

# Data Mining

**John Samuel**

CPE Lyon

**Year:** 2020-2021

**Email:** john(dot)samuel(at)cpe(dot)fr



## Objectifs

1. Régularités
2. Exploration des données
3. Algorithmes
4. Sélection de caractéristiques

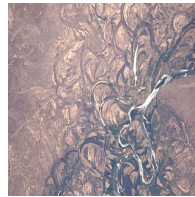
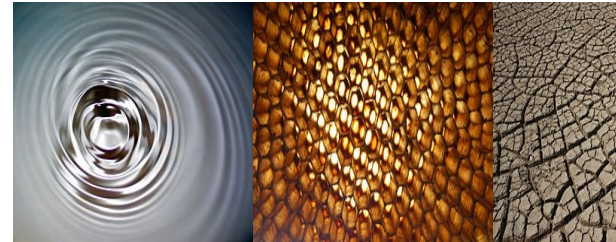
# 1. Régularités



# 1. Régularités

## Régularités naturelles

- Symétrie
- Arbres, fractales
- Spirales
- Chaos
- Ondes
- Bulles, mousse
- Pavages
- Ruptures
- Taches, bandes



# 1. Régularités

## Créations humaines

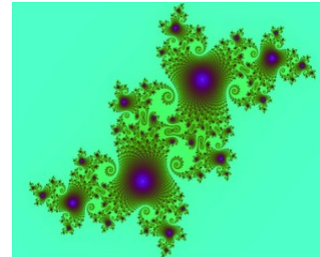
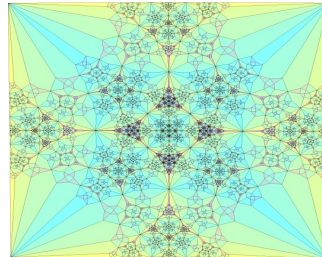
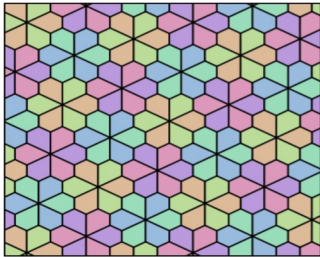
- Bâtiments (Symétrie)
- Villes
- Environnement virtuel (e.g., jeux de vidéo)
- Les artefacts humains



# 1. Régularités

## Création

- Répétition
- Fractales
  - Ensemble de Julia:  $f(z) = z^2 + c$



# 1. Régularités

## Synonymes

- Fouille de données
- Forage de données
- Extraction de connaissances à partir de données
- Data mining
- Machine learning
- Apprentissage automatique

# 1. Régularités

## Reconnaissance de formes

- Identifier des motifs informatiques à partir de données brutes
- Approches
  1. **Apprentissage supervisé**: Apprentissage automatique qui utilise un ensemble de données étiquetées
  2. **Apprentissage non-supervisé**: Apprentissage automatique qui utilise un ensemble de données non-étiquetées
  3. **Apprentissage semi-supervisé**: Apprentissage automatique qui utilise un ensemble de données étiquetées et non-étiquetées



## Formalisation

- **Vecteur euclidien**: objet géométrique avec magnitude et direction
- **Espace vectoriel**: collection de vecteurs qui peuvent être additionnés et multipliés par des nombres
- **Vecteur de caractéristiques**: vecteur n-dimensionnel
- **Espace de caractéristiques**: Espace vectoriel associé aux vecteurs

## Exemples de caractéristiques

- **Images**: les valeurs des pixels.
- **Textes**: Fréquence d'apparition des phrases textuelles.

## Formalisation

- **Construction de caractéristiques**<sup>1</sup>: construction of new features from already available features
- **Opérateurs de construction pour les caractéristiques**
  - Opérateurs d'égalité, opérateurs arithmétiques, opérateurs de tableau (min, max, moyenne, etc.)...

## Exemple

- Soit **Année de naissance** et **Année de décès** deux caractéristiques existantes.
- Une nouvelle caractéristique appelée **âge** est créée. **âge** = **Année de décès** - **Année de naissance**

1. [https://en.wikipedia.org/wiki/Feature\\_vector](https://en.wikipedia.org/wiki/Feature_vector)

## Formalisation: Supervised learning

- Soit  $N$  le nombre d'exemples d'entraînement
- Soit  $X$  l'espace de saisie des caractéristiques
- Soit  $Y$  l'espace des caractéristiques de sortie (des étiquettes)
- Soit  $(x_1, y_1), \dots, (x_N, y_N)$  les  $N$  exemples d'entraînement, où
  - $x_i$  est le vecteur de caractéristiques de  $i^{\text{ème}}$  exemple d'entraînement.
  - $y_i$  est son label.
- L'objectif de l'algorithme d'apprentissage supervisé est de trouver  $g : X \rightarrow Y$ , où
  - $g$  est l'une des fonctions de l'ensemble des fonctions possibles  $G$  (espace des hypothèses)
- **Fonction d'évaluation  $F$**  indiquent l'espace des fonctions d'évaluation, où
  - $f : X \times Y \rightarrow \mathbb{R}$  telle que  $g$  renvoie la fonction d'évaluation la plus élevée.

## Formalisation: Apprentissage non supervisé

- Soit  $X$  l'espace de saisie des caractéristiques
- Soit  $Y$  l'espace des caractéristiques de sortie (des étiquettes)
- L'objectif de l'algorithme d'apprentissage non supervisé est
  - trouver la mise en correspondance  $X \rightarrow Y$

## Formalisation: Apprentissage semi-supervisé

- Soit  $X$  l'espace de saisie des caractéristiques
- Soit  $Y$  l'espace des caractéristiques de sortie (des étiquettes)
- Soit  $(x_1, y_1), \dots, (x_l, y_l)$  l'ensemble d'exemples d'exercices étiquetés
- Soit  $x_{l+1}, \dots, x_{l+u}$  sont les  $u$  ensembles des vecteurs de caractéristiques non étiquetées de  $X$ .
- L'objectif de l'algorithme d'apprentissage semi-supervisé est de faire
  - **l'apprentissage transductif**, c'est-à-dire trouver des étiquettes correctes pour  $x_{l+1}, \dots, x_{l+u}$ .
  - **l'apprentissage inductif**, c'est-à-dire trouver la bonne mise en correspondance  $X \rightarrow Y$

### Activités

1. Classification
2. Partitionnement de données (Clustering)
3. Régression
4. Étiquetage des séquences
5. Règles d'association
6. Détection d'anomalies
7. Récapitulation

# 2.1. Classification

## 2.1.1 Introduction

- Catégorisation algorithmique d'objets.
- Attribuer une classe ou catégorie à chaque objet (ou individu)
- Classification binaire ou classification en classes multiples

### Applications

- Filtrage de contenu (e.g., spam/pourriel)
- Classification de documents
- Reconnaissance de l'écriture manuscrite
- Reconnaissance automatique de la parole
- Moteurs de recherche



## 2.1. Classification

### 2.1.2 Définition formelle

- Soit  $X$  l'espace de saisie des caractéristiques
- Soit  $Y$  l'espace des caractéristiques de sortie (des étiquettes)
- L'objectif de l'algorithme de classification (ou classificateur) est de trouver  $(x_1, y_1), \dots, (x_l, y_k)$ , c'est-à-dire l'attribution d'une étiquette connue à chaque vecteur de caractéristique d'entrée, où
  - $x_i \in X$
  - $y_i \in Y$
  - $|X| = l$
  - $|Y| = k$
  - $l \geq k$

## 2.1. Classification

### 2.1.3. Classificateurs

- Algorithme de classification
- Deux types de classificateurs:
  - **Classificateurs binaires** attribue un objet à l'une des deux classes
  - **Classificateurs multiclassés** attribue un objet à une ou plusieurs classes

## 2.1. Classification

### 2.1.4 Linear Classificateurs

- Fonction linéaire attribuant un score à chaque catégorie possible en combinant le vecteur de caractéristiques d'une instance avec un vecteur de poids, en utilisant un produit de points.
- Formalisation :
  - Soit  $X$  être l'espace de saisie des caractéristiques et  $x_i \in X$
  - Soit  $\beta_k$  un vecteur de poids pour la catégorie  $k$
  - $\text{score}(x_i, k) = x_i \cdot \beta_k$ , score pour l'attribution de la catégorie  $k$  à l'instance  $x_i$ . La catégorie qui donne le score le plus élevé est attribuée à la catégorie de l'instance.

## 2.1. Classification

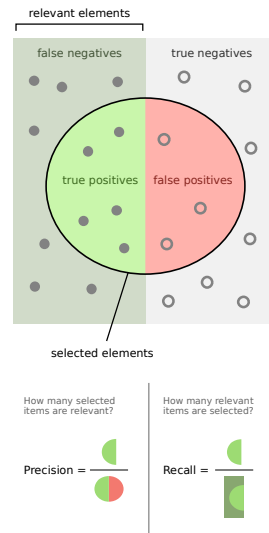
### 2.1.5. Précision et rappel

		Real Value	
		True	False
Predicted Value	True	True Positive	False Positive
	False	False Negative	True Negative

Les vrais positifs et les vrais négatifs

## 2.1. Classification

### 2.1.5. Précision et rappel



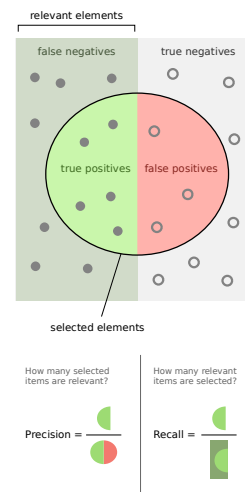
Précision et rappel

## 2.1. Classification

### 2.1.5. Précision et rappel

Soit

- $tp$ : nombre de vrais positifs
- $fp$ : nombre de faux positifs
- $fn$ : nombre de faux négatifs



### 2.1.5. Précision et rappel

Alors

- Précision

$$p = \frac{tp}{(tp + fp)}$$

- Rappel (Recall)

$$r = \frac{tp}{(tp + fn)}$$

## 2.1. Classification

### 2.1.5. Précision et rappel

- score F1 est la moyenne harmonique de la précision et du rappel :
- F1-score

$$f1 = 2 * \frac{(p * r)}{(p + r)}$$

- F1-score: meilleure valeur à 1 (précision et rappel parfaits) et pire à 0.



## 2.1. Classification

### 2.1.5. Précision et rappel

- $F_\beta$ -score utilise un facteur réel positif  $\beta$ , où  $\beta$  est choisi de telle sorte que le rappel est considéré comme  $\beta$  fois plus important que la précision, est :

- $F_\beta$ -score

$$F_\beta = (1 + \beta^2) \cdot \frac{p \cdot r}{(\beta^2 \cdot p) + r}$$

- Exemple:  $F_2$  score

## 2.1. Classification

### Matrice de confusion

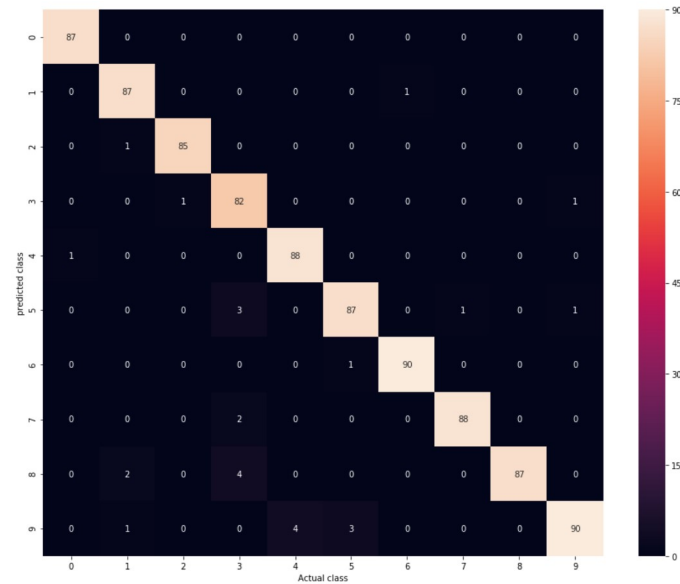
- une matrice qui mesure la qualité d'un système de classification
- chaque ligne de la matrice représente les instances d'une classe prédite
- chaque colonne représente les instances d'une classe réelle
- Toutes les prédictions correctes sont situées dans la diagonale du tableau
- Les erreurs de prédiction seront représentées par des valeurs situées en dehors de la diagonale.

		Real Value	
		True	False
Predicted Value	True	True Positive	False Positive
	False	False Negative	True Negative

Les vrais positifs et les vrais négatifs

## 2.1. Classification

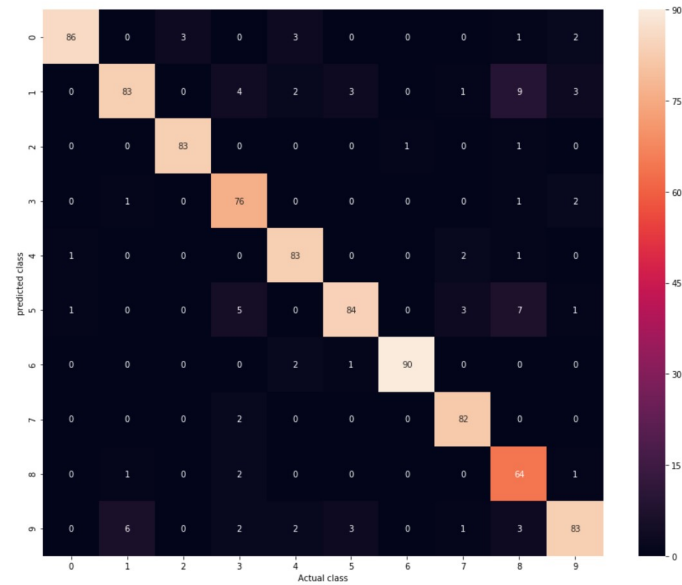
### Matrice de confusion



Matrice de confusion pour un classificateur SVM pour les chiffres manuscrits (MNIST)

## 2.1. Classification

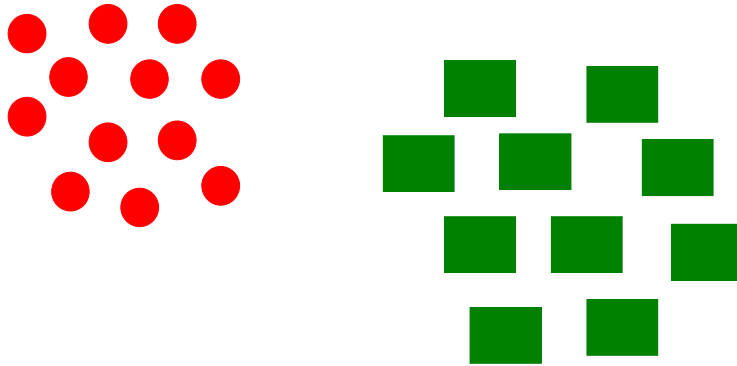
### Matrice de confusion



Matrice de confusion pour un perceptron pour les chiffres manuscrits (MNIST)

## 2.1. Classification

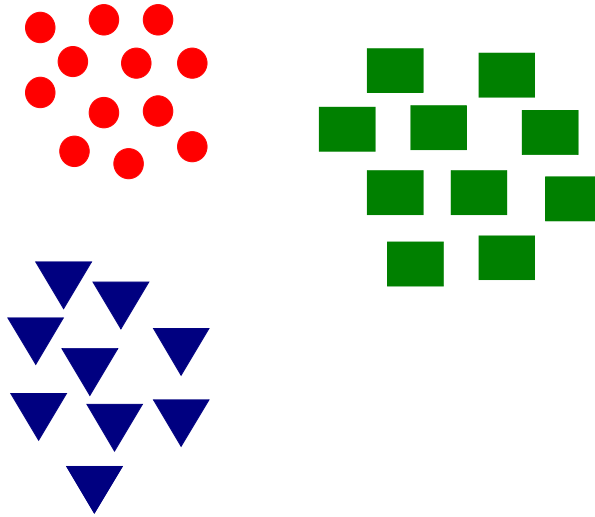
Classification binaire



Classification binaire

## 2.1. Classification

Classification multiclasse



Classification multiclasse

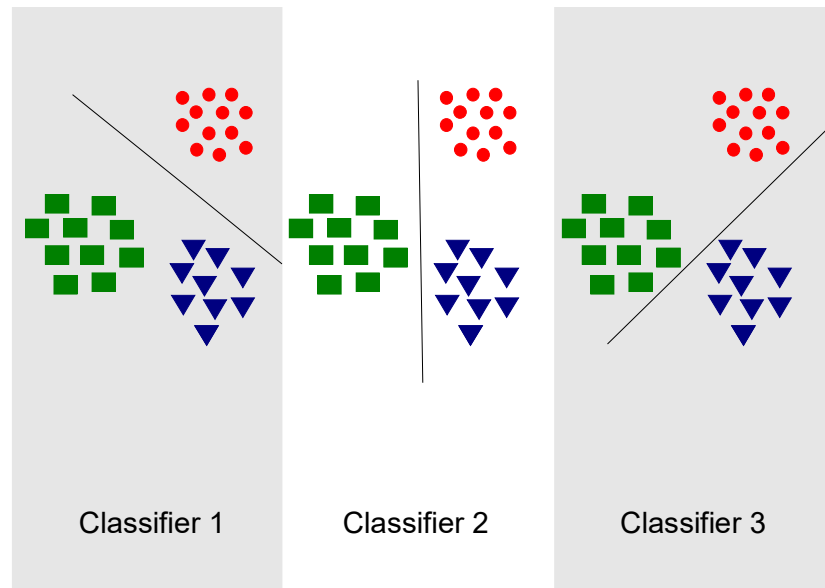
## 2.1. Classification

### Classification multiclasse [Aly 2005]

- Transformation en classification binaire
  - L'approche un contre le reste (Un contre tous)
  - L'approche un-contre-un
- Extension de la classification binaire
  - Réseaux de neurones
  - k-voisins les plus proches
- la classification hiérarchique.

## 2.1. Classification

### One-vs.-rest (One-vs.-all) strategy



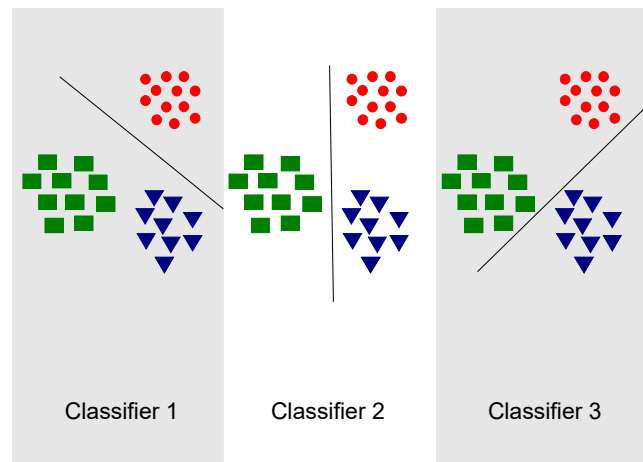
La strategie un-contre le rest pour la classification multiclasse



## 2.1. Classification

### One-vs.-rest (One-vs.-all) strategy

- Entraîner un seul classificateur par classe, avec les échantillons de cette classe comme échantillons positifs et tous les autres comme négatifs.
- Chaque classificateur produit un score de confiance réel pour sa décision



La strategie un-contre le rest pour la classification multiclasse

## 2.1. Classification

### One-vs.-rest or One-vs.-all (OvR, OvA)

- Entrées :
  - $L$ , un apprenant (algorithme d'entraînement pour les classificateurs binaires)
  - échantillons  $X$
  - étiquettes  $y$ , où  $y_i \in \{1, \dots, K\}$  est l'étiquette de l'échantillon  $X_i$
- Sortie :
  - une liste de classificateurs  $f_k$ , où  $k \in \{1, \dots, K\}$

## 2.1. Classification

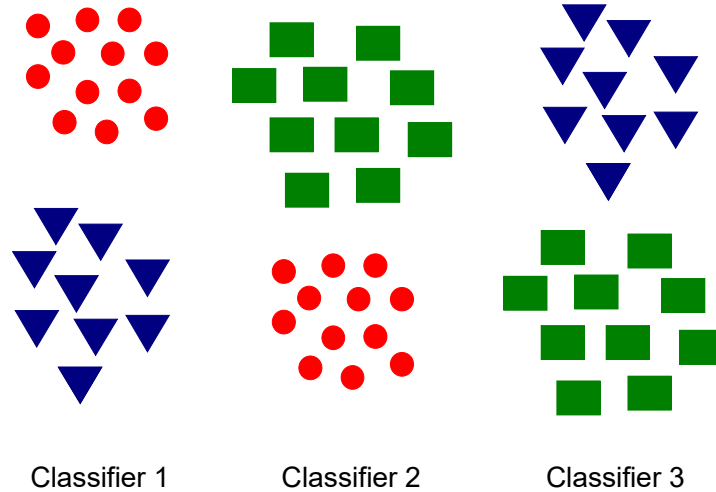
### One-vs.-rest or One-vs.-all (OvR, OvA)

Prendre des décisions signifie appliquer tous les classificateurs à un échantillon invisible  $x$  et prédire l'étiquette  $k$  pour laquelle le classificateur correspondant rapporte le score de confiance le plus élevé :

$$\hat{y} = \arg \max_{k \in \{1 \dots K\}} f_k(x)$$

## 2.1. Classification

### One-vs.-one strategy

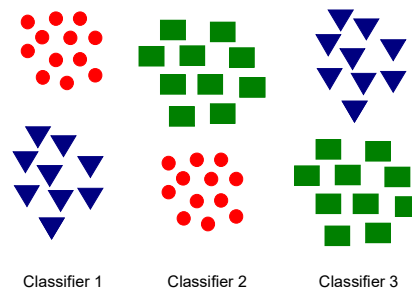


La strategie un-contre-un pour la classification multiclasse

## 2.1. Classification

### One-vs.-one strategy

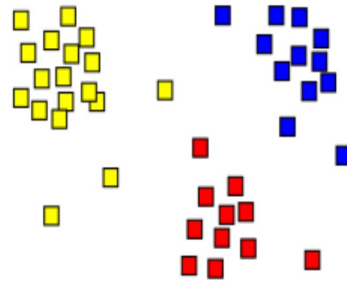
- nécessite l'entraînement des  $\frac{K(K-1)}{2}$  classificateurs binaires
- chaque classificateur reçoit les échantillons d'une paire de classes du jeu de formation original, et doit apprendre à distinguer ces deux classes.
- Au moment de la prédiction, un système de vote est appliqué : tous les  $\frac{K(K-1)}{2}$  classificateurs sont appliqués à un échantillon non vu et la classe qui a obtenu le plus grand nombre de prédictions est prédite par le classificateur combiné.



La stratégie un-contre-un pour la classification multiclasse

### 2.2.1. Introduction

- Diviser un ensemble de données en différents « paquets » homogènes,
- Les données de chaque sous-ensemble partagent des caractéristiques communes



### Applications

- Analyse des réseaux sociaux
- Segmentation d'image
- Systèmes de recommandation



### Définition formelle

- Soit  $X$  être l'espace de saisie des caractéristiques
- L'objectif du regroupement est de trouver  $k$  des sous-ensembles de  $X$ , de façon à ce que

$$C_1 \cup \dots \cup C_k \cup C_{\text{outliers}} = X$$

et

$$C_i \cap C_j = \emptyset, i \neq j; 1 \leq i, j \leq k$$

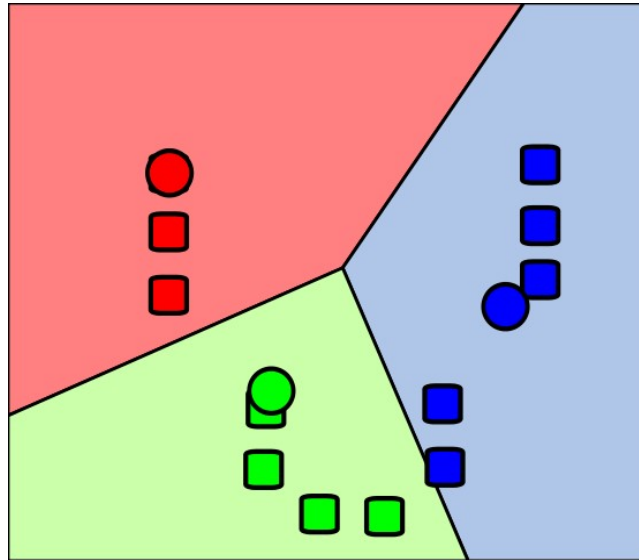
$C_{\text{outliers}}$  peut consister en des cas extrêmes (anomalie de données)



### Modèles de regroupement

- **Modèles de centroïdes** : groupe représenté par un seul vecteur moyen
- **Modèles de connectivité** : proximité de la connectivité
- Modèles de distribution : regroupements modélisés à l'aide de distributions statistiques
- Modèles de densité : regroupements de régions denses connectées dans l'espace de données
- Modèles de sous-espace
- Modèles de groupes
- Modèles graphiques
- Modèles neuronaux

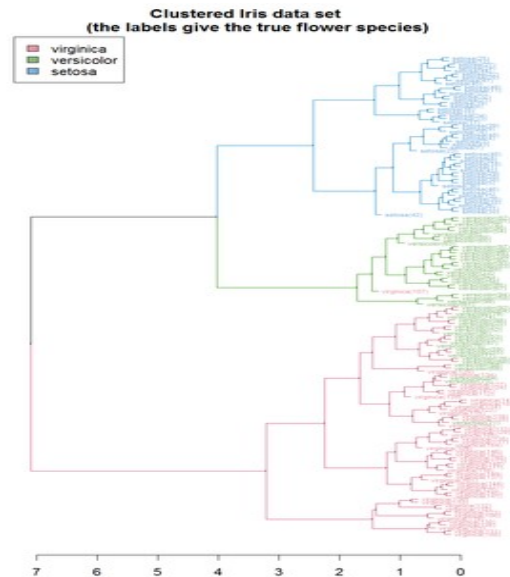
### Modèles de regroupement



k-means regroupement (voir section 3.3)

## 2.2. Partitionnement de données

### Modèles de regroupement



Dendrogramme de regroupement hiérarchique de l'ensemble de données Iris

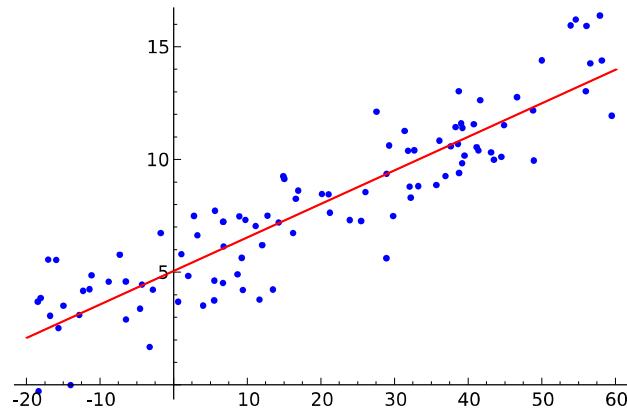
### 2.3 Régression

- Trouver une fonction qui modélise les données
- Estimer les relations entre les variables
- Analyser la relation d'une variable par rapport à une ou plusieurs autres.
- Attribuer une valeur réelle à chaque entrée

## 2.3 Régression

### Applications

- Prévisions météorologiques
- Prévisions de ventes
- Apprentissage machine
- Finance



### Définition formelle

- Une fonction qui associe un élément de données à une variable de prédiction
- Soit  $X$  les variables indépendantes
- Soit  $Y$  les variables dépendantes
- Soit  $\beta$  les paramètres inconnus (scalaires ou vectoriels)
- Le but du modèle de régression est d'approximer  $Y$  avec  $X, \beta$ , c'est à dire,

$$Y \cong f(X, \beta)$$

### Régression linéaire

- ligne droite:  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  OR
- parabole:  $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

### Régression linéaire

- ligne droite:  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  OR
- $\hat{y}_i = \beta_0 + \beta_1 x_i$  OR
- Résiduels:  $e_i = \hat{y}_i - y_i$
- Somme des carrés des résidus,  $SSE = \sum e_i^2$ , where  $1 \leq i \leq n$
- L'objectif est de minimiser l'SSE



## 2.4. Étiquetage des séquences

- Attribuer une classe à chaque membre d'une séquence de valeurs

spaCy: Reconnaissance d'entités nommées

Paris **GPE** is the capital of France **GPE** . In 2015 **DATE** , its population was recorded  
as 2,206,488 **CARDINAL**

## 2.4. Étiquetage des séquences

### Reconnaissance d'entités nommées (spaCy)

Paris **GPE** is the capital of France **GPE** . In 2015 **DATE** , its population was recorded as 2,206,488 **CARDINAL**

Balise	Signification
GPE	Pays, villes, états.
DATE	Dates ou périodes absolues ou relatives
CARDINAL	Les chiffres qui ne correspondent à aucun autre type.

### Applications

- Étiquetage de la partie du discours
- Traduction linguistique
- Analyse vidéo
- Reconnaissance de l'écriture manuscrite
- Extraction d'informations

### Définition formelle

- Soit  $X$  l'espace de saisie des caractéristiques
- Soit  $Y$  l'espace des caractéristiques de sortie (des étiquettes)
- Soit  $\langle x_1, \dots, x_T \rangle$  une séquence de longueur  $T$ .
- L'objectif de l'étiquetage des séquences est de générer une séquence correspondante
  - $\langle y_1, \dots, y_T \rangle$  des étiquettes
  - $x_i \in X$
  - $y_j \in Y$

### Association Rules

- Recherche de relations entre les variables

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

### Applications

- Exploitation de l'utilisation du web
- Détection d'intrusion
- Analyse d'affinité

### Définition formelle

- Soit  $I$  un ensemble de  $n$  attributs binaires appelés items
- Soit  $T$  un ensemble de  $m$  transactions appelé base de données
- Soit  $I = \{(i_1, \dots, i_n)\}$  et  $T = (t_1, \dots, t_m)$
- L'objectif de l'apprentissage des règles d'association est de trouver
  - $X \Rightarrow Y$ , where  $X \Rightarrow Y \subseteq I$
  - $X$  est l'antécédent
  - $Y$  est la conséquence

### Définition formelle

- Support: how frequently an itemset appears in the database

- $$\text{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

- Confidence: how frequently the rule has been found to be true.

- $$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$



### Définition formelle

- Lift: the ratio of the observed support to that of the expected if X and Y were independent

- $$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{(\text{supp}(X) \times \text{supp}(Y))}$$

## 2.5. Règles d'association

### Example

- $\{\text{bread, butter}\} \Rightarrow \{\text{milk}\}$

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

## 2.6. Détection d'anomalies

- Identification de données inhabituelles
- Approches
  1. Détection supervisé
  2. Détection non-supervisé
  3. Détection semi-supervisé

## 2.6. Détection d'anomalies

### Applications

- Détection d'intrusion
- Détection de fraude
- System health monitoring
- Détection d'événements dans les réseaux de capteurs
- Détection d'abus dans un système d'information


### Characteristics

- Des sursauts inattendus

### Formalisation

- Soit  $Y$  un ensemble de mesures
- Soit  $P_Y(y)$  un modèle statistique pour la distribution des  $Y$  dans des conditions "normales"..
- Soit  $T$  un seuil défini par l'utilisateur..
- Une mesure  $x$  est une valeur isolée si  $P_Y(x) < T$

## 2.7. Récapitulation

- 
- Synthèse courte d'un ensemble de données
  - Génération de rapports

## 2.7. Récapitulation

### Applications

- Extraction des mots-clès
- Récapitulation de documents
- Moteurs de recherche
- Récapitulation d'images
- Récapitulation de vidéos: découvrir des événements principaux dans une vidéo



### Formalisation: Synthèse multi-documents

- Soit  $\{D = D_1, \dots, D_k\}$  une collection de  $k$  documents
- Un document  $\{D = t_1, \dots, t_m\}$  se compose de  $m$  unités textuelles (mots, phrases, paragraphes, etc.)
- Soit  $\{D = t_1, \dots, t_n\}$  être l'ensemble complet de toutes les unités textuelles de tous les documents, où
  - $t_i \in D$ , si et seulement si  $\exists D_j$  de sorte que  $t_i \in D_j$
- $S \subseteq D$  constitue un résumé
- Deux fonctions de scoring
  - $Rel(i)$ : pertinence de l'unité textuelle  $i$  dans le résumé
  - $Red(i, j)$ : Redondance entre deux unités textuelles  $t_i, t_j$

### Formalisation: Multidocument summarization

- La note pour un résumé  $S$ 
  - $s(S)$  note pour un résumé  $S$
  - $l(i)$  est la longueur de l'unité textuelle  $i$
  - $K$  est la longueur maximale fixée du résumé

$$\begin{aligned} S &= \arg \max_{S \subseteq \mathcal{D}} s(S) \\ &= \arg \max_{S \subseteq \mathcal{D}} \sum_{t_i \in S} Rel(i) - \sum_{t_i, t_j \in S, i < j} Red(i, j) \\ &\quad \text{such that } \sum_{t_i \in S} l(i) \leq K \end{aligned}$$

## 2.7. Récapitulation

- Trouver un sous-ensemble à partir de l'ensemble du sous-ensemble
- Approches
  1. **Extraction**: Sélection d'un sous-ensemble de mots, de phrases ou d'expressions existants dans le texte original sans aucune modification
  2. **Abstraction**: construire une représentation sémantique interne et utiliser ensuite les techniques de génération du langage naturel

### Résumé extractif

- Approches
  1. **Résumé générique**: Obtenir un résumé générique
  2. **Résumé pertinent pour la recherche**

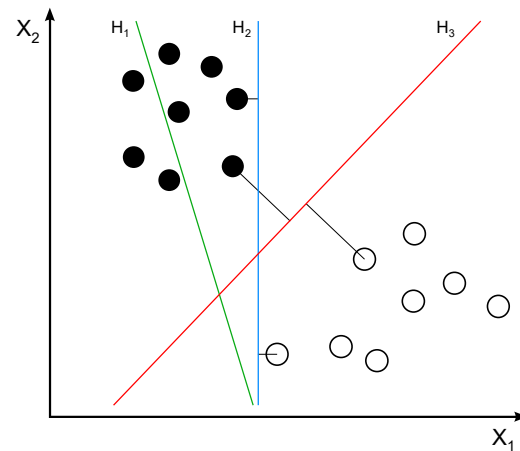
## 3. Algorithmes

1. Support Vector Machines (SVM)
2. Descente du gradient stochastique
3. Voisins proches
4. Bayes naïfs
5. Arbres de décision
6. Ensemble Methods (Forêt d'arbres décisionnels)

## 3.1. Machine à vecteurs de support (SVM)

### Introduction

- Approche d'apprentissage supervisé
- Algorithme de classification binaire
- Construit un hyperplan assurant la séparation maximale entre deux classes

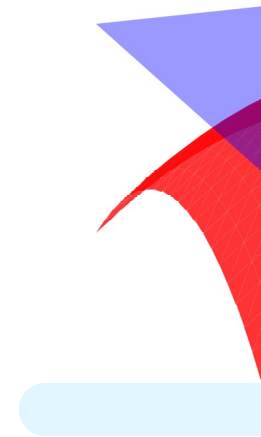


### Hyperplane

- L'hyperplan de l'espace  $n$ -dimensionnel est un sous-espace de dimension  $n-1$
- Exemples
  - L'hyperplan d'un espace à deux dimensions est une ligne à une dimension
  - L'hyperplan d'un espace tridimensionnel est un plan bidimensionnel

### Définition formelle

- Le but d'un SVM est d'estimer une fonction  $f : \mathbb{R}^N \times \{+1, -1\}$ , c'est à dire,
  - Si  $x_1, \dots, x_l \in \mathbb{R}^N$  sont les  $N$  points de données d'entrée,
  - L'objectif est de trouver  $(x_1, y_1), \dots, (x_l, y_l) \in \mathbb{R}^N \times \{+1, -1\}$
- Tout hyperplan peut être écrit par l'équation en utilisant un ensemble de points d'entrée  $x$ 
  - $w \cdot x - b = 0$ , où
  - $w \in \mathbb{R}^N$ , un vecteur normal à la plane
  - $b \in \mathbb{R}$
- Une fonction de décision est donnée par  $f(x) = \text{sign}(w \cdot x - b)$





### Définition formelle

- Si les données de formation sont séparables linéairement, deux hyperplans peuvent être sélectionnés
- Ils séparent les deux classes de données, afin que la distance entre elles soit la plus grande possible.
- Les hyperplans peuvent être donnés par les équations
  - $w \cdot x - b = 1$
  - $w \cdot x - b = -1$
- La distance entre les deux hyperplans peut être donnée par  $\frac{2}{||w||}$
- La région située entre ces deux hyperplans est appelée marge.
- L'hyperplan à marge maximale est l'hyperplan qui se trouve à mi-chemin entre eux.

### Définition formelle

- Afin d'éviter que les points de données ne tombent dans la marge, les contraintes suivantes sont ajoutées
  - $w \cdot x_i - b \geq 1$ , si  $y_i = 1$
  - $w \cdot x_i - b \leq -1$ , si  $y_i = -1$
- $y_i(w \cdot x_i - b) \geq 1, 1 \leq i \leq n$
- L'objectif est de minimiser  $\|w\|$  sous réserve de  $y_i(w \cdot x_i - b) \geq 1, 1 \leq i \leq n$
- Une solution pour les deux  $w$  et  $b$  donne le classificateur  $f(x) = \text{sign}(w \cdot x - b)$
- L'hyperplan à marge maximale est entièrement déterminé par les points qui en sont les plus proches, appelés vecteurs de soutien

### Data mining

- Classification (classification multi-classes)
- Régression
- Détection des anomalies

### Applications

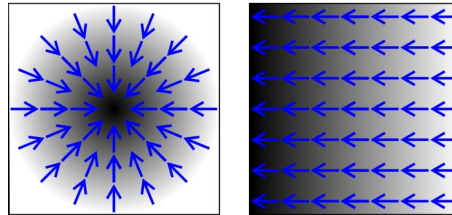
- Catégorisation des textes et des hypertextes
- Classification des images
- Reconnaissance de l'écriture manuscrite

## 3.2. Gradient stochastique de descente

- Une approximation stochastique de l'optimisation de la descente du gradient
- Méthode itérative pour minimiser une fonction objective qui s'écrit comme une somme de fonctions différentiables.
- Trouve des minima ou des maxima par itération

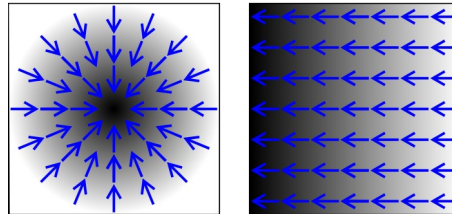
### Gradient

- Généralisation multi-variable du dérivé.
- Donne la pente de la tangente du graphe d'une fonction
- Le gradient pointe dans la direction du plus grand taux d'augmentation d'une fonction
- L'amplitude du gradient est la pente du graphique dans cette direction



### Gradient ou dérivé

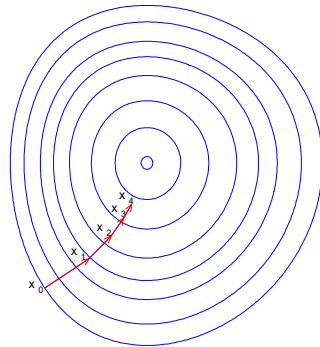
- Dérivés définis sur des fonctions d'une seule variable
- Gradient défini sur des fonctions de variables multiples
- Le gradient est une fonction à valeur vectorielle (la plage est un vecteur)
- Le dérivé est une fonction à valeur scalaire



## 3.2. Gradient stochastique de descente

### Algorithme du gradient

- Algorithme d'optimisation itératif du premier ordre pour trouver le minimum d'une fonction.
- Trouver un minimum local implique de prendre des mesures proportionnelles à le négatif du gradient de la fonction au point courant.





### Méthode standard de descente de gradient

- Prenons le problème de la minimisation d'une fonction objective
  - $Q(w) = \frac{1}{n}(\sum Q_i(w))$ ,  $1 \leq i \leq n$
  - $Q_i(w)$  est la valeur de la fonction objectif pour le  $i$ -ème exemple.
  - $Q(w)$  est le risque empirique.
- $w = w - \eta \cdot \nabla Q(w)$
- $w = w - \frac{\eta}{n} \sum_{i=1}^n \nabla Q_i(w)$ ,  $\eta$  est le pas de l'itération

### Méthode itérative

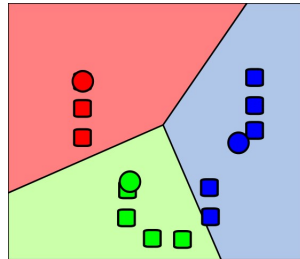
- Choisissez un vecteur initial de paramètres  $w$  et le taux d'apprentissage  $\eta$ .
- Répétez l'opération jusqu'à l'obtention d'un minimum approximatif :
  - Mélangez aléatoirement les exemples dans le jeu de formation.
  - $w = w - \eta \cdot \nabla Q_i(w)$ ,  $i = 1 \dots n$

### Applications

- Classification
- Régression

### partitionnement en k-moyennes (k-means clustering)

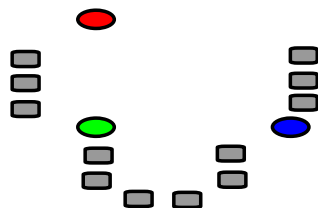
- méthode de partitionnement de données
- L'entrée est un ensemble de points et un nombre k et l'objectif est de diviser ces points en k groupes



### partitionnement en k-moyennes (k-means clustering)

#### Étape 1 (Initialisation)

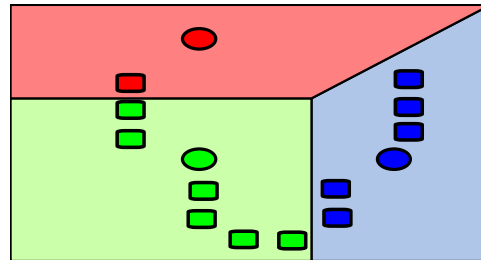
- k "moyens" initiaux (dans ce cas  $k=3$ ) sont générés de manière aléatoire



### partitionnement en k-moyennes (k-means clustering)

#### Étape 2 (Étape d'affectation)

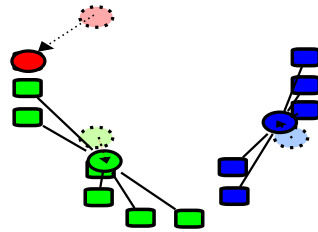
- k clusters sont créés en associant chaque observation à la moyenne la plus proche. Les partitions représentent ici le diagramme de Voronoï généré par les moyennes.



### partitionnement en k-moyennes (k-means clustering)

#### Étape 3 (Étape de mise à jour et calcul du

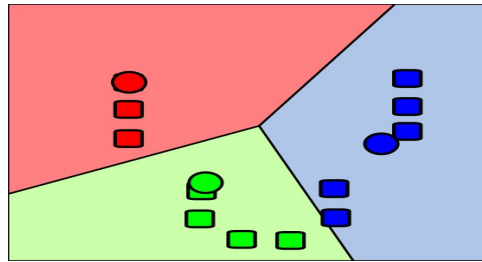
- Le centroïde de chacun des k agrégats devient la nouvelle moyenne.



### partitionnement en k-moyennes (k-means clustering)

#### Étape 4 (Répéter jusqu'à la convergence)

- Les étapes 2 et 3 sont répétées jusqu'à ce que la convergence soit atteinte.
- L'algorithme a convergé lorsque les affectations ne changent plus.

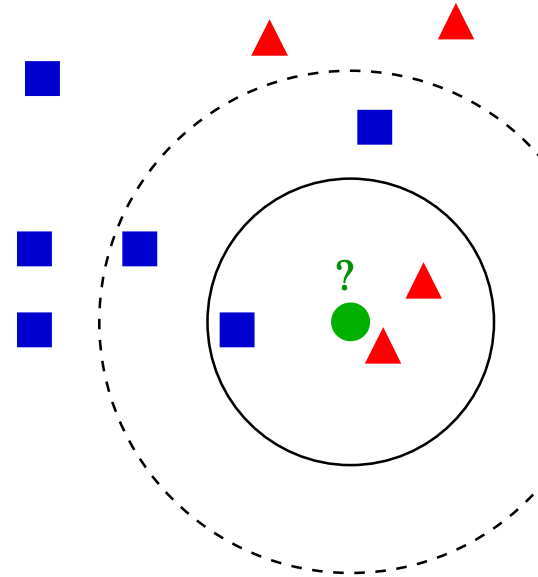




## 3.3. Méthode des plus proches voisins

### Méthode des k plus proches voisins

- Classification k-NN : la sortie est une appartenance à une classe (l'objet est classé par un vote majoritaire de ses voisins).
- Régression k-NN : la sortie est la valeur de propriété de l'objet (valeurs moyennes de ses k plus proches voisins)



### Applications

- Régression
- Détection des anomalies

## 3.4. Classification naïve bayésienne

- Collection de classificateurs probabilistes simples basés sur l'application du théorème de Bayes avec une forte hypothèse d'indépendance entre les caractéristiques.

### Applications

- Classification des documents (spam/non-spam)

### Théorème de Bayes

- Si A and B sont des événements.
- $P(A)$ ,  $P(B)$  sont des probabilités d'observer A et B indépendamment l'un de l'autre.
- $P(A|B)$  est une probabilité conditionnelle, la probabilité que l'événement A se produise étant donné que B est vrai
- $P(B|A)$  est une probabilité conditionnelle, la probabilité que l'événement B se produise étant donné que A est vrai
- $P(B) \neq 0$

$$P(A|B) = \frac{(P(B|A) \cdot P(A))}{P(B)}$$

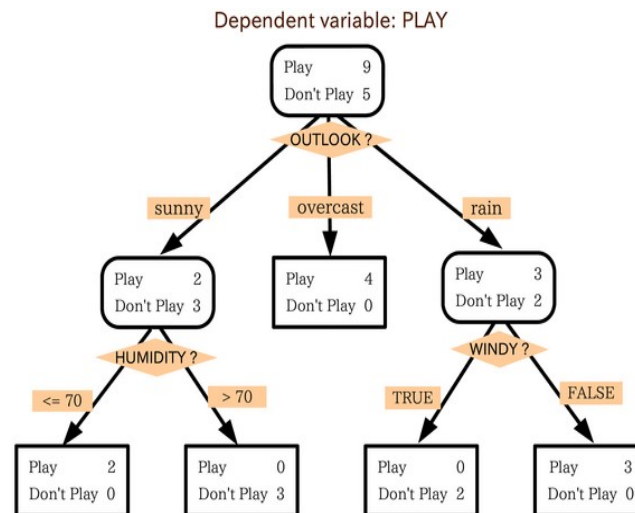
### Théorème de Bayes: Classification d'un message

- Si A and B sont des événements.
- $P(S)$  est la probabilité globale qu'un message donné soit un spam.
- $P(H)$  est la probabilité globale qu'un message donné ne soit pas du spam.
- $P(S|W)$  est la probabilité qu'un message soit un spam, sachant que le mot s'y trouve ;
- $P(W|S)$  est la probabilité que le mot apparaisse dans les messages de spam ;
- $P(W|H)$  est la probabilité que le mot "réplique" apparaisse dans les messages ham.

$$P(S|W) = \frac{P(W|S) \cdot P(S)}{P(W|S) \cdot P(S) + P(W|H) \cdot P(H)}$$

## 3.5. Arbres de décision

- Outil d'aide à la décision
- Modèle arborescent des décisions et de leurs conséquences possibles



## 3.5. Arbres de décision

- Les données sont disponibles sous la forme

$$(x, Y) = (x_1, x_2, x_3, \dots, x_k, Y)$$

- Le vecteur  $x$  est composé des caractéristiques suivantes  $x_1, x_2, x_3, \dots$
- $Y$  est la variable dépendante qui peut dépendre de  $x$



### Applications

- Classification
- Régression
- Analyse de la décision : identifier les stratégies pour atteindre un objectif
- Recherche opérationnelle

### Définition

- Collecte de plusieurs algorithmes d'apprentissage pour obtenir de meilleures performances prédictives qu'un seul des algorithmes constitutifs
- Les forêts aléatoires sont obtenues en construisant des arbres de décision multiples au moment de la formation

## 3.6. Apprentissage ensembliste (Forêt d'arbres décisionnels)

### Algorithme

- Soit  $X = x_1, x_2, \dots, x_n$  un ensemble de données avec des réponses  $Y = y_1, y_2, \dots, y_n$
- Soit  $b = 1, 2, \dots, B$ 
  - Échantillon, avec remplacement (un élément peut apparaître plusieurs fois dans un même échantillon),  $n$  exemples de formation de  $X, Y$  ; appelez-les  $X_b, Y_b$ .
  - Former un arbre de classification ou de régression  $f_b$  sur  $X_b, Y_b$ .
- Après entraînement, les prédictions pour les échantillons non vus  $x'$  peuvent être faites en faisant la moyenne des prédictions de tous les arbres de régression individuels sur  $x'$

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

ou par un vote à la majorité dans le cas des arbres de classification.

### Applications

- Classification multiclasse
- Classification multilabel (problème de l'attribution d'un ou plusieurs labels à chaque instance. Il n'y a pas de limite au nombre de classes auxquelles une instance peut être assignée).
- Régression
- Détection des anomalies

### Définition

- Processus de sélection d'un sous-ensemble de caractéristiques pertinentes
- Utilisé dans des domaines présentant un grand nombre de caractéristiques et relativement peu de points d'échantillonnage
- une méthode de réduction de la dimensionnalité

### Applications

- Analyse des textes écrits
- Analyse des données des puces à ADN

### Définition formelle[8]

- Soit  $X$  l'ensemble original de  $n$  caractéristiques, c'est-à-dire,  $|X| = n$
- Soit  $w_i$  le poids attribué à l'élément  $x_i \in X$
- La sélection binaire attribue des poids binaires tandis que la sélection continue attribue des poids en préservant l'ordre de sa pertinence.
- Soit  $J(X')$  soit une mesure d'évaluation, définie comme  $J : X' \subseteq X \rightarrow \mathbb{R}$
- Le problème de la sélection des caractéristiques peut être défini de trois façons
  1.  $|X'| = m < n$ . Trouver  $X' \subset X$  tel que  $J(X')$  est le maximum
  2. Choisir  $J_0$ , Trouver  $X' \subseteq X$ , tel que  $J(X') \geq J_0$
  3. Trouver un compromis entre la minimisation de  $|X'|$  et la maximisation du  $J(X')$

## Articles de recherche

1. From data mining to knowledge discovery in databases, Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, AI Magazine Volume 17 Number 3 (1996)
2. Survey of Clustering Data Mining Techniques, Pavel Berkhin
3. Mining association rules between sets of items in large databases, Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD 1993. p. 207.
4. Comparisons of Sequence Labeling Algorithms and Extensions, Nguyen, Nam, and Yunsong Guo. Proceedings of the 24th international conference on Machine learning. ACM, 2007.



## Articles de recherche

5. An Analysis of Active Learning Strategies for Sequence Labeling Tasks, Settles, Burr, and Mark Craven. Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2008.
6. Anomaly detection in crowded scenes, Mahadevan; Vijay et al. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010
7. A Study of Global Inference Algorithms in Multi-Document Summarization. McDonald, Ryan. European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2007.
8. Feature selection algorithms: A survey and experimental evaluation., Molina, Luis Carlos, Lluís Belanche, and Àngela Nebot. Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002.
9. Support vector machines, Hearst, Marti A., et al. IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

## Ressources en ligne

- Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on?
  - Patterns in Nature
  - Data Mining
  - Statistical classification
  - Regression analysis
  - Cluster analysis
  - Association rule learning
  - Anomaly detection
- 
- Sequence labeling
  - Automatic summarization
  - Pattern recognition
  - Scikit-learn

## Ressources en ligne

- [Support Vector Machines](#)
- [Decision tree learning](#)
- [Stochastic gradient descent](#)

## Couleurs

- [Color Tool - Material Design](#)

## Images

- [Wikimedia Commons](#)