
- Sensibilité aux individus hors normes
- Sensibilité à un trop grand nombre de variables non discriminantes (contrairement aux arbres de décision)
- Convergence vers la meilleure solution globale pas toujours assurée
- Paramètres nombreux et délicats à régler (nb et taille des couches cachées, taux d'apprentissage, etc...)
- Ne s'appliquent naturellement qu'aux variables continues dans l'intervalle $(0,1)$ – Nécessité de normaliser les données.

Limitations des réseaux de neurones

- Les réseaux de neurones sont bons pour la prédiction et l'estimation seulement quand :
 - Les entrées sont bien comprises
 - La sortie est bien comprise
 - L'expérience est disponible pour un grand nombre d'exemples à utiliser pour entraîner le réseau.
- Les réseaux de neurones sont seulement aussi bons que l'ensemble d'apprentissage utilisé.
- Le modèle construit est statique et doit être continuellement mis à jour avec des exemples plus récents.

Perceptron : exercice 1

Apprentissage d'un ensemble linéairement séparable :

les descriptions appartiennent à \mathbb{R}^2 , le concept cible est défini à l'aide de la droite d'équation $y=x/2$. Les couples (x,y) tels que $y > x/2$ sont de classe **1** ; Les couples (x,y) tels que $y \leq x/2$ sont de classe 0.

L'échantillon d'entrée est :

$$S=\{((0,2),1), ((1,1),1), ((1,2.5),1), ((2,0),0), ((3,0.5),0)\}.$$

On suppose qu'à l'initialisation, les poids ont été choisis :

$$w_0=0 ; w_1 = 0 \text{ et } w_2 = 0.$$

On choisit de présenter tous les exemples en alternant exemple positif (de classe 1) et exemple négatif.

Perceptron : exercice 1

L'échantillon d'entrée est :

- $S = \{((0,2),1), ((2,0),0), ((1,1),1), ((3,0.5),0), ((1,2.5),1)\}$.

On suppose qu'à l'initialisation, les poids ont été choisis :

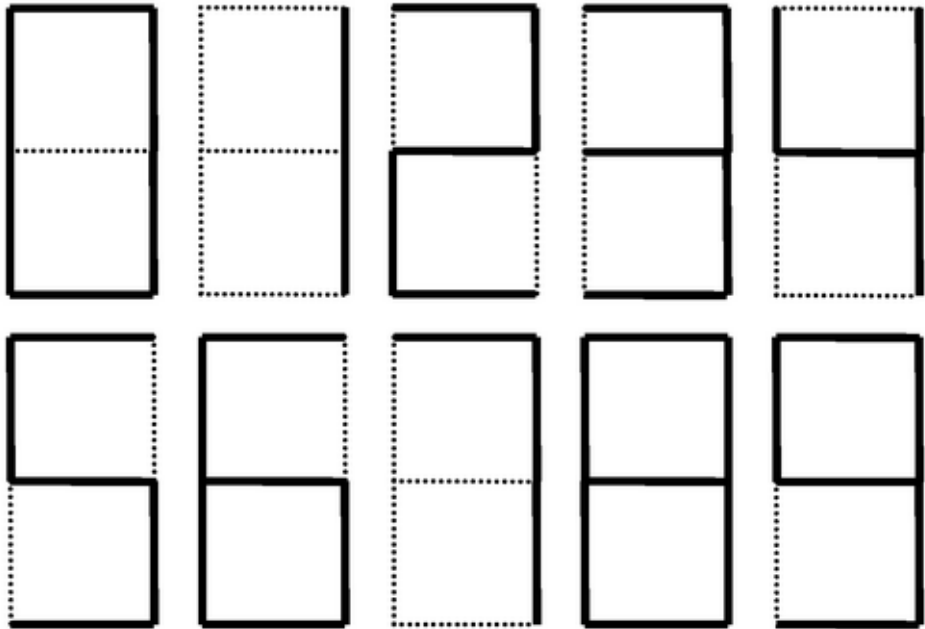
$w_0=0$; $w_1 = 0$ et $w_2 = 0$.

<i>étape</i>	w_0	w_1	w_2	<i>Entrée</i>	$\sum_0^2 w_i x_i$	o	c	w_0	w_1	w_2
<i>init</i>								0	0	0
1	0	0	0	(1,0,2)	0	0	1	1	0	2
2	1	0	2	(1,2,0)	1	1	0	0	-2	2
3	0	-2	2	(1,1,1)	0	0	1	1	-1	3
4	1	-1	3	(1,3,0.5)	-0.5	0	0	1	-1	3
5	1	-1	3	(1,1,2.5)	7.5	1	1	1	-1	3

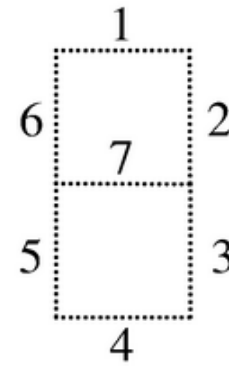
Perceptron : exercice 2

Apprentissage de parité :

Considérons un afficheur numérique à sept segments et formons un perceptron donnant la parité du chiffre écrit, à savoir 0 s'il est pair et 1 sinon.

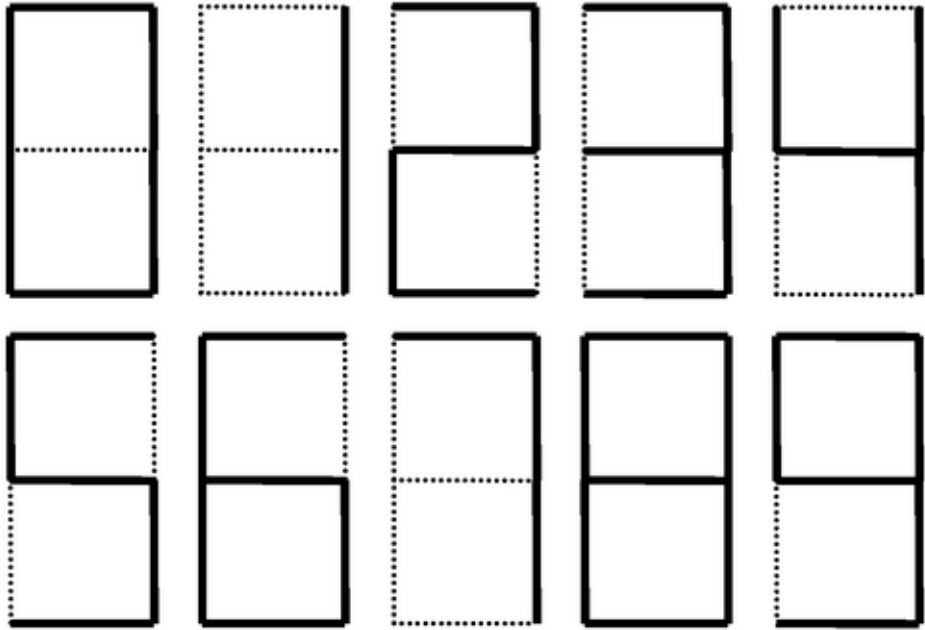


On commence par coder chaque chiffre en une liste de sept 0 ou 1 selon les segments allumés, liste qui constituera les neurones d'entrée du perceptron

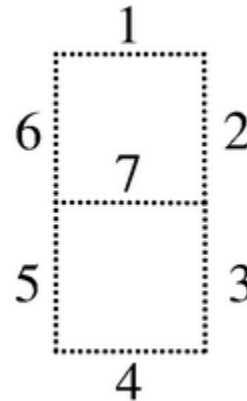


0 se code 1111110,
1 se code 0110000, etc.

Perceptron : exercice 2



On commence par coder chaque chiffre en une liste de sept 0 ou 1 selon les segments allumés, liste qui constituera les neurones d'entrée du perceptron



0 se code 1111110,
1 se code 0110000, etc.

On considère un ensemble complet

$$S = \left\{ \begin{array}{l} (1111110, 0), (0110000, 1), (1101101, 0), \\ (1111001, 1), (0110011, 0), (1011011, 1), \\ (0011111, 0), (1110000, 1), (1111111, 0), (1111011, 1) \end{array} \right\}$$

But : Apprendre si
un chiffre est pair ou
impair.

Les poids initiaux :
 $w=(1,1,1,1,1,1,1,1)$

Perceptron : exercice 2

Trace de l'algorithme

<i>Etape</i>	<i>w</i>	<i>x</i>	<i>c</i>	<i>o</i>
1	(1, 1, 1, 1, 1, 1, 1, 1)	(1, 1, 1, 1, 1, 1, 1, 0)	0	1
2	(0, 0, 0, 0, 0, 0, 0, 1)	(1, 0, 1, 1, 0, 0, 0, 0)	1	0
3	(1, 0, 1, 1, 0, 0, 0, 1)	(1, 1, 1, 0, 1, 1, 0, 1)	0	1
4	(0, -1, 0, 1, -1, -1, 0, 0)	(1, 1, 1, 1, 1, 0, 0, 1)	1	0
5	(1, 0, 1, 2, 0, -1, 0, 1)	(1, 0, 1, 1, 0, 0, 1, 1)	0	1
6	(0, 0, 0, 1, 0, -1, -1, 0)	(1, 1, 0, 1, 1, 0, 1, 1)	1	0
7	(1, 1, 0, 2, 1, -1, 0, 1)	(1, 0, 0, 1, 1, 1, 1, 1)	0	1
8	(0, 1, 0, 1, 0, -2, -1, 0)	(1, 1, 1, 1, 0, 0, 0, 0)	1	1
9	(0, 1, 0, 1, 0, -2, -1, 0)	(1, 1, 1, 1, 1, 1, 1, 1)	0	0
10	(0, 1, 0, 1, 0, -2, -1, 0)	(1, 1, 1, 1, 1, 0, 1, 1)	1	1