Algorithmique des données

Introduction à l'apprentissage automatique

Charlotte Pelletier
(Basé sur le cours de N. Courty, L. Chapel, et C. Friguet,
inspiré d'un ensemble de cours de R. Flamary, A. Rakotomamonjy, C. Richard, H. Larochelle)

16 janvier 2020

Charlotte Pelletier

Enseignante-Chercheuse - Univ. Bretagne-Sud

- Recherche : IRISA - équipe Obelix
  - Classification de séries temporelles (arbres de décision et apprentissage profond)
  - Traitement d'images, détection de données aberrantes, traitement de données manquantes
  - Applications : analyse d'images satellitaires (exemple d'une carte d'occupation des sols)
- Enseignement : département SSI
  - Cours de programmation : Programmation orientée objet en Java (L2)
  - Apprentissage automatique (vous)
  - Co-responsable du Master Copernicus in Digital Earth

Contact : `charlotte.pelletier@univ-ubs.fr`
IRISA (bâtiment ENSIBS) bureau C019

Introduction

Ce cours est un cours d'**introduction** aux méthodes informatiques permettant d'exploiter des données dans le cadre de plusieurs problèmes fondamentaux :

- description et exploration des données, visualisation
- discrimination et classification.
- régression et prédiction.

Il est constitué des bases théoriques et pratiques des fondamentaux de ce que l'on appelle **l'intelligence artificielle (IA)**. Il établit un certain nombre de passerelles avec les UEs suivantes :

- Calcul Haute Performance pour le Big Data (INF2245)
- Deep learning et IA (au premier semestre du M2)
- Fouille de données (au premier semestre du M2)

Ce qui se passe sur Internet toutes les 60 secondes

Source : `https://visual.ly/community/infographic/technology/things-happen-internet-every-60-seconds.gif`

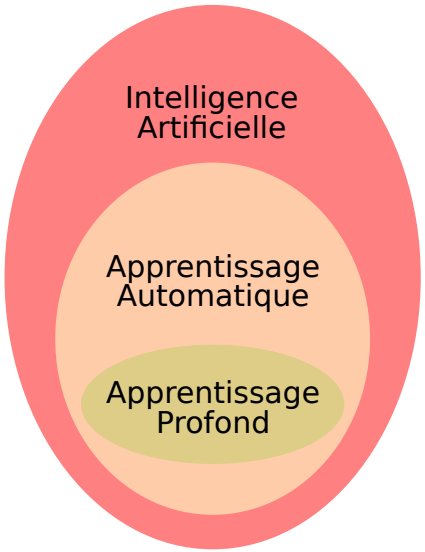
Introduction

Ce cours est motivé par la présence abondante de données rendue possible par la numérisation

- données images, textuelles, séries temporelles
- mesures et nouvelles technologies de capteurs
- réseaux sociaux, etc.

On trouve le contenu de ce cours sous plusieurs noms, dépendant de la communauté scientifique associée

- apprentissage automatique (*machine learning* ML)
  - communautés statistique et informatique
- reconnaissance de formes (*pattern recognition* PR)
  - communauté signal et image
- intelligence artificielle
  - grand public



On va surtout s'intéresser aux aspects informatiques du problème, mais il y aura aussi un peu de maths pour comprendre et expliquer ce qui se passe.

Pratique

Le cours suivra une progression classique : **10 séances de cours (2h) + 11 séances de TP (2h)**.
Divisé en trois grands thèmes

I. Introduction
- Introduction
- Rappel de maths
- Analyse en Composantes Principales

II. Classification
- apprentissage non-supervisé
- apprentissage supervisé

III. Régression

Modalités de contrôle des connaissances : note finale$=(CP+3CT)/4$

Pratique

La partie pratique de ce cours sera réalisé en **Python**, en utilisant notamment la bibliothèque `Scikit-Learn` (`http://scikit-learn.org`).

A la fin de ce cours, vous serez en mesure de :

- comprendre et distinguer les grandes catégories de problèmes se posant avec les données
- programmer et étudier des algorithmes permettant d'étiqueter ou prédire des données
- appréhender la complexité de certains problèmes ainsi que les outils mathématiques nécessaires

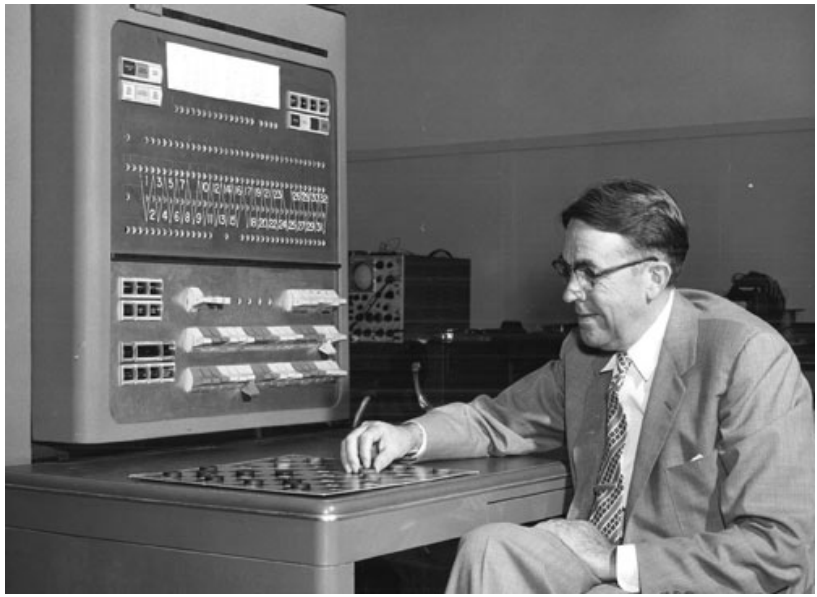
Section

Source : https://xkcd.com/1838/

Qu'est ce que l'apprentissage automatique ?

Quelques définitions provenant de la littérature

- Le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmé
  Samuel, 1959.
- Un programme informatique se dit d'apprendre de l'expérience E par rapport à une catégorie de tâches T et une mesure de la performance P, si sa performance à des tâches T, telle que mesurée par P, s'améliore avec l'expérience E
  Mitchel, 1997.



Qu'est ce que la reconnaissance de formes ?

Quelques définitions provenant de la littérature

- Le processus d'affectation d'un objet physique ou un évènement à une ou plusieurs catégories pré-spécifiées (*Duda et Hart, 1973*).
- Étant donné, plusieurs exemples de signaux complexes et d'étiquettes (ou décisions) associées, la reconnaissances de forme est le processus de prise de décision automatique pour un ensemble d'autres signaux *Ripley, 1993*.
- Le processus d'affectation d'un nom w à une observation x *Schuermann, 1993*.

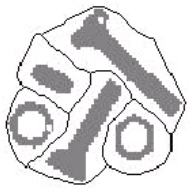
But de la reconnaissance de formes

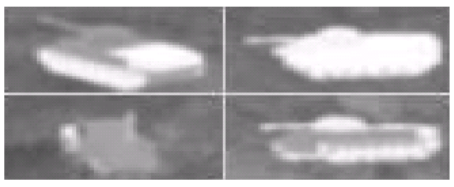
Permettre à la machine de traiter automatiquement des masses de données (signaux, images) pour résoudre un problème donné.

Exemples de problèmes

Vision

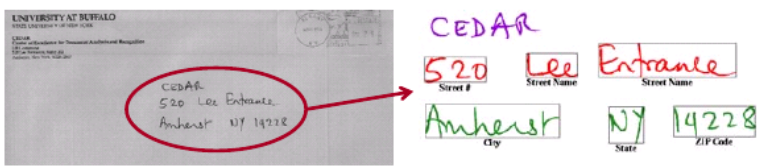
- Inspection de pièces
- Cibles militaires





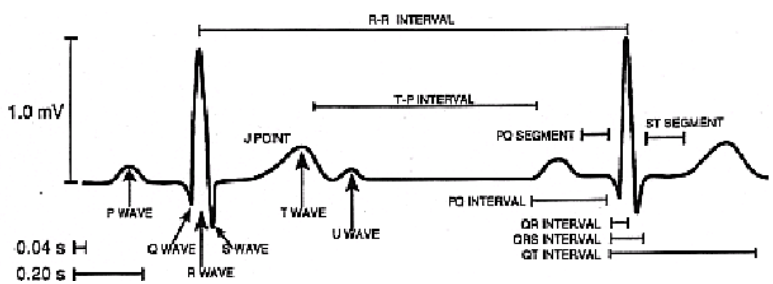
Reconnaissance de caractères

- Classement de courrier
- Traitement de chèques



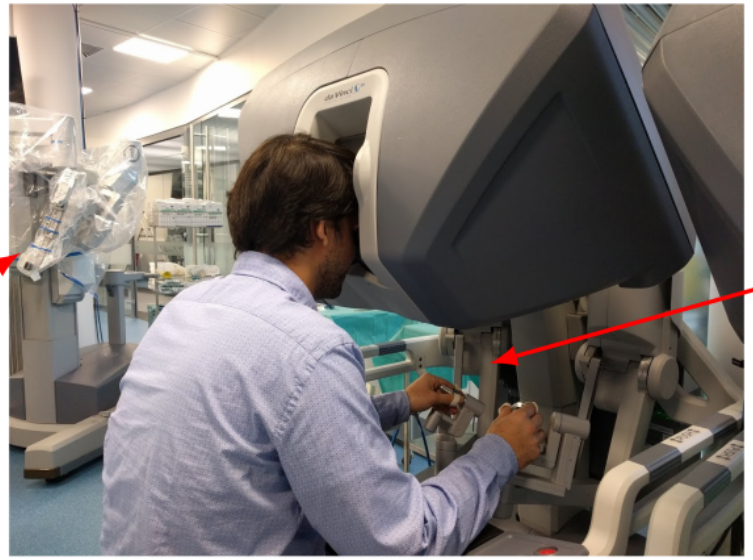
Aide au diagnostic

- Imagerie médical, EEG, ECG
- Pour assister les médecins (et non les remplacer)



Exemples spécifique “Le robot chirurgical”

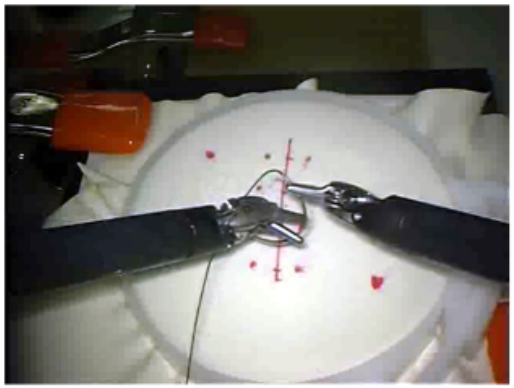
Da Vinci Surgical System



master manipulators controlled by the surgeon

slave robot end effectors controlled by the master manipulators

Photo taken during a visit to IRCAD-Strasbourg (France)



Sources : Pr. Germain Forestier, Université Haute-Alsace, IRIMAS &
Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., & Chen, C. C. G. (2014). JHU-ISI gesture and skill assessment working set (JIGSAWS) : A surgical activity dataset for human motion modeling. In MICCAI Workshop : M2CAI (Vol. 3, p. 3).

Types de problèmes

Apprentissage non-supervisé

- **Clustering** Organiser les objets en des groupes présentant une certaine similarité (taxonomie des espèces animales).
- **Estimation de densité de probabilité** Estimer la loi de probabilité des données d'apprentissage (estimer la loi d'un bruit).
- **Réduction de dimension** Diminuer la dimensionnalité des données pour pouvoir mieux les interpréter/visualiser (recommandation).

Types de problèmes

Apprentissage non-supervisé

- **Clustering** Organiser les objets en des groupes présentant une certaine similarité (taxonomie des espèces animales).
- **Estimation de densité de probabilité** Estimer la loi de probabilité des données d'apprentissage (estimer la loi d'un bruit).
- **Réduction de dimension** Diminuer la dimensionnalité des données pour pouvoir mieux les interpréter/visualiser (recommandation).

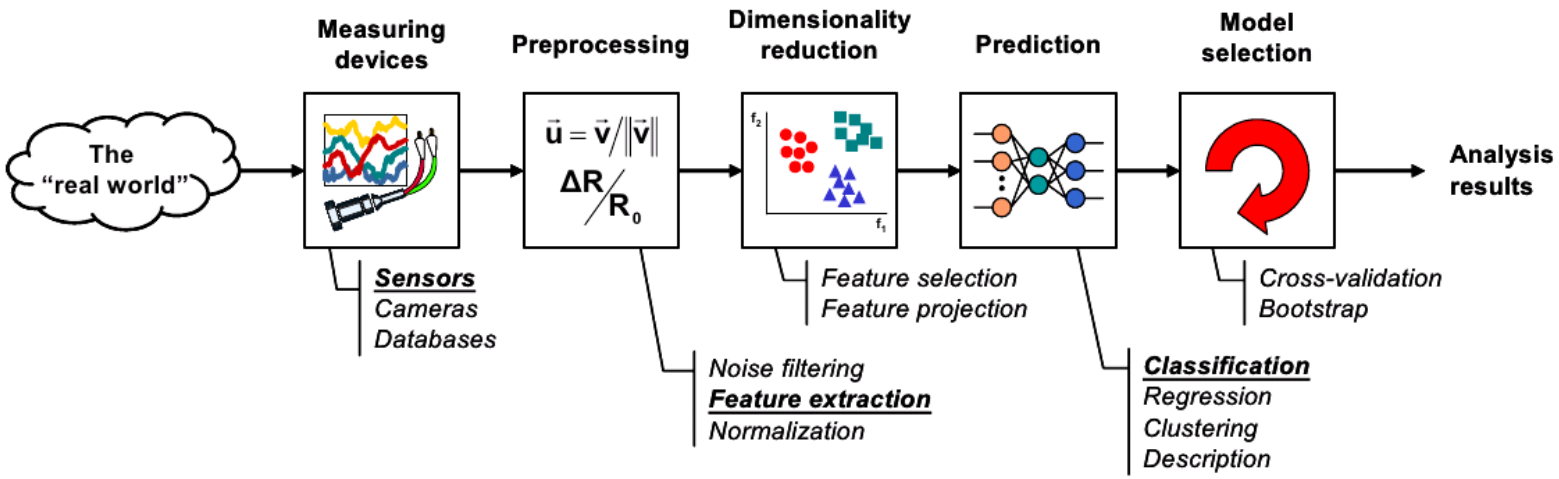
Apprentissage supervisé

- **Discrimination/Classification** Affecter une classe à une observation (reconnaissances de caractères, météo pluie).
- **Régression** Prédire une valeur réelle à partir d'une observation (météo température).

Les composantes d'un système d'apprentissage automatique

Un système classique est composé

- d'un capteur
- d'un ensemble de pré-traitements des données
- d'un système d'extraction de caractéristiques
- d'un algorithme
- d'un ensemble d'exemples (observations), les données d'apprentissage



Données d'apprentissage

Apprentissage non-supervisé

- $\mathbf{x} \in \mathbb{R}^d$ est une observation de d caractéristiques réelles.
- L'ensemble d'apprentissage définit par les observations $\{\mathbf{x}_i\}_{i=1}^n$ où n est le nombre d'exemples d'apprentissages (de points).
- Les exemples sont souvent mis sous la forme d'une matrice $\mathbf{X} \in \mathbb{R}^{n \times d}$ définie par $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ contenant les exemples d'apprentissage en lignes et les caractéristiques en colonnes.
- d et n définissent la dimensionnalité du problème d'apprentissage.

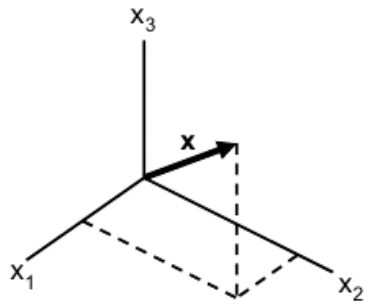
Apprentissage supervisé

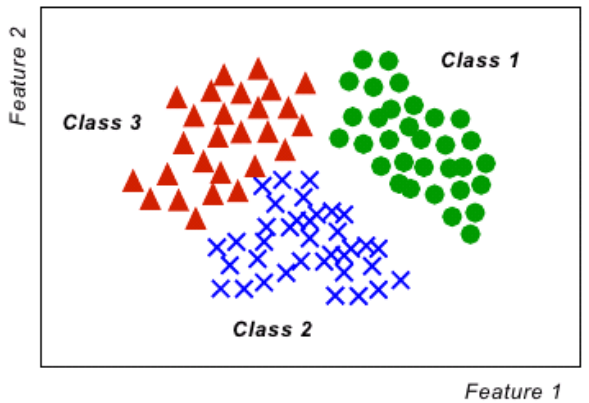
- On associe à chaque observation $\mathbf{x}_i$ une valeur à prédire $y_i \in \mathcal{Y}$.
- Tout comme pour les observation les valeurs à prédire (label) peuvent être concaténées en un vecteur $\mathbf{y} \in \mathcal{Y}^n$
- L'espace des valeurs à prédire $\mathcal{Y}$ sera :
  - $\mathcal{Y} = \{-1, 1\}$ pour la classification binaire ou $\mathcal{Y} = \{1, \ldots, m\}$ pour la classification multi-classes (m classes).
  - $\mathcal{Y} = \mathbb{R}$ pour la régression.

Caractéristiques et formes

- Une caractéristique est un trait distinctif, ou caractéristique d'un objet. Il peut être symbolique (ex : une couleur) ou numérique (ex : taille).
- **Définition**
  - Une combinaison de caractéristiques est représentée à l'aide d'un vecteur $\mathbf{x}$ de dimension d.
  - L'espace de dimension d contenant est appelé l'espace de représentation
  - Les objets sont représentés comme des points dans l'espace de représentation. On appelle cette représentation diagramme de dispersion

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$



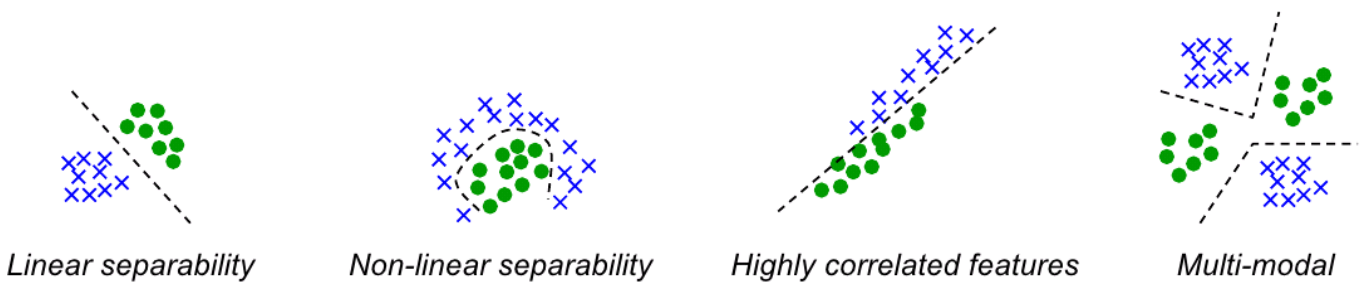


- Une forme est un ensemble de trait de caractéristiques d'une observation donnée. Dans les problèmes de discrimination, une forme est composée d'un vecteur de caractéristiques et d'un label

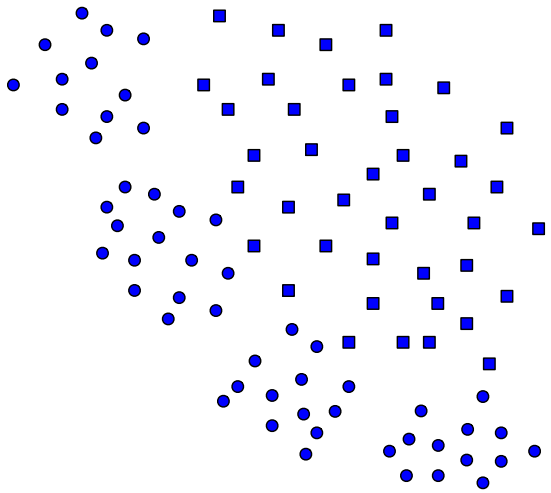
Qu'est ce qu'une "bonne" caractéristique ?
La qualité d'une caractéristique dépend du problème d'apprentissage.

- **Discrimination** Les exemples d'une même classe devraient avoir des caractéristiques similaires alors que les exemples de classes différentes devraient avoir des caractéristiques différentes.
- **Régression** La caractéristique doit aider à mieux prédire la valeur (elle doit être corrélée avec les valeurs à prédire).

Autres propriétés



Section



Soit un ensemble d'apprentissage $\{\mathbf{x}_i\}_{i=1}^n$ composé d'exemples de dimension d

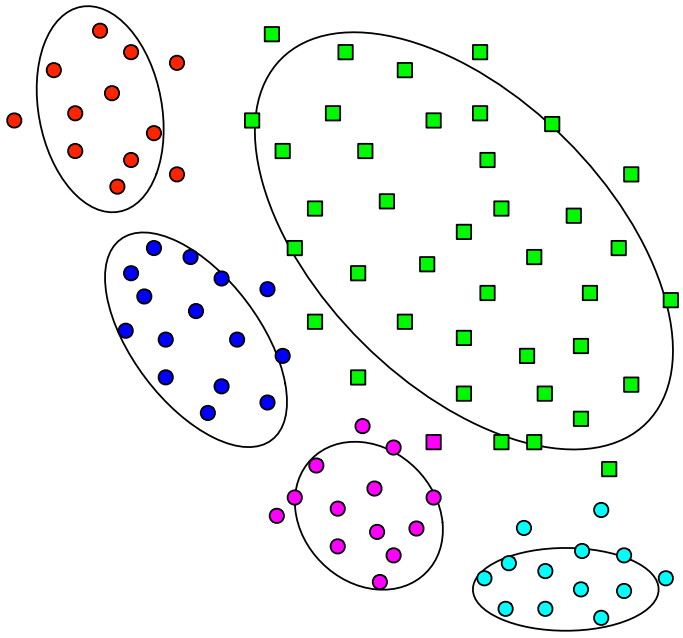
Objectifs

- **Clustering** $\{\mathbf{x}_i\}_{i=1}^n \Rightarrow \{\hat{y}_i\}_{i=1}^n$ où $\hat{y}$ est une appartenance à un groupe.
- **Estimation de densité de probabilité** $\{\mathbf{x}_i\}_{i=1}^n \Rightarrow p(\mathbf{x})$.
- **Réduction de dimension** $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n \Rightarrow \{\tilde{\mathbf{x}}_i \in \mathbb{R}^{d'}\}_{i=1}^n$ avec $d' \ll d$.

Clustering

Objectif

- Organiser les exemples d'apprentissage par groupes.
- $\{\mathbf{x}_i\}_{i=1}^n \Rightarrow \{\hat{y}_i\}_{i=1}^n$ où $\hat{y} \in \mathcal{Y}$ représente un groupe (un cluster) $\{1, \ldots, m\}$
- Paramètres :
  - m nombre de groupes
  - mesure de similarité (caractériser les similarités entre les observations)



Méthodes

- k-means (k-moyennes).
- Mélange de gaussiennes
- Clustering hiérarchique

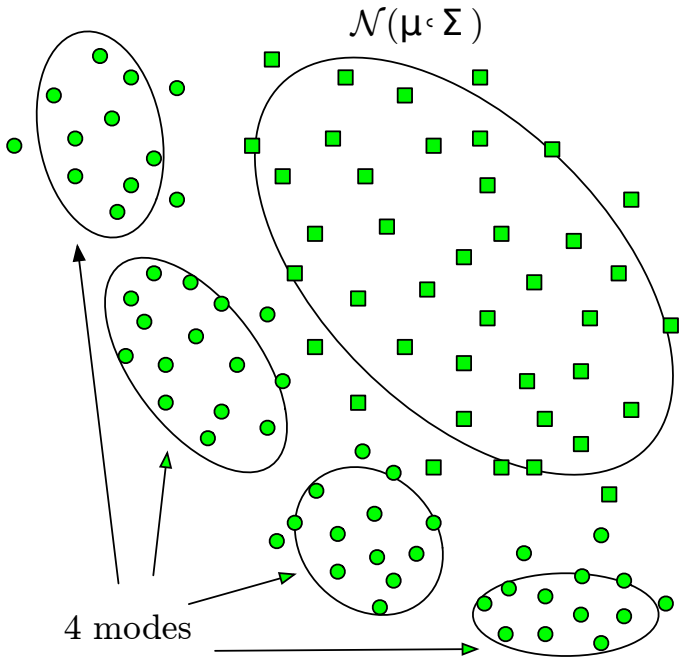
Exemples

- Taxonomie d'animaux
- Regroupement de gènes
- Réseaux sociaux

Estimation de densité de probabilité

Objectif

- Estimer la loi de proba des données
- $\{\mathbf{x}_i\}_{i=1}^n \Rightarrow p(\mathbf{x})$ où $p(\mathbf{x})$ est une densité de proba ($\int p(\mathbf{x})d\mathbf{x} = 1$)
- Modèle peut être génératif
- Paramètres :
  - Type de loi (gaussienne, ...)
  - Paramètres de la loi (μ, Σ)



Méthodes

- Fenêtres de Parzen
- Histogramme
- Mélange de gaussiennes

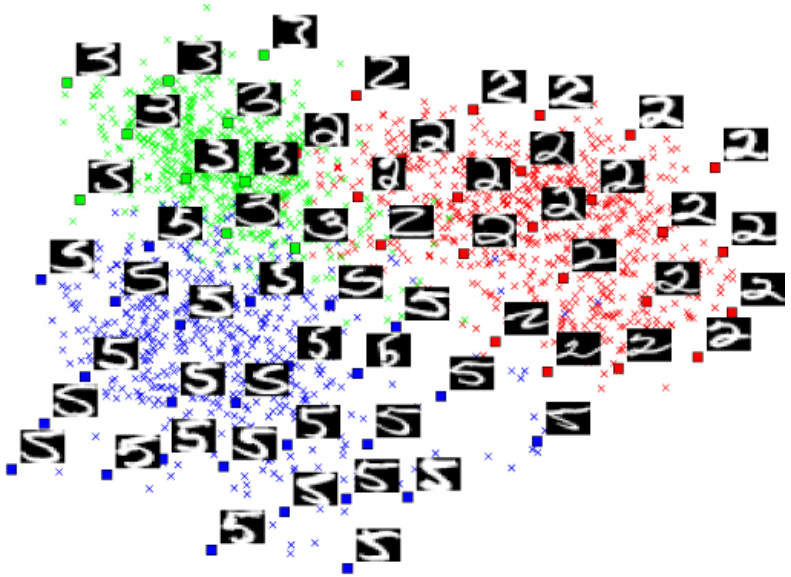
Exemples

- Estimation de bruit
- Génération de données (visage,...)
- Détection de nouveauté

Réduction de dimension

Objectif

- Projeter les données dans un espace de faible dimension.
- $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n \Rightarrow \{\tilde{\mathbf{x}}_i \in \mathbb{R}^{d'}\}_{i=1}^n$ avec $d' \ll d$ (souvent $d' = 2$).
- Paramètres :
  - Type de projection.
  - Mesure de similarité.



Méthodes

- Sélection de caractéristiques.
- Analyse en composantes principales (ACP, PCA).
- Réduction non-linéaire.

Exemples

- Pré-traitements des données
- Visualisation de vecteurs.
- Interprétation des données.
- Systèmes de recommandation.

Section

Prédiction

Soit un ensemble d'apprentissage $\{\mathbf{x}_i, y_i\}_{i=1}^n$ composé de n observations $\mathbf{x}_i \in \mathbb{R}^d$ de dimension d et de valeurs à prédire $y_i \in \mathcal{Y}$.

Objectif

- On cherche à apprendre à partir des données d'apprentissage une fonction de prédiction $f(\cdot) : \mathbb{R}^d \to \mathcal{Y}$.
- Types de prédiction :
  - **Classification**
    $f(\cdot)$ prédit une classe / une catégorie (sortie discrète) soit en classification binaire $\mathcal{Y} = \{-1, 1\}$ soit multiclasses $\mathcal{Y} = \{1, \ldots, m\}$.
  - **Régression**
    $f(\cdot)$ prédit une valeur réelle ($\mathcal{Y} = \mathbb{R}$).

Fonction linéaire

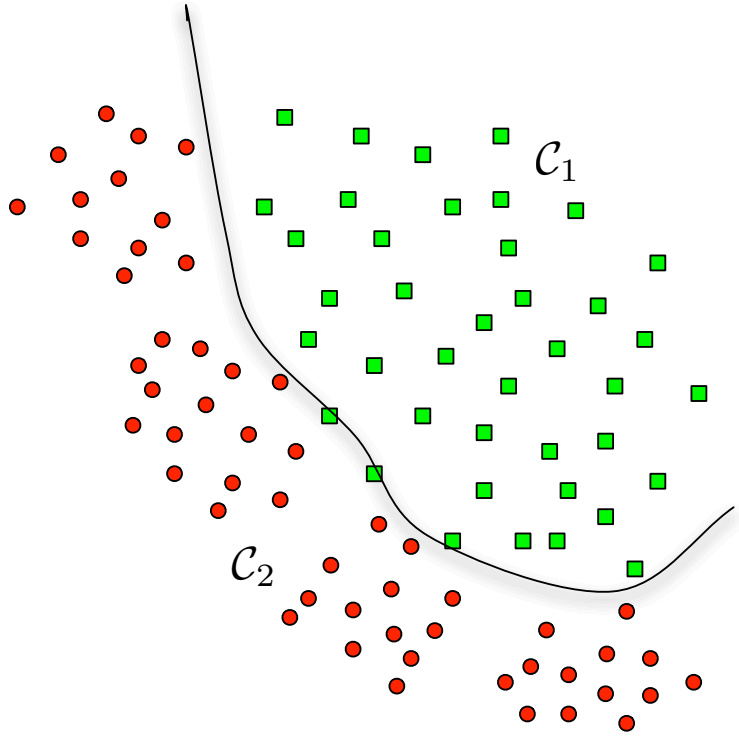
$$f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j + b = \mathbf{w}^\top \mathbf{x} + b$$

paramétrée par $\mathbf{w} \in \mathbb{R}^d$ et $b \in \mathbb{R}$

Classification binaire

Objectif

- Apprendre une fonction qui prédit la classe -1 ou 1.
- $\{\mathbf{x}_i, y_i\}_{i=1}^n \Rightarrow f(\mathbf{x})$.
- Prédiction : signe de $f(\cdot)$
- $f(\mathbf{x}) = 0$: frontière de décision.
- Paramètres :
  - Type de fonction·s
  - Mesure de performance



Méthodes

- Méthodes bayésiennes
- Séparateur linéaire discriminant
- Séparateur à Vaste Marge
- Arbre de décision

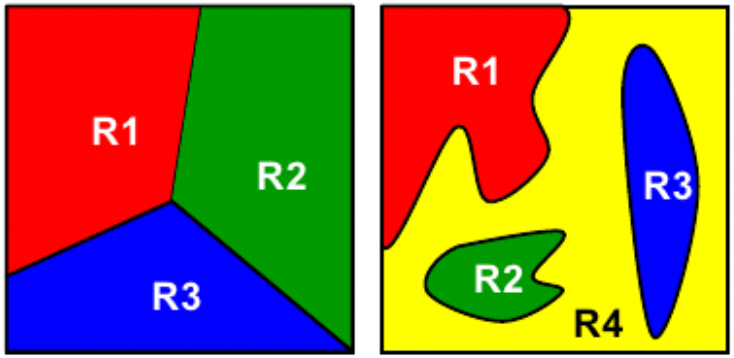
Exemples

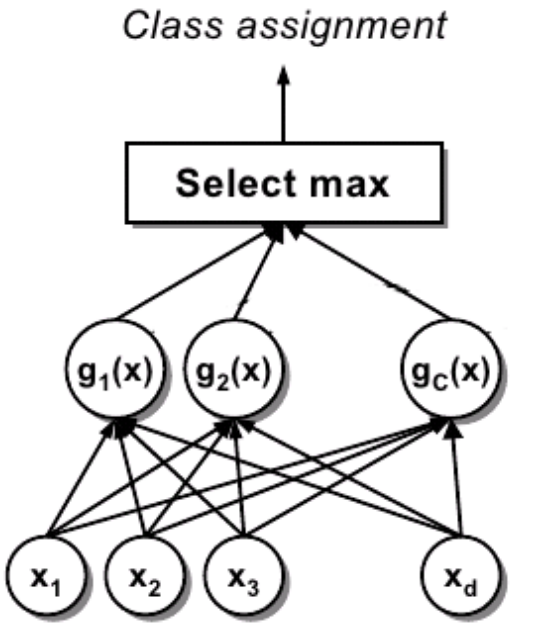
- Reconnaissance de caractères.
- Aide au diagnostique.
- Inspection de pièces.
- Météo (pluie)

Classification multiclasse

- Le rôle d'un classifieur est de partitionner l'espace des caractéristiques en plusieurs régions auxquels sont assignés des classes
  - les frontières s'appellent des frontières de décision
  - la discrimination d'un vecteur de caractéristiques $\mathbf{x}$ consiste à déterminer à quelle région il appartient et lui assigner l'étiquette (le label) de la région

- Le classifieur peut être représenter par un ensemble de fonctions discriminantes : le classifieur affecte $\mathbf{x}$ à la classe j si $g_j(x) > g_i(x)$ pour tout $i \neq j$

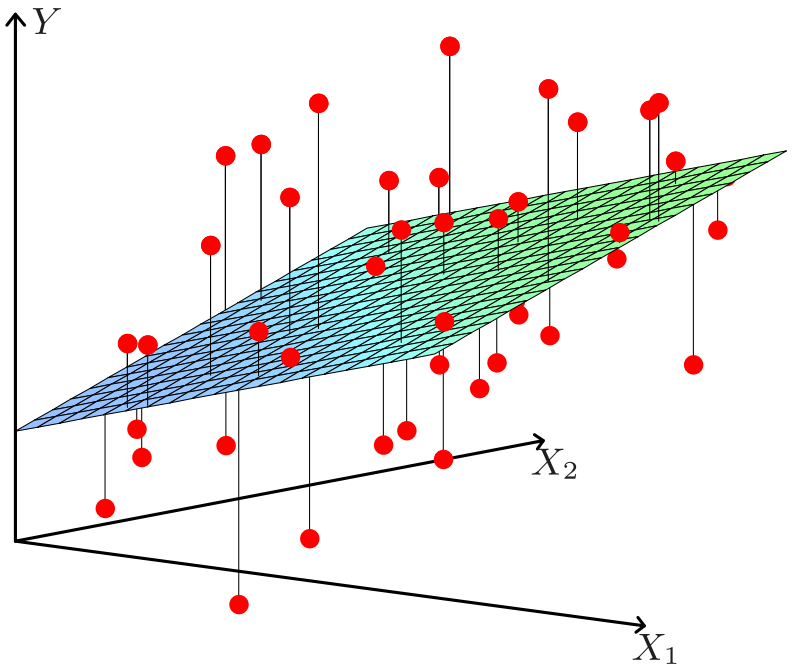




Régression

Objectif

- Apprendre une fonction qui prédit une valeur réelle.
- $\{\mathbf{x}_i, y_i\}_{i=1}^n \Rightarrow f(\mathbf{x})$.
- Paramètres :
  - Type de fonction.
  - Mesure de performance.
  - Erreur de prédiction.



Méthodes

- Moindres carrés.
- Régression ridge.
- Régression à noyaux.

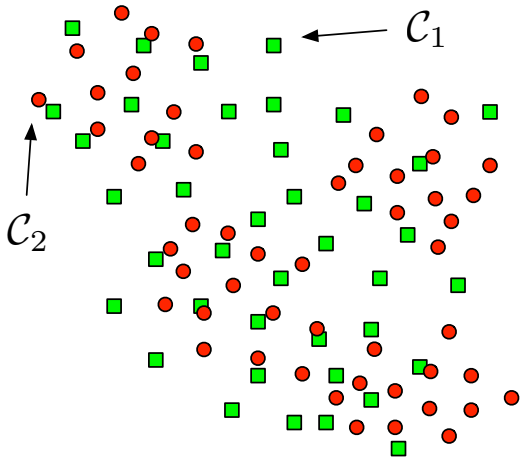
Exemples

- Prédiction mouvement.
- Prédiction taux de cholestérol.
- Météo (température).

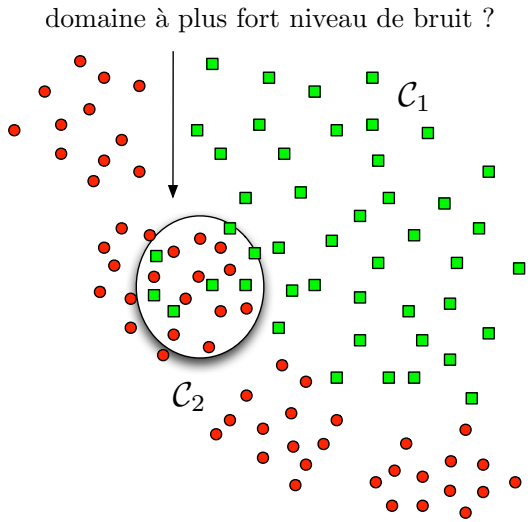
Section

Données réelles (1)

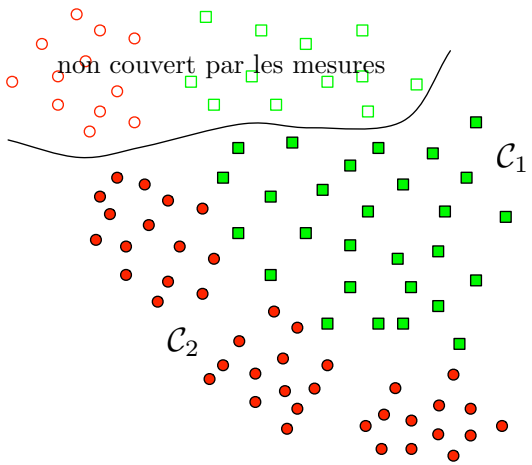
- Inadaptées



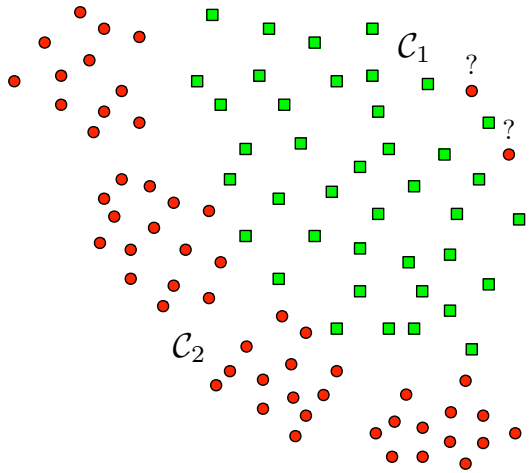
- Entachées de bruit



- Non représentative



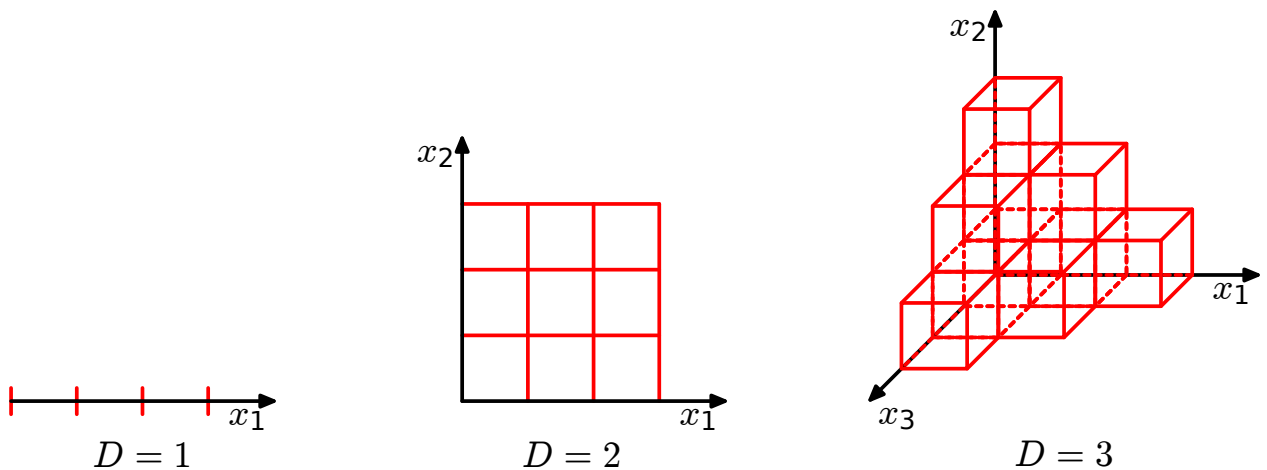
- Données aberrantes



Données réelles (2)

Taille des jeux de données
On a toujours un nombre fini n de points d'apprentissage.

Malédiction de la dimensionnalité

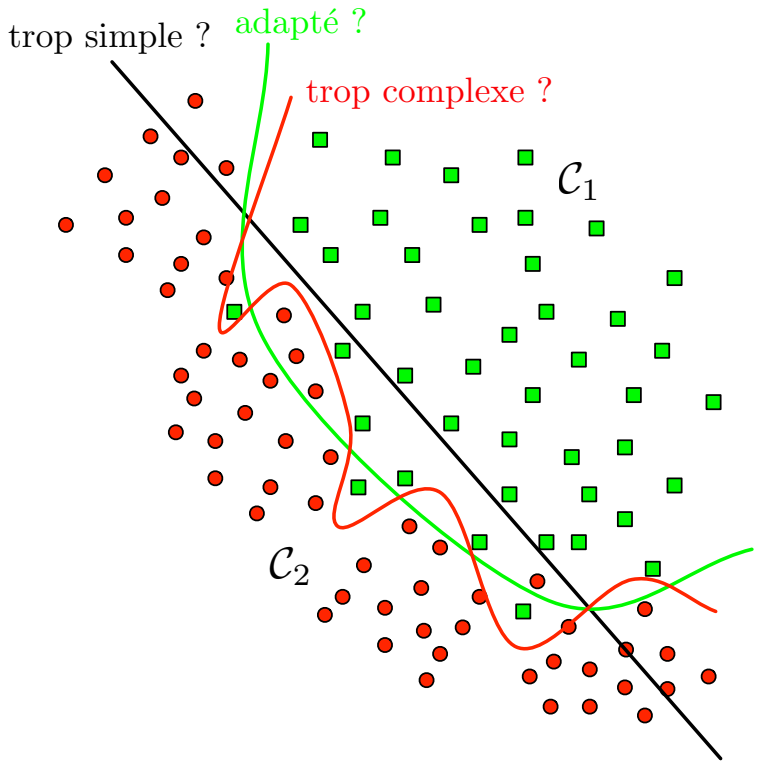


La malédiction de la dimensionnalité exprime le fait que le nombre de données doit croître exponentiellement avec la dimension pour conserver une densité équivalente.

Comment sélectionner ?

Modèle	Apprentissage	Prédiction
Trop simple	- -	- -
Adapté	+	+
Trop complexe	++	- -

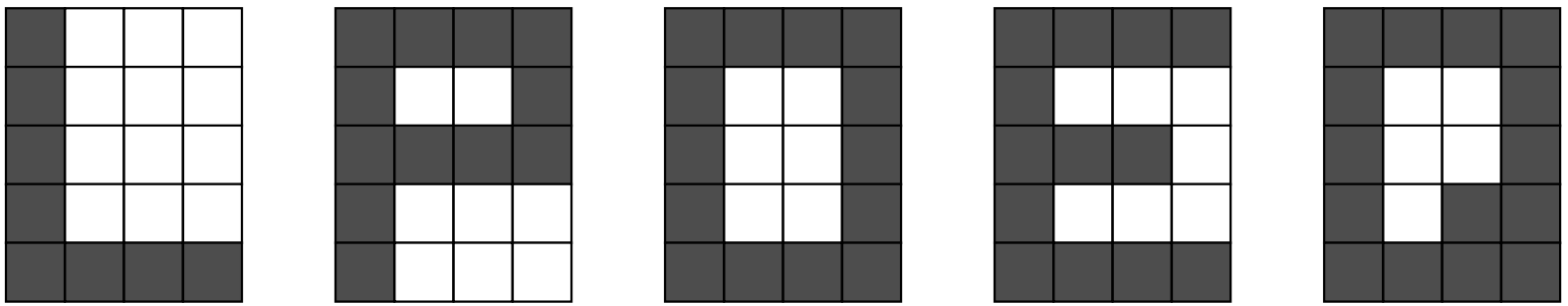
- Un modèle trop complexe provoque ce qui s'appelle du sur-apprentissage.
- On veut apprendre à prédire !



Validation

- Découpage des données en ensembles d'apprentissage/validation.
- Maximisation des performances sur les données de validation.
- La validation nécessite une bonne mesure de performances

Un exemple de tâche de reconnaissance de forme



- Développer un algorithme permettant de discriminer les lettres majuscules L,P,O,E,Q
  - Déterminer un ensemble de caractéristiques
  - Proposer un méthode de classification basé sur un arbre binaire.

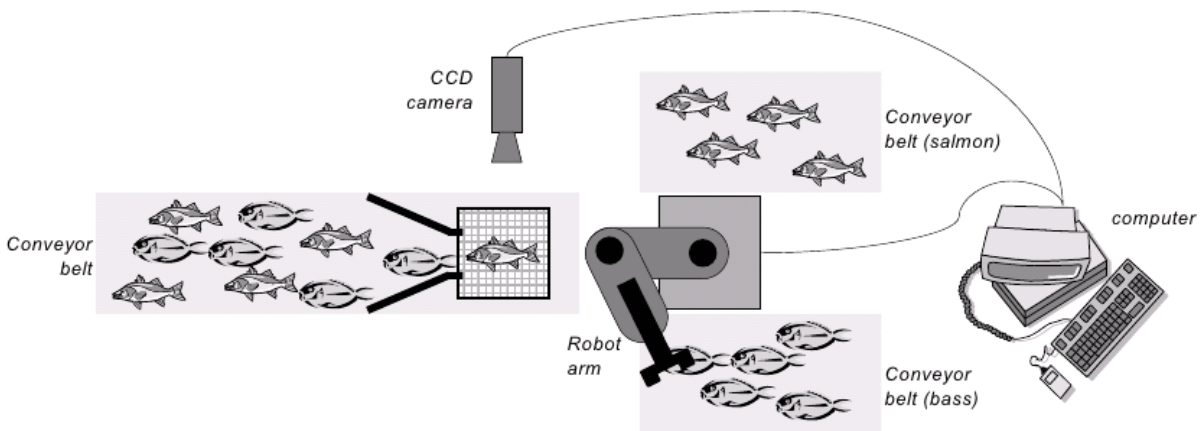
Cycle de conception

- Collecte de données
  - fastidieux et chronophage mais essentielle
  - Combien d'exemples suffisent ?
- Choix des caractéristiques
  - critique
  - peuvent être construit manuellement à partir de connaissances a priori ou automatiquement
- Choix du classifieur
  - quel modèle ?
  - comment ajuster ses paramètres ?
- Apprentissage
  - entraîner le modèle à bien "répondre" sur les données d'apprentissage
- Évaluation
  - est ce que mon modèle est bon ?
  - dilemme sur-apprentissage *vs* généralisation

Cycle de conception (2)

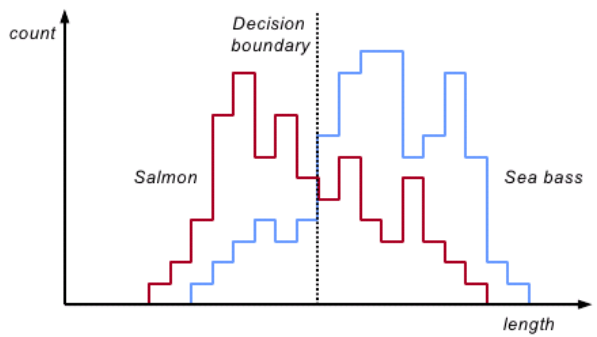
Scénario

- Une poissonnerie cherche à mettre au point un système de vision permettant de faire le tri automatique de poissons en fonctions de leurs types (saumon ou bar).
- le système est composé
  - Un tapis roulant permettant de convoyer les poissons
  - 2 tapis roulant permettant de convoyer les deux espèces de poisson
  - un bras robotisé permettant de faire le tri
  - un système de vision
  - un ordinateur permettant d'analyser les images et de contrôler le bras robotisé en fonction de la décision.



Cycle de conception (3)

- Capteur
  - le système de vision capture une image d'un poisson arrivant sur le système de tri
- Traitement d'images
  - ajustement des niveaux de gris
  - segmentation pour séparer le poisson du fond de l'image
- Extraction de caractéristiques
  - on suppose qu'en moyenne, le bar est plus long qu'un saumon
  - à partir de l'image segmentée, on estime la longueur du poisson
- Discrimination
  - Recueillir des spécimens de poissons des deux classes
  - tracer des histogrammes de longueurs pour les deux classes
  - choisir un seuil de longueur permettant de minimiser l'erreur de discrimination
  - on obtient un score décevant de 40%
  - et maintenant ?

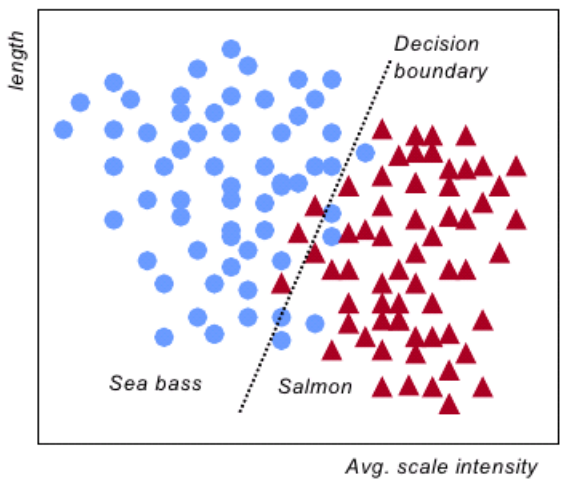


Cycle de conception (4)

Amélioration du système RdF

- visant un taux de reconnaissance de 95%, on essaye plusieurs caractéristiques
  - largeur, aire, position des yeux par rapport à la bouche ...
  - caractéristiques ne portant pas d'information discriminante
- finalement, on trouve une "bonne" caractéristique : le niveau de gris moyen des écailles.

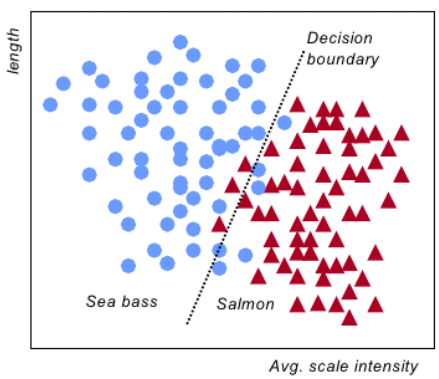
- on combine "longueur" et "niveau de gris" pour améliorer la séparabilité des classes
- on calcule une fonction de décision linéaire permettant de séparer les deux classes et obtenir un taux de reconnaissance de 95.7%

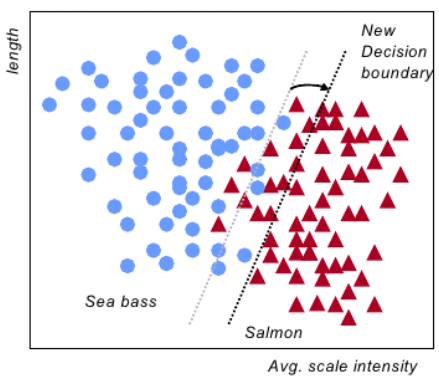


Cycle de conception d'un système de RdF (5)

Coût et taux de reconnaissance

- Notre classifieur a été construit de sorte à minimiser l'erreur de discrimination
- Est ce que c'est le meilleur choix pour notre poissonnerie ?
  - le coût de classer un saumon comme étant un bar est que le client final trouve un "bon" goût de saumon alors qu'il a acheté un bar
  - le coût de classer un bar comme étant un saumon est celui du client mécontent d'avoir acheté du bar au prix du saumon
  - les coût de mauvaise classification peuvent être différent
- Intuitivement, on aimerait prendre ce coût en compte lorsqu'on construit notre frontière de décision.

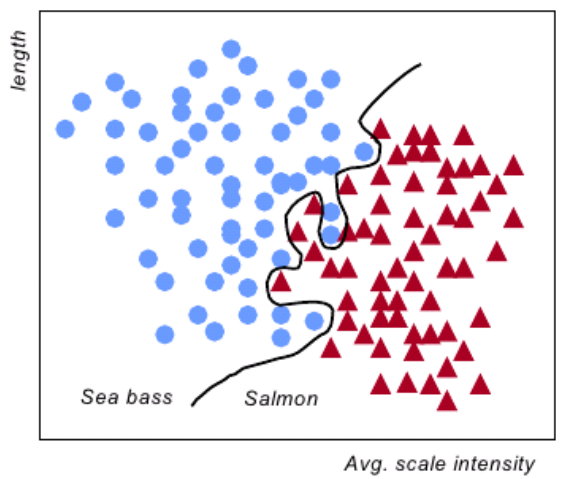




Cycle de conception d'un système de RdF (6)

Généralisation

- Notre système remplit le cahier des charges avec un pourcentage de reconnaissance des exemples de 95.7%.
- En améliorant encore le système par l'utilisation d'une méthode permettant une fonction de décision non-linéaire, on aboutit à un taux de 99.9975%, avec la fonction de décision suivante



- satisfait, nous déployons notre système dans l'usine de traitement. Mais quelques semaines, le responsable de l'usine nous rappelle pour signifier qu'en pratique, le système ne reconnaît correctement que 75% des poissons
- où s'est on trompé ?

Section

Pas seulement une question de terminologies

modèles statistiques	apprentissage automatique
points	échantillons
(co)variable	caractéristiques
paramètres	poids
estimation/fitting	apprentissage
régression/classification	apprentissage supervisé
clustering/estimation de densité	apprentissage non-supervisé
réponse	étiquette/label
performance	généralisation

Pour aller plus loin : `https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3`