

Fouille de Données

Data Mining

Classification - Partie 2

Plan du cours

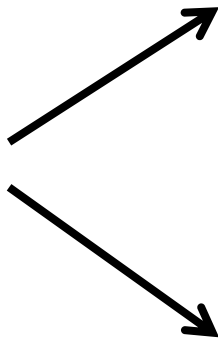
1. Classification associative
2. Méthodes d'évaluation d'un modèle
3. Combinaison de modèles
4. K plus proches voisins

Contexte

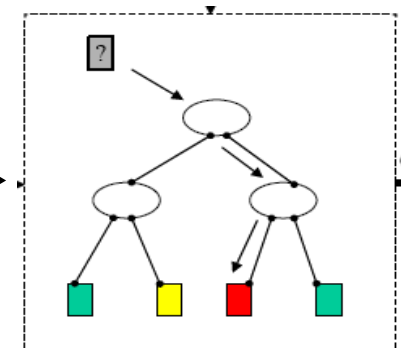
SAVOIR - PREDIRE - DECIDER



Données



Arbres de décision



Connaissances

Les arbres de décision

Algorithmes de construction d'arbres de décision

Plusieurs problèmes :

- Comment choisir l'attribut qui sépare le mieux l'ensemble d'exemples?
On parle souvent de la variable de segmentation (**Split**).
- Comment choisir les critères de séparation d'un ensemble selon l'attribut choisi, et comment ces critères varient selon que l'attribut soit numérique ou nominal ?
- Quel est le nombre optimal du nombre de critères qui minimise la taille de l'arbre et maximise la précision ?
- Quels sont les critères d'arrêt de ce partitionnement, sachant que souvent l'arbre est d'une taille gigantesque ?

Les arbres de décision

Algorithmes de construction d'arbres de décision

La bonne taille de l'arbre:

- Eviter l'overfitting – Sur-apprentissage: anomalies, bruits, erreurs, etc.
- L'arbre construit peut être d'une taille importante.
- => Opérations d'élagage.
- Eliminer les branches les moins significatives.
- Elagage avant ou après apprentissage : pré-élagage ou post-élagage.

Les arbres de décision

Choix de la bonne taille de l'arbre

Pré-élagage:

- Effectué **lors** de la construction de l'arbre.
- Au moment du calcul du gain d'information, décider de l'importance ou non de sa subdivision.
- Arrêter la construction lorsqu'il n'y a pas d'association statistiquement significative entre un attribut et la classe d'un nœud particulier.
- Couper complètement des branches qui peuvent être générées.

Les arbres de décision

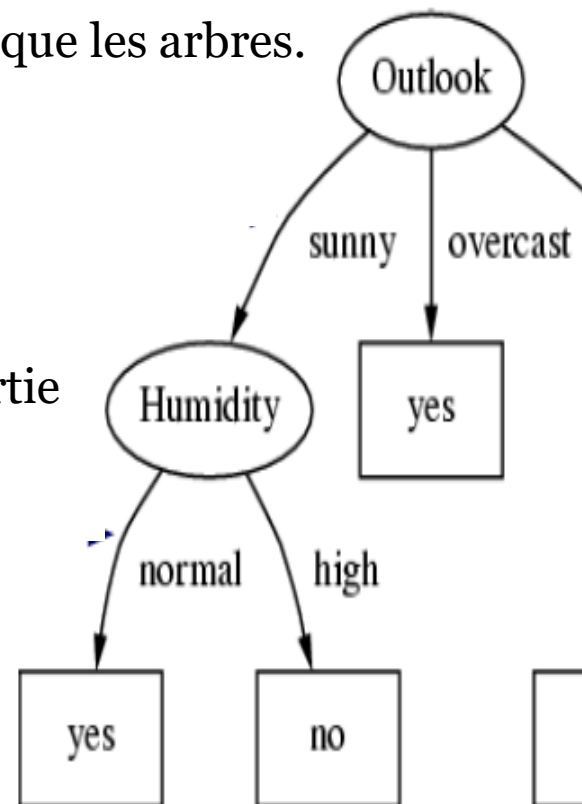
Choix de la bonne taille de l'arbre

Post-élagage:

- Effectué **après** la construction de l'arbre en coupant des sous arbres entiers et en les remplaçant par des feuilles représentant la classe la plus fréquente dans l'ensemble des données de cet arbre.
- On commence de la racine et on descend.
- Pour chaque nœud interne (non feuille), on mesure sa performance avant et après sa coupure (son remplacement par une feuille).
- Si la différence est peu importante, on coupe le sous arbre et on le remplace par une feuille.

D'arbres de décision aux règles d'association

- Les règles sont plus simples à lire et à interpréter que les arbres.
- Une règle d'association pour chaque feuille.
- Connaissance sous forme de : IF-THEN
- Extraction des règles solides qui ont dans leur partie droite l'attribut classe.
- Exemple :

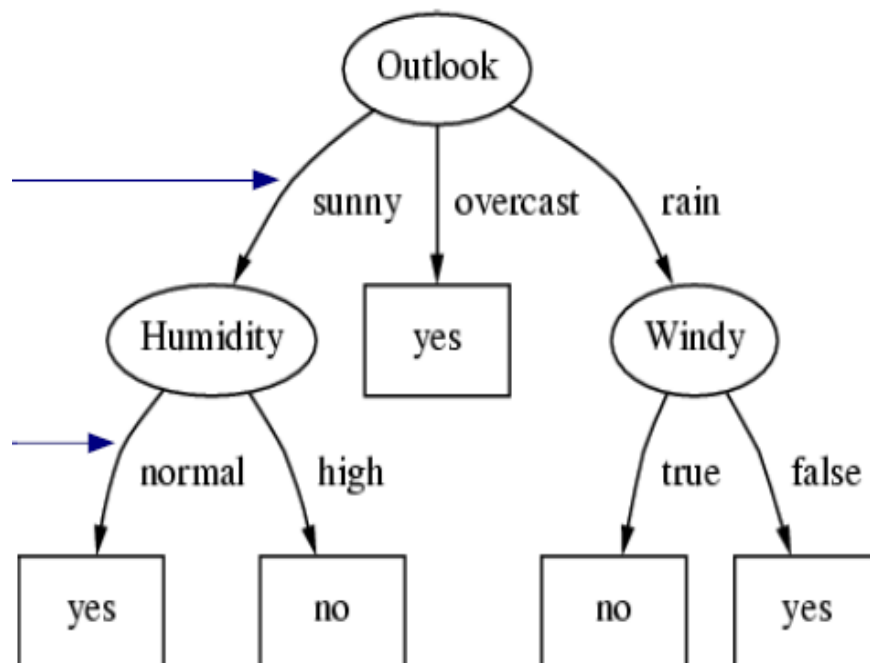


IF (Outlook=sunny) **AND** (Humidity=normal) **THEN** Play =yes

D'arbres de décision aux règles d'association

➤ Exemple :

- ✓ **IF** (Outlook=sunny) **AND** (Humidity=high) **THEN** Play =no
- ✓ **IF** (Outlook=rain) **AND** (Wind=true) **THEN** Play =no
- ✓ **OTHERWISE** play=yes



D'arbres de décision aux règles d'association

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

IF (Outlook=sunny) **and** (Humidity=high) **THEN** Play =no

D'arbres de décision aux règles d'association

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

IF (Outlook=rainy) **and** (Wind=true) **THEN** Play =no

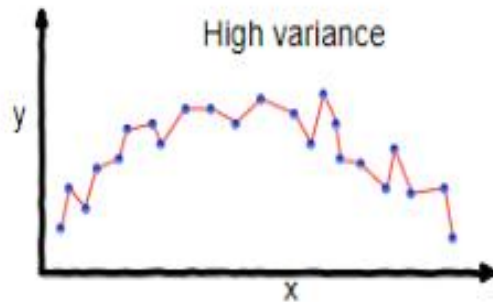
D'arbres de décision aux règles d'association

Outlook	Temp	Humidity	Windy	Play
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

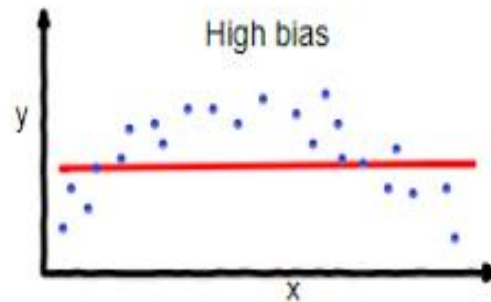
OTHERWISE play=yes

Evaluation du modèle

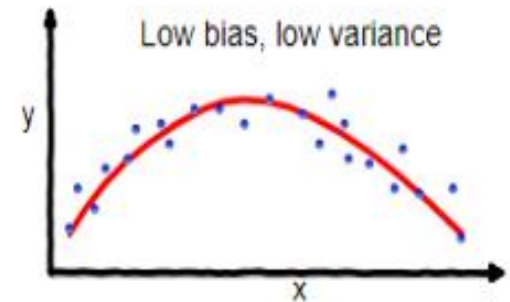
- Evaluer pour s'assurer de la capacité de généralisation en dehors des données d'entraînement.
- Overfitting/Underfitting – Variance/biais tradeoff
- Permet de qualifier le comportement du modèle appris sur les données non utilisées lors de l'apprentissage.



overfitting



underfitting



Good balance

Evaluation du modèle

- Evaluer pour s'assurer de la capacité de généralisation en dehors des données d'entraînement.
- Overfitting/Underfitting – Variance/biais tradeoff
- Permet de qualifier le comportement du modèle appris sur les données non utilisées lors de l'apprentissage.
- Sur les exemples d'entraînement ou autres exemples réservés pour le test.
- Différentes mesures :
 - ✓ Précision (Accuracy) / Taux d'erreurs
 - ✓ Sensitivité
 - ✓ Spécificité
 - ✓ Moyenne harmonique
 - ✓ Etc.

Evaluation du modèle

Précision P (Accuracy) d'un modèle

- Métrique intuitive, qui représente le rapport entre le nombre d'exemples correctement classés et le nombre total des exemples testés.
- = > Pourcentage des exemples correctement classés.
- Taux d'erreurs = $100 - P$

$$P = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$$

avec :

$$L = \begin{cases} 1 & \text{si } y_i = \hat{f}(x_i) \\ 0 & \text{sinon} \end{cases}$$

Evaluation du modèle

Matrice de confusion

- Aucune information sur la nature des erreurs.
- Exemple : considérer un échantillon non cancéreux alors qu'il l'est, est beaucoup plus grave de considérer un échantillon cancéreux alors qu'il ne l'est pas.
- Cas de classification binaire :

$$\left\{ \begin{array}{llll} \hat{f}(x_i) = \text{Yes} & \text{et} & y_i = \text{Yes} & \text{correcte positive} \\ \hat{f}(x_i) = \text{Yes} & \text{et} & y_i = \text{No} & \text{fausse positive} \\ \hat{f}(x_i) = \text{No} & \text{et} & y_i = \text{No} & \text{correcte négative} \\ \hat{f}(x_i) = \text{No} & \text{et} & y_i = \text{Yes} & \text{fausse négative} \end{array} \right.$$

Evaluation du modèle

Matrice de confusion

- **CP** : classe positive considérée positive
- **CN** : classe negative considérée negative
- **FP** : classe negative considérée positive
- **FN** : classe positive considérée negative

		Prédiction	
		Yes	No
Observation	Yes	CP	FN
	No	FP	CN

Evaluation du modèle

Matrice de confusion

- Modèle sans erreurs : $CP + CN = N$
- Multi classes : nombre de colonnes = nombre de classes
- Précision (Accuracy):

$$P = \frac{CP + CN}{CP + FP + CN + FN}$$

		Prédiction	
		Yes	No
Observation	Yes	CP	FN
	No	FP	CN

Evaluation du modèle

Matrice de confusion

Exemple:

- Chat ou Chien ?
- N = 14 exemples/animaux testés
- Deux Classes : Chat, Chien
- En réalité : **8** Chats et **6** Chiens.
- Le modèle a prédit :
 - ✓ À partir des 8 chats : 5 l'étaient seulement, les 3 restants étaient Chiens.
 - ✓ A partir des 6 Chiens : 4 l'étaient seulement, les 2 autres étaient Chats.

		Prédiction	
		Chat	Chien
Observation	Chat	CP	FN
	Chien	FP	CN

Evaluation du modèle

Matrice de confusion

Exemple:

- Chat ou Chien ?
- N = 14 exemples/animaux testés
- Deux Classes : Chat, Chien
- En réalité : **8** Chats et **6** Chiens.
- Le modèle a prédit :
 - ✓ Depuis les 8 chats : 5 l'étaient seulement, les 3 restants étaient Chiens.
 - ✓ Depuis les 6 Chiens : 4 l'étaient seulement, les 2 autres étaient Chats.

		Prédiction	
		Chat	Chien
Observation	Chat	5	3
	Chien	2	4

Evaluation du modèle

Matrice de confusion

Exemple 2:

		Prédiction		
		Chat	Chien	Lapin
Observation	Chat	5	3	0
	Chien	2	3	1
	Lapin	0	2	11

- Chat, Chien, ou Lapin ?
- N = 27 exemples/animaux testés
- Deux Classes : Chat, Chien, Lapin
- En réalité : **8** Chats et **6** Chiens, **13** Lapins.

Evaluation du modèle

Matrice de confusion

Exemple 2:

		Prédiction		
		Chat	Chien	Lapin
Observation	Chat	5	3	0
	Chien	2	3	1
	Lapin	0	2	11

		Prédiction	
		Chat	Chien&Lapin
Observation	Chat	5	3
	Chien&Lapin	2	17

Evaluation du modèle

Matrice de confusion

Autres mesures :

- Sensitivité – True Positive Rate:

$$Sv = \frac{CP}{CP + FN}$$

- Spécificité – True Negative Rate:

$$Sp = \frac{CN}{CN + FP}$$

- Moyenne harmonique :

$$Mh = \frac{2 * Sv * Sp}{Sv + Sp}$$

		Prédiction	
		Yes	No
Observation	Yes	CP	FN
	No	FP	CN

Evaluation du modèle

Matrice de confusion

Autres mesures :

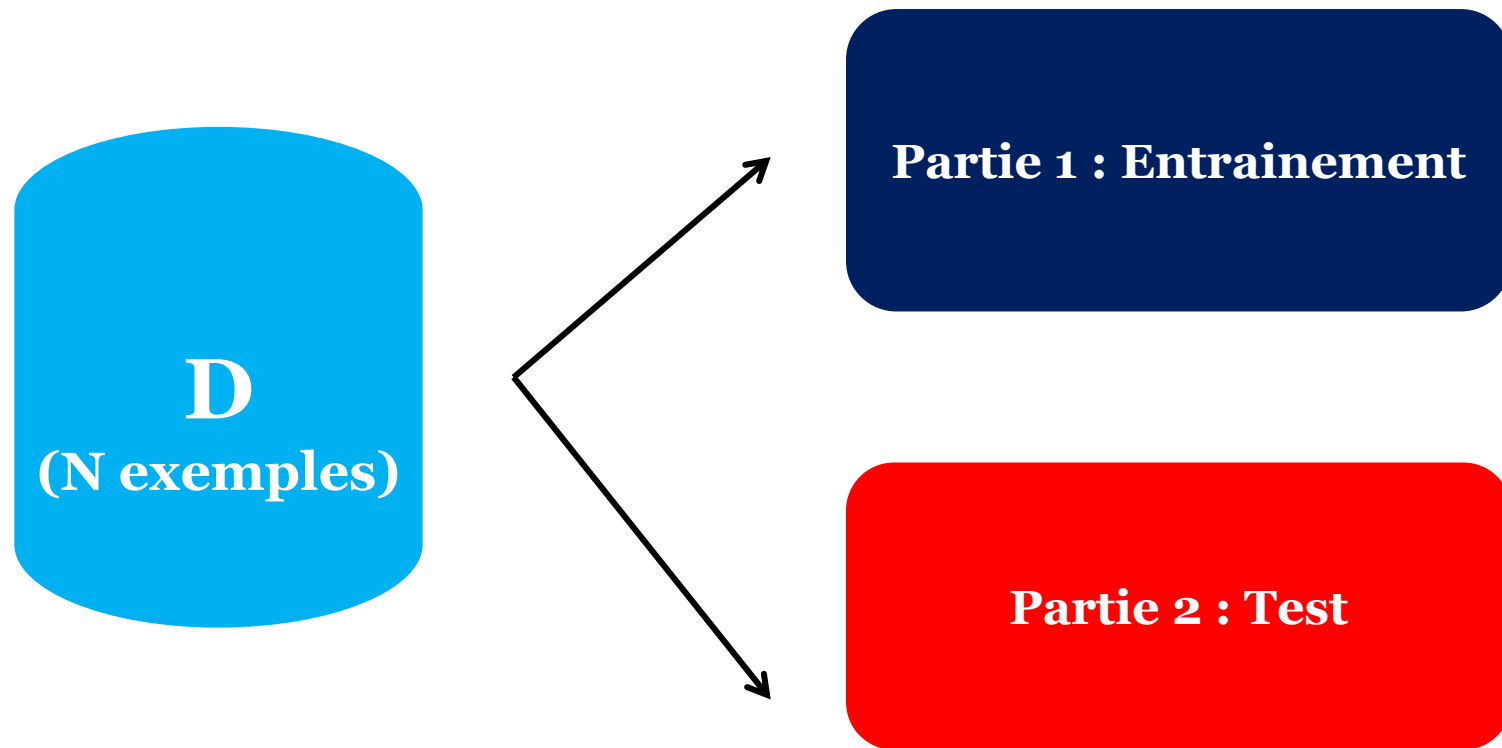
		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$ Recall
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Méthodes d'évaluation

- L'apprentissage d'un modèle se fait à base de plusieurs paramètres.
- Le choix de leurs valeurs se fait à travers plusieurs essais et évaluations.
- Paramètres optimaux => précision de 100%.
- Problème d'overfitting => Mesure de précision non suffisante.
- D'où les méthodes d'évaluation.
- Tirer des conclusions sur le comportement d'un modèle face à tout l'espace d'exemples en limitant l'influence des exemples d'entraînement.

Méthodes d'évaluation

Méthode HoldOut

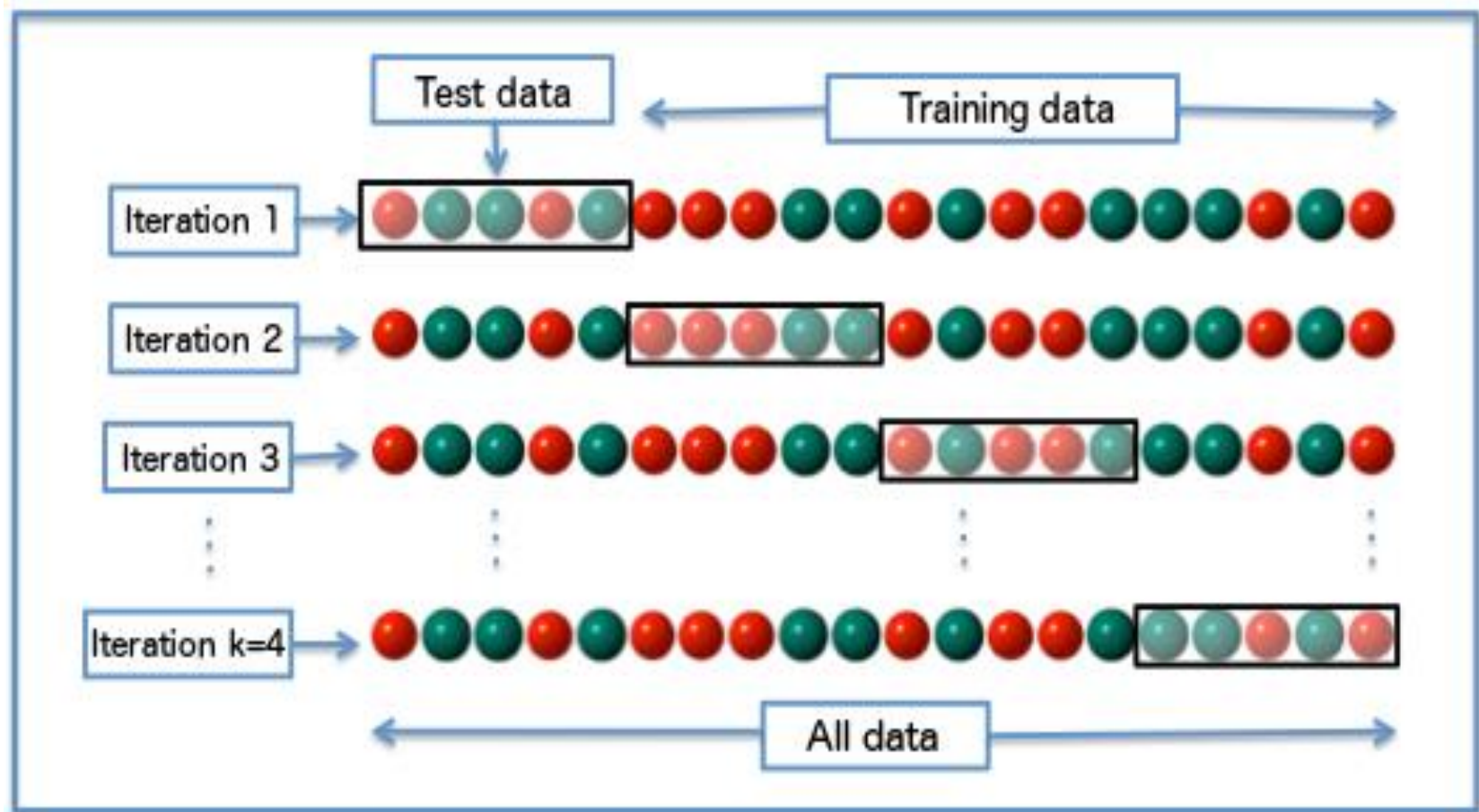


Maximiser précision Test => Maximiser précision Modèle

Méthodes d'évaluation

Méthode validation croisée

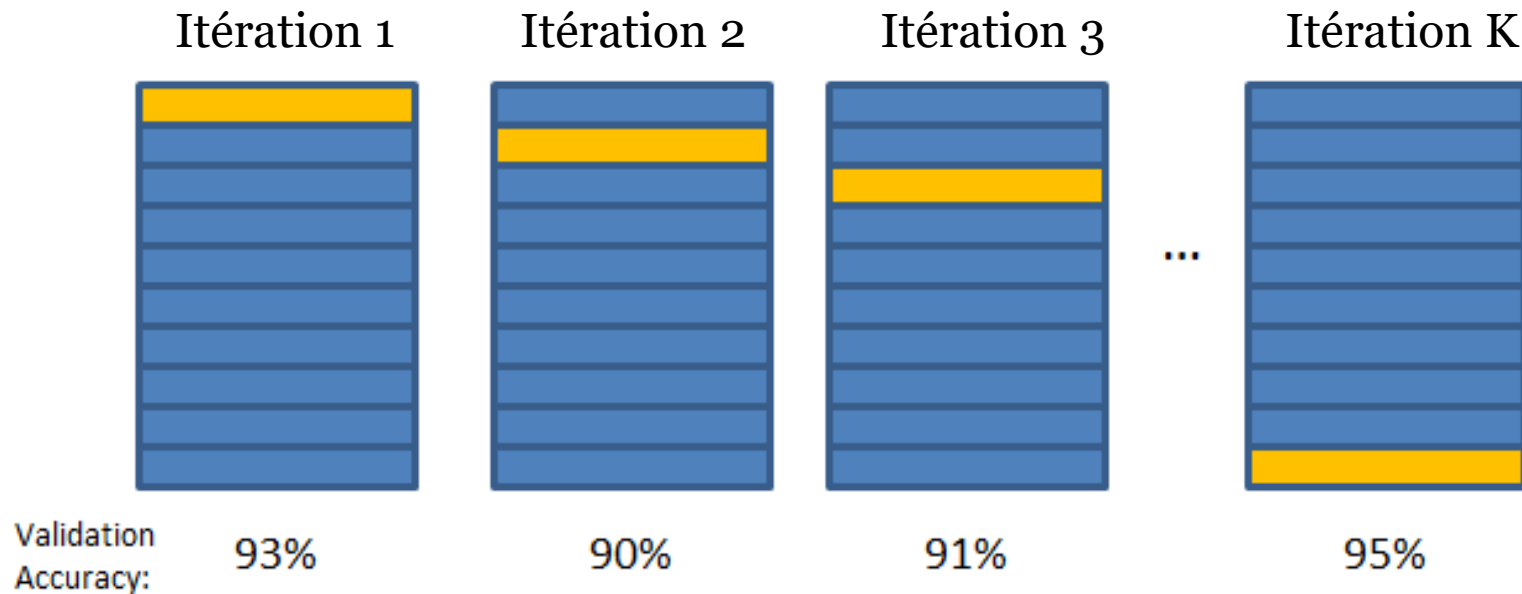
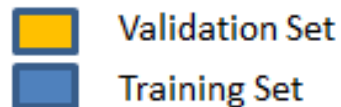
Subdiviser D en k sous-ensembles de même taille - Folds.



Méthodes d'évaluation

Méthode validation croisée

Si $k=N$ (i.e. test sur un seul exemple exclu) =>
Méthode **Leave-One-Out**



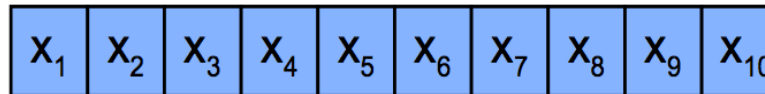
Précision Finale = Moyenne(Itération 1, Itération 2, ...)

Méthodes d'évaluation

Subdiviser ***D*** en ***k*** sous-ensembles
aléatoires – Par remplacement.

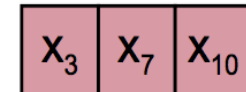
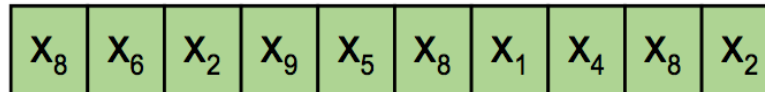
Méthode Bootstrap

Original Dataset

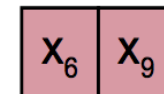
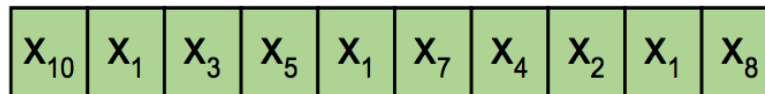


Sampling with
Replacement

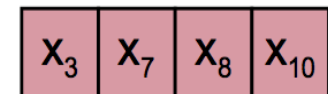
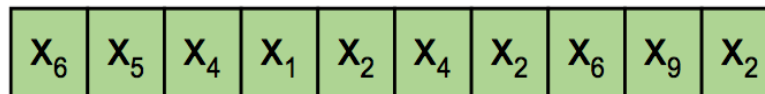
Bootstrap 1



Bootstrap 2



Bootstrap 3



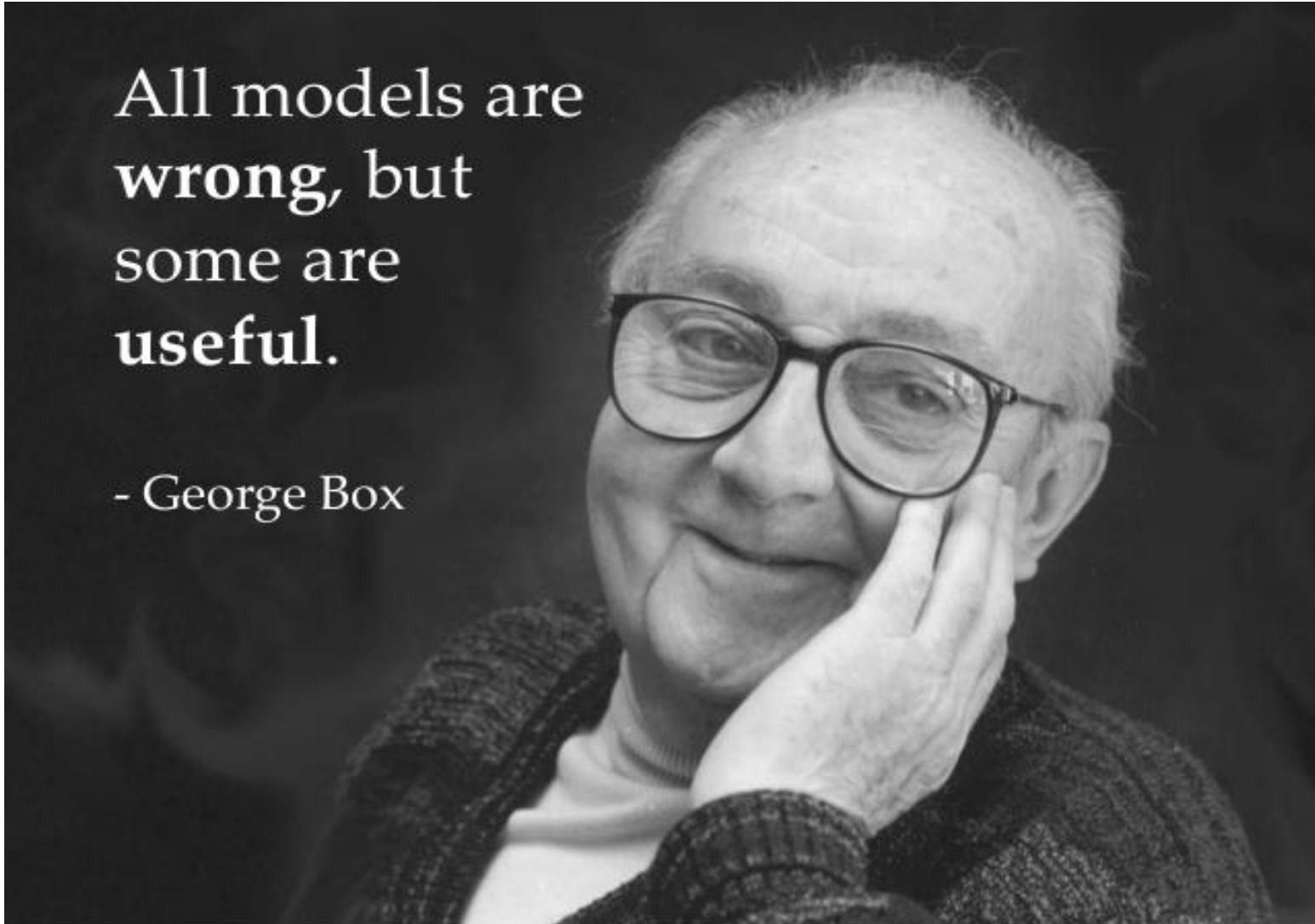
Training Sets

Test Sets

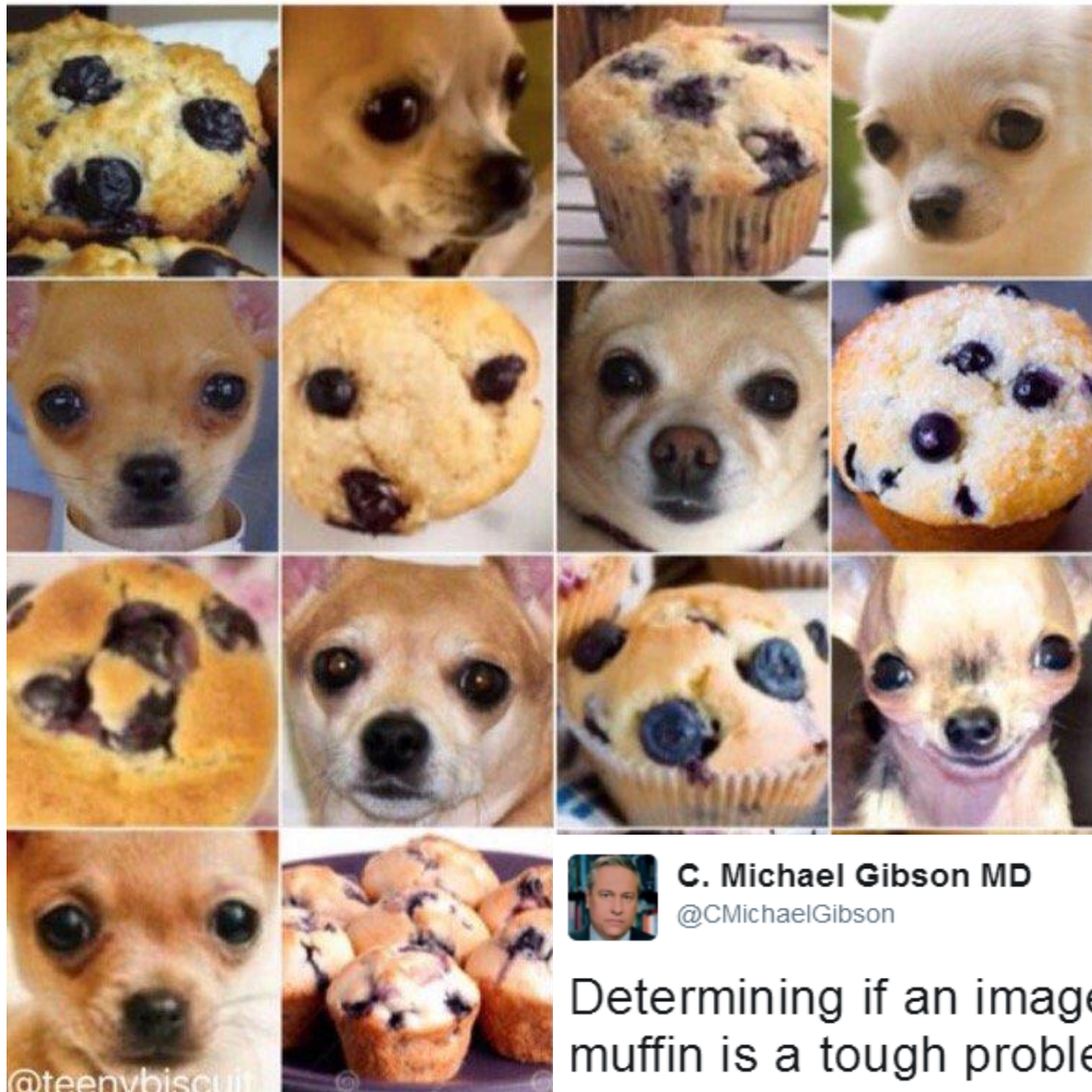
Combinaison de modèles - Ensemble Learning

All models are
wrong, but
some are
useful.

- George Box



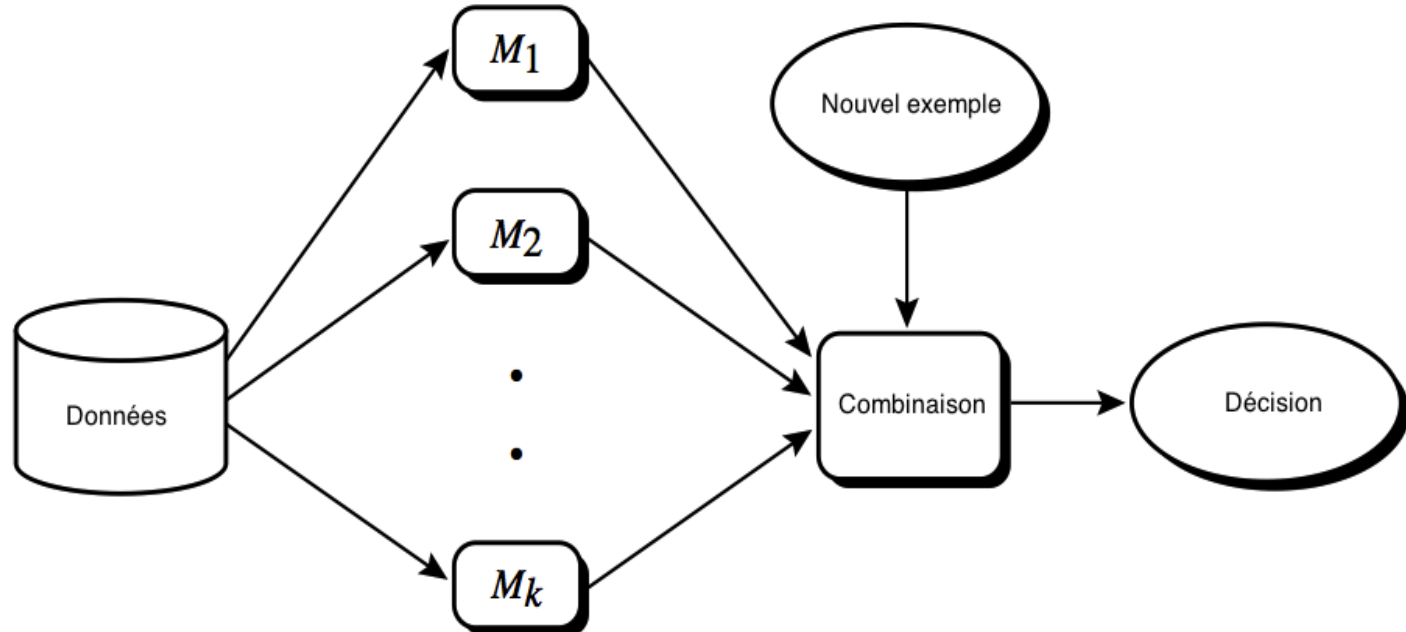
Combinaison de modèles



Determining if an image is a Chihuahua or muffin is a tough problem in artificial intelligence

Combinaison de modèles - Ensemble Learning

- Combiner plusieurs modèles avec des performances faibles (weak learners) permettant d'obtenir un modèle plus efficace (meta-learner).
- Créer un grand nombre de petits modèles rapidement puis développer un modèle qui les rassemble.



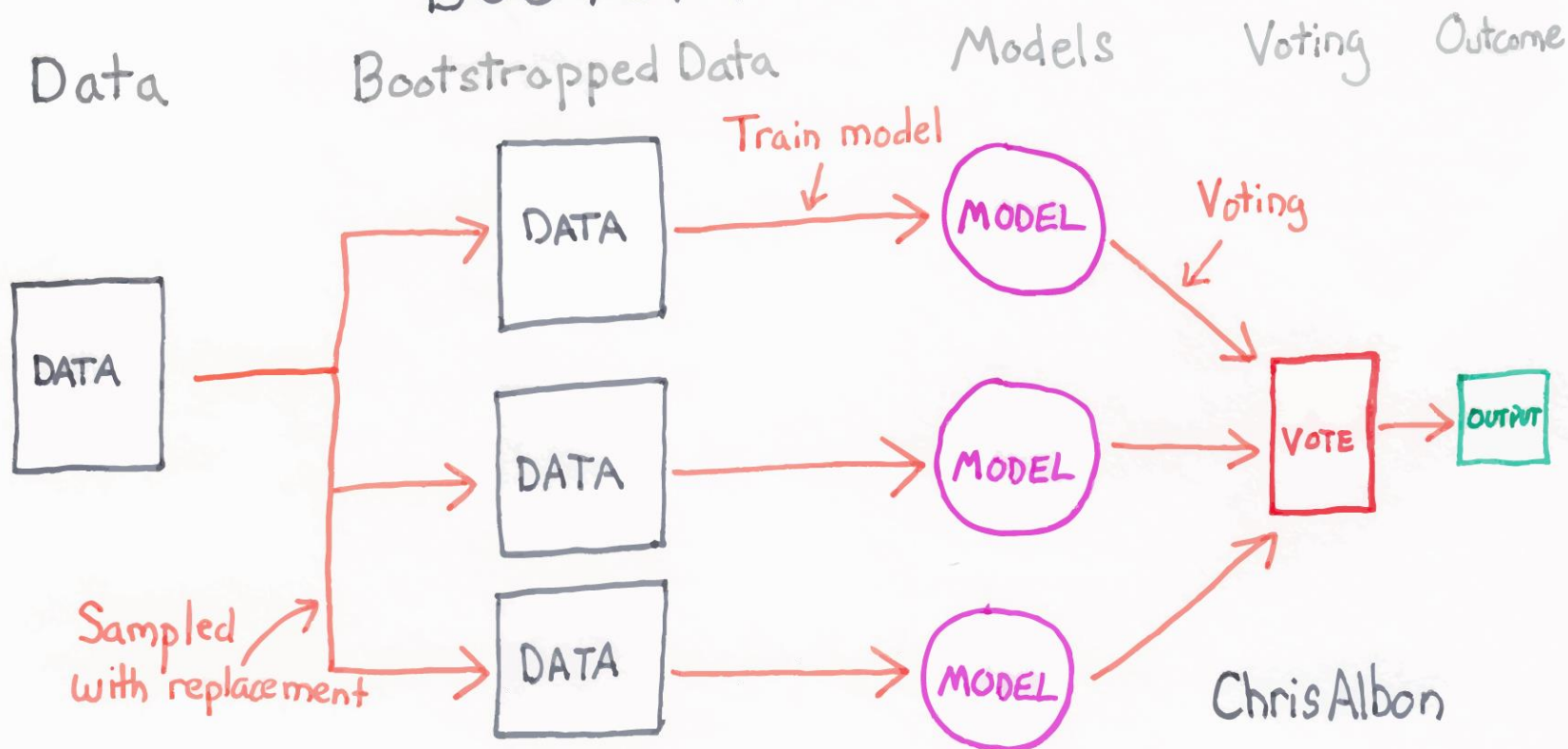
Combinaison de modèles - Ensemble Learning

- Deux méthodes : Parallèles et séquentielles.
- Méthodes **parallèles** : Bagging et Random Forest.
- Méthodes **séquentielles** : Boosting (AdaBoost, Gradient Boosting, XGBoost etc.)
- **Stacking**, Stacked Generalization (Wolpert, 1992).

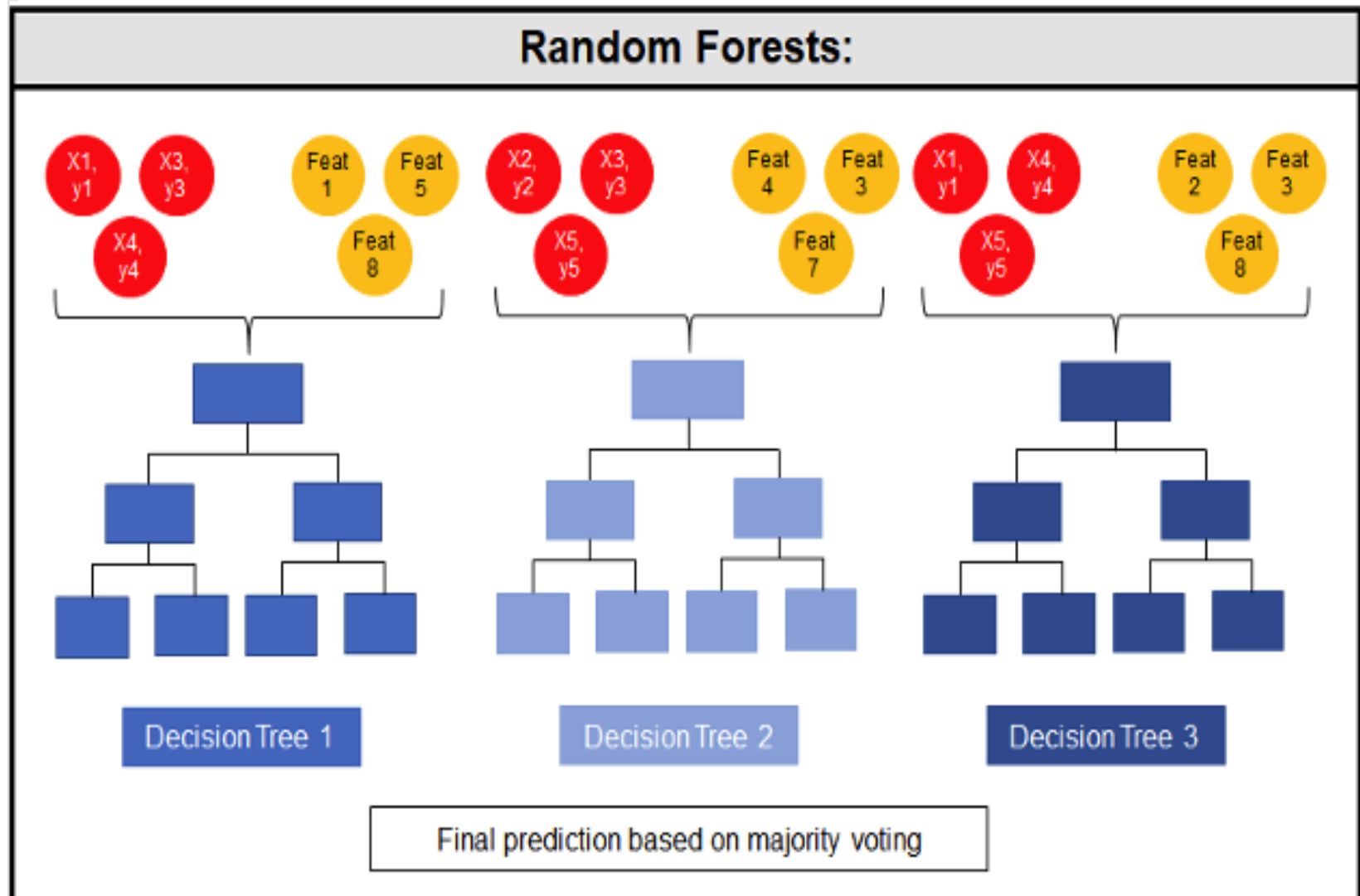
Combinaison de modèles - Ensemble Learning

BAGGING

BOOTSTRAP AGGREGATION

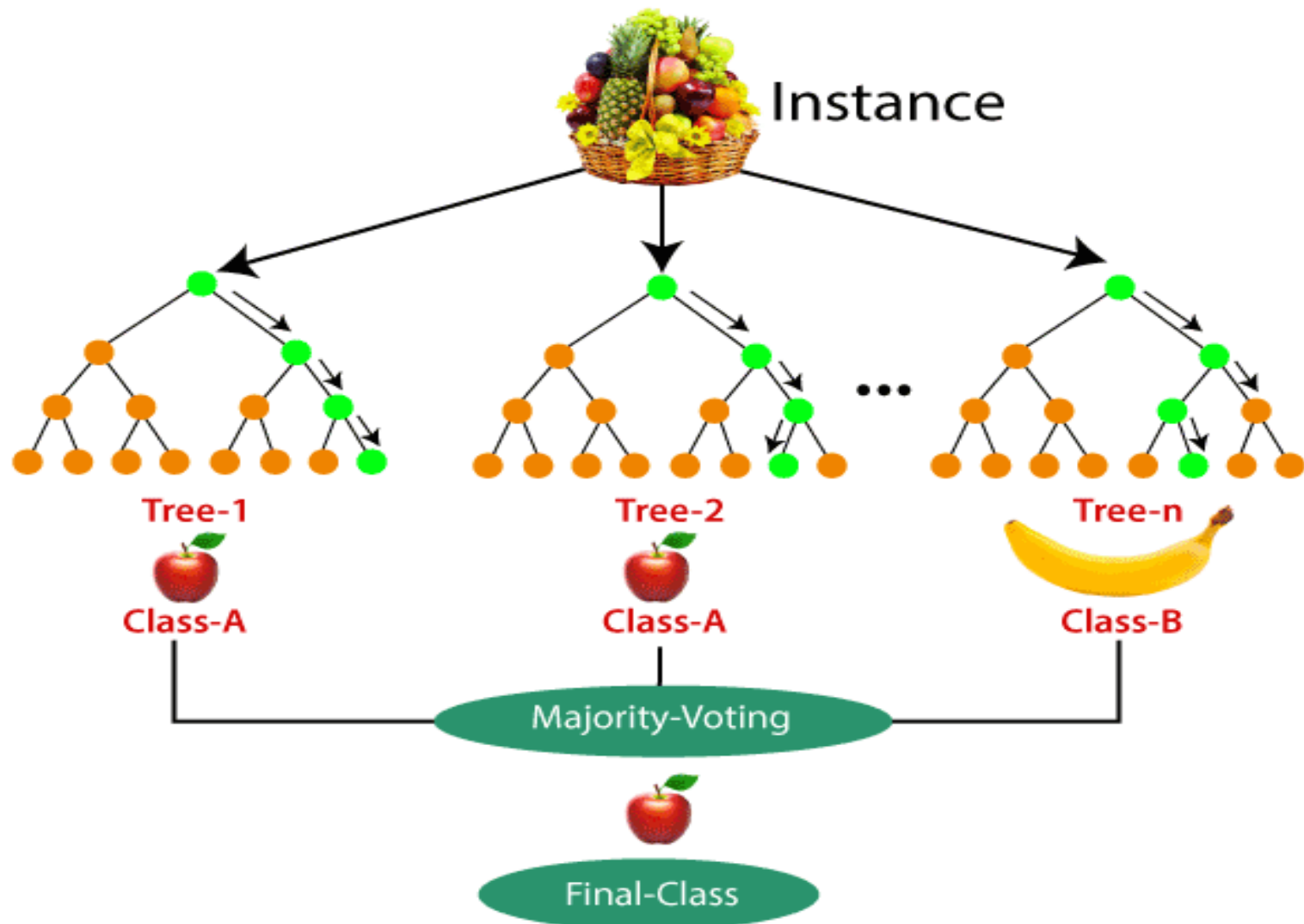


Combinaison de modèles - Ensemble Learning



Combinaison de modèles - Ensemble Learning

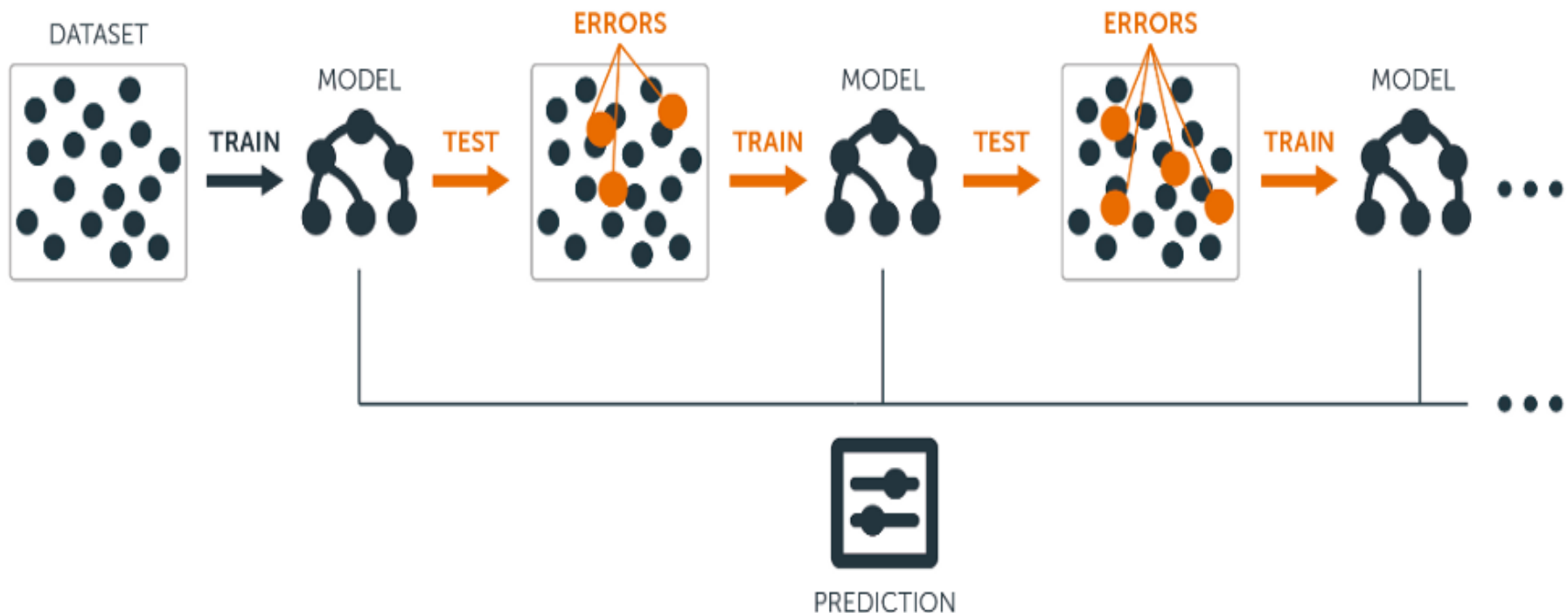
Random Forest Simplified



Combinaison de modèles - Ensemble Learning

Boosting

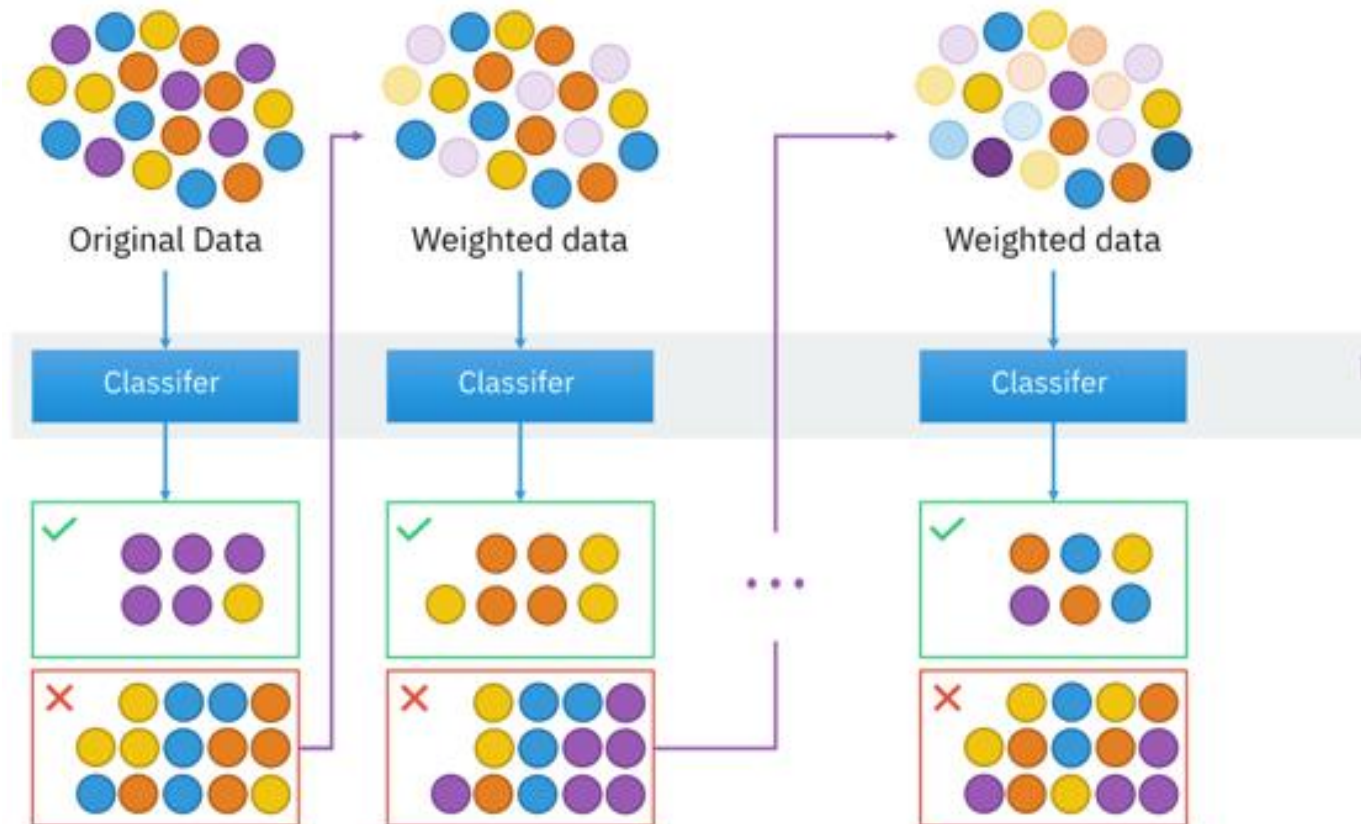
- Les exemples mal classés sont *boostés* (mettre à jour leur **poids**) pour qu'ils aient davantage d'importance vis-à-vis du classifieur faible au prochain tour, afin qu'il pallie le manque.



Combinaison de modèles - Ensemble Learning

Boosting

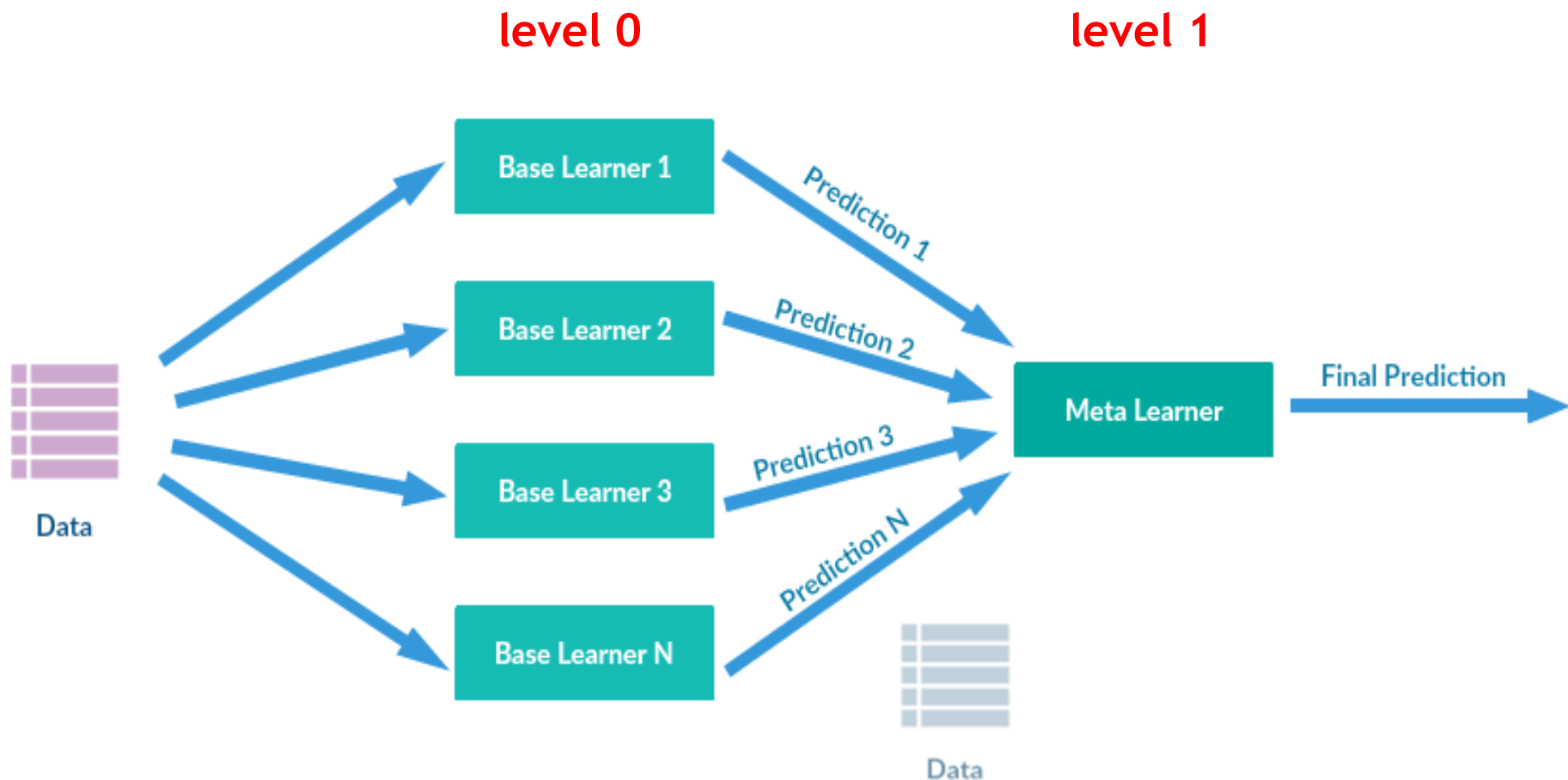
- Les exemples mal classés sont *boostés* (mettre à jour leur **poids**) pour qu'ils aient davantage d'importance vis-à-vis du classifieur faible au prochain tour, afin qu'il pallie le manque.



Combinaison de modèles - Ensemble Learning

Stacking/Blending

- Deux niveaux d'algorithmes. Principalement, algos de types différents.



Combinaison de modèles - Ensemble Learning

Stacking/Blending

- Deux niveaux d'algorithmes. Principalement, algos de types différents.

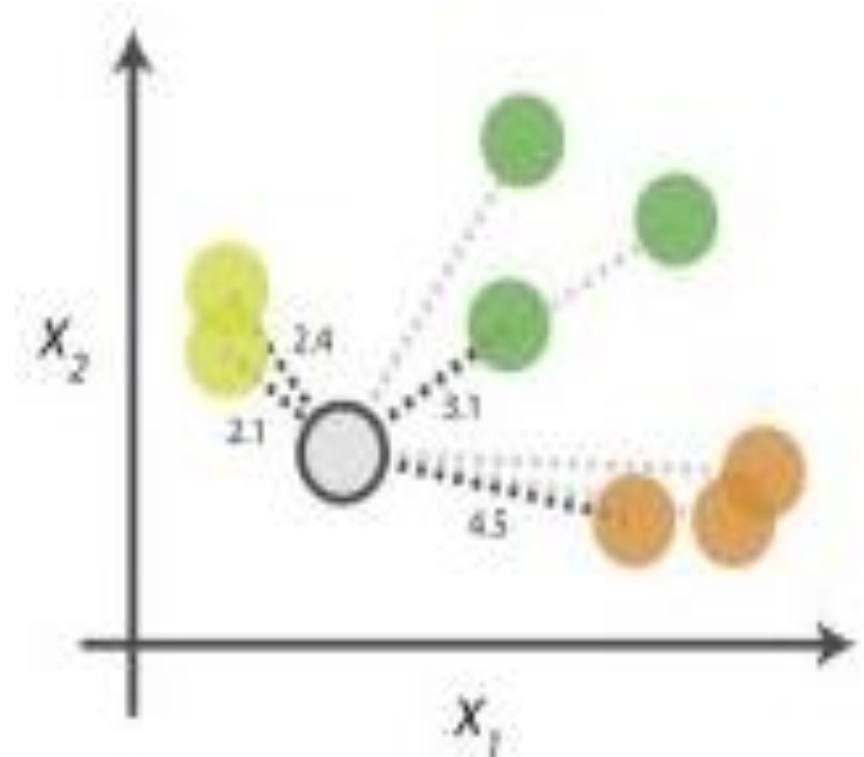
Level 1 training data

Data Point #	prediction from base learner 1	prediction from base learner 2	prediction from base learner 3	prediction from base learner M	actual
1	$y_{11}^{\hat{}}$	$y_{12}^{\hat{}}$	$y_{13}^{\hat{}}$	$y_{1M}^{\hat{}}$	y_1
2	$y_{21}^{\hat{}}$	$y_{22}^{\hat{}}$	$y_{23}^{\hat{}}$	$y_{2M}^{\hat{}}$	y_2
...
N	$y_{N1}^{\hat{}}$	$y_{N2}^{\hat{}}$	$y_{N3}^{\hat{}}$	$y_{NM}^{\hat{}}$	y_N



K plus proche voisins

- KNN – K Nearest Neighbors
- Algorithme de classification le plus simple.
- Principe : Calcul la distance entre tous les exemples de la base et le nouvel exemple qu'on cherche à classer.
- Choisir la classe majoritaire parmi les K-distances les plus petites.
- Les exemples sont représentés par des vecteurs de coordonnées.
- Distance euclidienne, Manhattan, Minkowski, Hamming, etc.



K plus proches voisins

Exemple : On pose **K = 3**

Scénario	Jeu d'acteurs	Classe
7	7	Good
7	4	Good
3	4	Bad
1	4	Bad

Test Data : Scénario = 3 , Jeu d'acteurs = 7, classe = ?

K plus proches voisins

Exemple : **K=3**

Calculer la distance :

Euclidean

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

Scénario	Jeu d'acteurs	Classe	Distance
7	7	Bon	$\text{sqrt}[(7-\mathbf{3})^2 + (7-\mathbf{7})^2] = 4$
7	4	Bon	$\text{sqrt}[(7-\mathbf{3})^2 + (4-\mathbf{7})^2] = 5$
3	4	Mauvais	3
1	4	Mauvais	3.60

Test Data : Scénario = **3** , Jeu d'acteurs = **7** , classe = **?**

K plus proches voisins

Exemple : **K = 3**

Les 3 plus proches exemples

Euclidean

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

3

1

2

Scénario	Jeu d'acteurs	Classe	Distance
7	7	Bon	$\text{sqrt}[(7-3)^2 + (7-7)^2] = 4$
7	4	Bon	$\text{sqrt}[(7-3)^2 + (4-7)^2] = 5$
3	4	Mauvais	3
1	4	Mauvais	3.60

Test Data : Scénario = **3** , Jeu d'acteurs = **7** , classe = **?**

K plus proches voisins

Exemple : **K=3**

Choix de la classe majoritaire

Euclidean

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

3

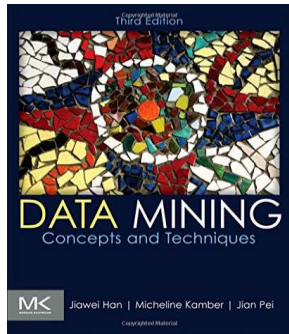
1

2

Scénario	Jeu d'acteurs	Classe	Distance
7	7	Bon	$\text{sqrt}[(7-\mathbf{3})^2 + (7-\mathbf{7})^2] = 4$
7	4	Bon	$\text{sqrt}[(7-\mathbf{3})^2 + (4-\mathbf{7})^2] = 5$
3	4	Mauvais	3
1	4	Mauvais	3.60

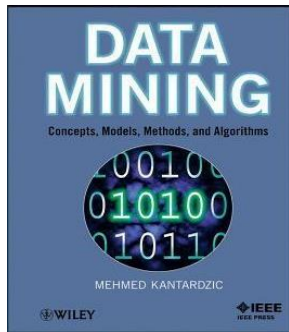
Test Data : Scénario = **3** , Jeu d'acteurs = **7** , classe = **Mauvais**

Ressources



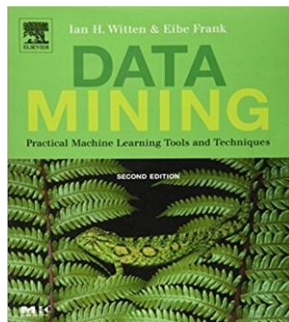
Data Mining : concepts and techniques, 3rd Edition

- ✓ Auteur : Jiawei Han, Micheline Kamber, Jian Pei
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2011 - 744 pages - ISBN 9780123814807



Data Mining : concepts, models, methods, and algorithms

- ✓ Auteur : Mehmed Kantardzi
- ✓ Éditeur : John Wiley & Sons
- ✓ Edition : Aout 2011 – 552 pages - ISBN : 9781118029121



Data Mining: Practical Machine Learning Tools and Techniques

- ✓ Auteur : Ian H. Witten & Eibe Frank
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2005 - 664 pages - ISBN : 0-12-088407-0

Ressources

Cours – Abdelhamid DJEFFAL – Fouille de données avancée

✓ www.abdelhamid-djeffal.net

WekaMOOC – Ian Witten – Data Mining with Weka

✓ <https://www.youtube.com/user/WekaMOOC/featured>

Cours - Laboratoire ERIC Lyon - DATA MINING et DATA SCIENCE

✓ https://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html

Gregory Piatetsky-Shapiro - KDNuggets

✓ <http://www.kdnuggets.com/>