Machine Learning

Introduction

Cédric RICHARD
Université Nice Sophia Antipolis

Reconnaissance des formes

Objectifs

Des observations expérimentales à l'élaboration de théories :

- Au 16^{e} siècle, J. Kepler élabore sa théorie du mouvement des planètes grâce aux observations de T. Brahé.
- Au début 20^{e} siècle, l'observation de la régularité des spectres atomiques permet le développement de la physique quantique.
- ...

L'objet de la reconnaissance des formes est l'extraction automatique d'informations d'un ensemble d'observations. Les objectifs sont : prédiction, classification, etc.

Les domaines de la reconnaissance des formes et du machine learning ont les mêmes objectifs mais relèvent de 2 communautés distinctes, l'ingénierie et l'informatique.

Reconnaissance des formes

Exemple : reconnaissance automatique de caractères manuscrits

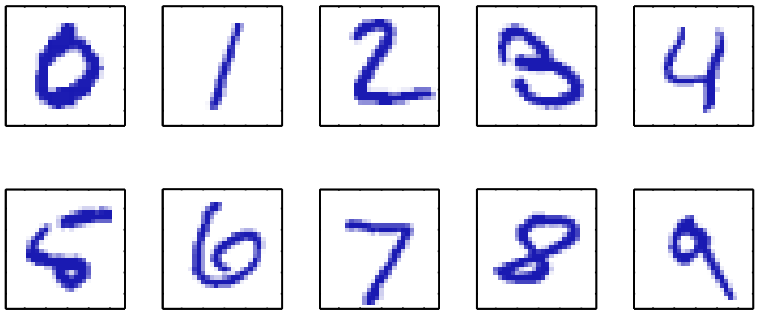


Figure 1 – Traitement des codes postaux US (28×28)

Objectif : reconnaissance automatique des codes postaux

Plusieurs stratégies sont envisageables, dont :

- Elaboration d'un système à base de règles...
- Extraction automatique d'informations pertinentes à partir d'un ensemble de données disponibles, dit d'apprentissage.

Reconnaissance des formes

Exemple : détection d'objets dans des images

Figure 2 – Détection et suivi d'objets

Objectifs : détecter des objets d'une certaine catégorie dans des images

Etapes de résolution, comportant souvent plusieurs cycles :

- Déterminer quelles données devraient permettre de discriminer la classe visée du reste du monde ;
- Récolter, analyser, débruiter/corriger les données ;
- Modéliser la frontière de discrimination à partir des données.

Reconnaissance des formes

Exemple illustratif simplifié

Figure 3 – Tri automatique des poissons

Objectifs : sur un bateau de pêche industrielle, séparer automatiquement les saumons des autres poissons péchés

- Matériel : système de vision et bras robotisé
- Hypothèse : on ne s'intéresse, ni à l'extraction des primitives par le système de vision, ni au contrôle du bras
- Problème : à partir de primitives à définir, définir la classe « saumon »

Reconnaissance des formes

Différentes finalités

On distingue les problèmes de reconnaissance des formes suivants, selon la nature du scénario rencontré :

Explorer une base de données vs. confirmer des hypothèses

- ▷ Exploration : représenter, identifier des particularités, ... ;
- ▷ Confirmer : répondre à des questions précises.

Décrire les données vs. élaborer une règle de décision

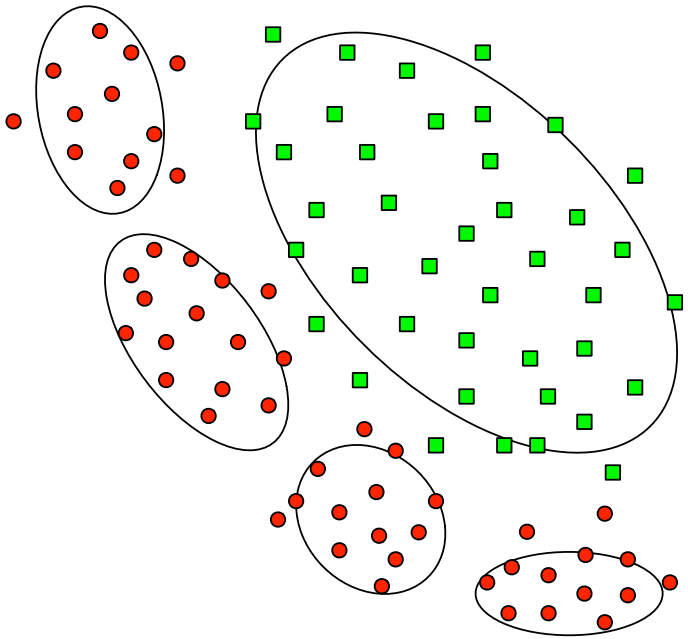
- ▷ Description : caractériser les données observées ;
- ▷ Décision : extrapoler à des données autres que celles dont on dispose.

Reconnaissance des formes

Différentes finalités

Exploration d'une base de données

▷ exemple : regroupement automatique des données en classes (clustering)

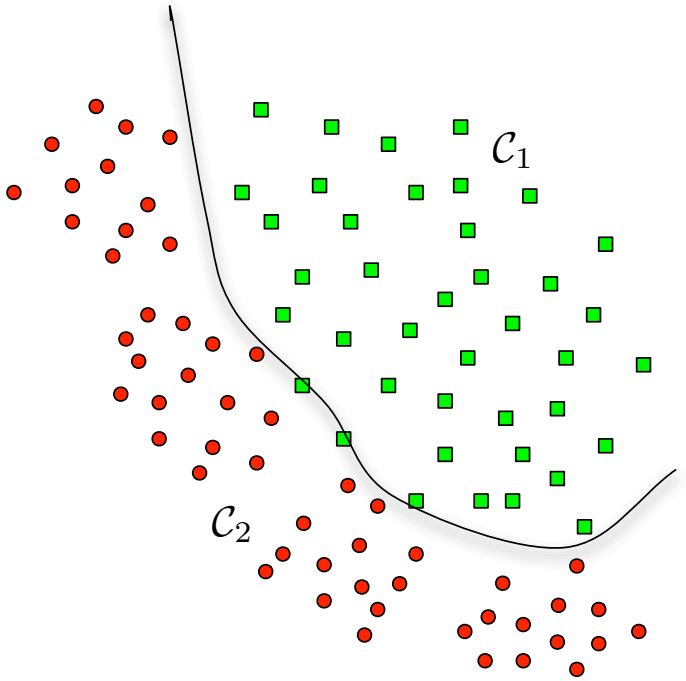


Reconnaissance des formes

Différentes finalités

Confirmation ou infirmation d'hypothèses

▷ exemple : étude de la séparabilité des classes $\mathcal{C}_1$ et $\mathcal{C}_2$

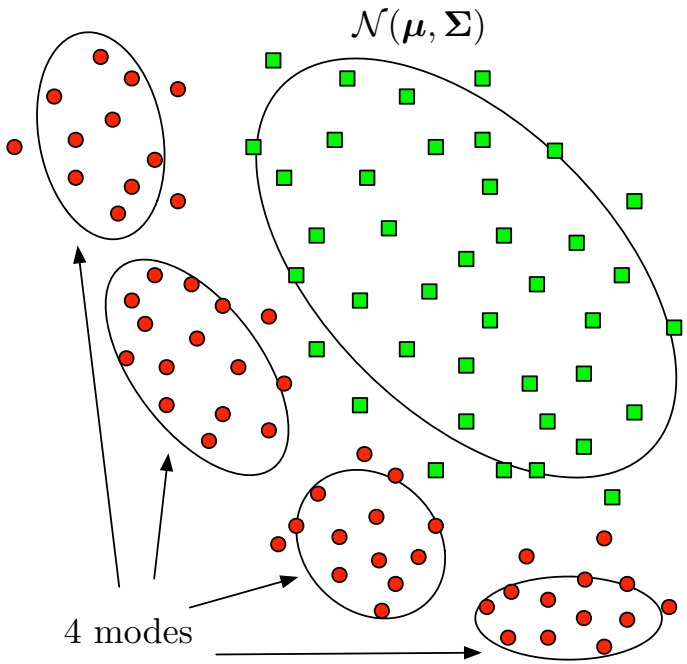


Reconnaissance des formes

Différentes finalités

Description des données

▷ exemple : modélisation probabiliste de chaque classe

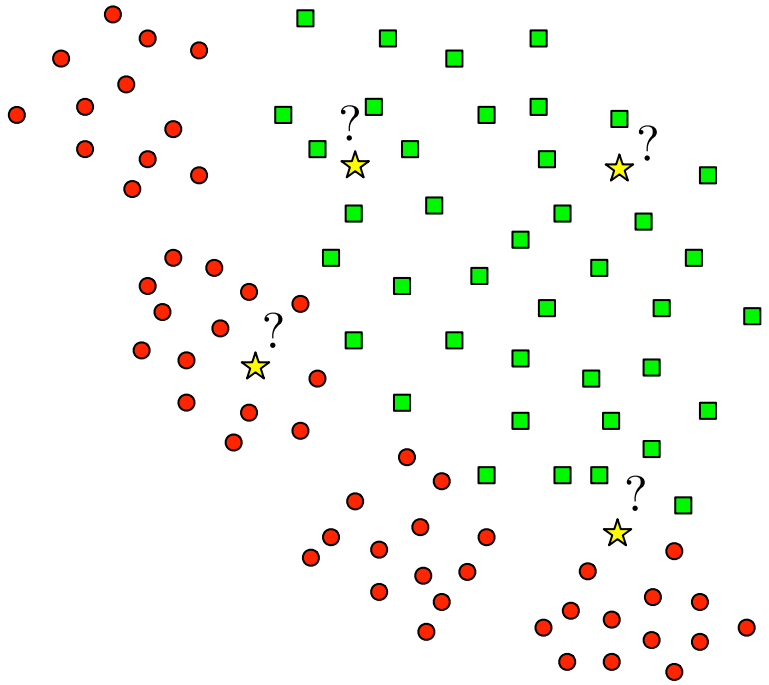


Reconnaissance des formes

Différentes finalités

Elaboration d'une règle de décision

▷ exemple : classification de nouvelles données



Reconnaissance des formes

Scénarios d'apprentissage

On distingue au moins trois scénarios possibles en matière d'apprentissage, et des stratégies différentes pour les aborder selon le problème sous-jacent :

Apprentissage non supervisé. On dispose dans ce cas d'un ensemble de données uniquement composé d'entrées $\boldsymbol{x}_n$. On distingue dans ce cas

- ▷ clustering : regrouper automatiquement les $\boldsymbol{x}_n$ en classes ;
- ▷ estimation de densité : déterminer la distribution des données ;
- ▷ visualisation : représenter les données dans un espace de faible dimension.

Apprentissage supervisé. On dispose d'un ensemble d'apprentissage composé de couples entrée-sortie $(\boldsymbol{x}_n, t_n)$. On distingue dans ce cas les problèmes de

- ▷ classification : les t_n désignent un nombre fini de classes ;
- ▷ régression : les t_n sont des réalisations d'une variable continue.

Apprentissage par renforcement. L'algorithme interagit avec son environnement par des séquences d'actions autorisées, à partir de desquelles il doit définir la meilleure stratégie.

Reconnaissance des formes

Nature des données

Données. Valeurs que prennent un ensemble de *variables* (attributs, traits, ...) pour un ensemble d'*observations* (objets, individus, enregistrements, ...)

Typologie

- ▷ quantitatives (continues, discrètes) vs. qualitatives (ordinales, nominales)
- ▷ séquentielles (ex. indice boursier) vs. non séquentielles
- ▷ spatiales (ex. fertilité d'un sol) vs. non spatiales
- ▷ structurées (ex. phrases) vs. non structurées

Remarque. Les variables retenues doivent être caractéristiques des données étudiées et du problème traité.

Reconnaissance des formes

Réalité des données

La qualité des données recueillies conditionne l'efficacité des traitements.
Cependant, le plus souvent, on déplore...

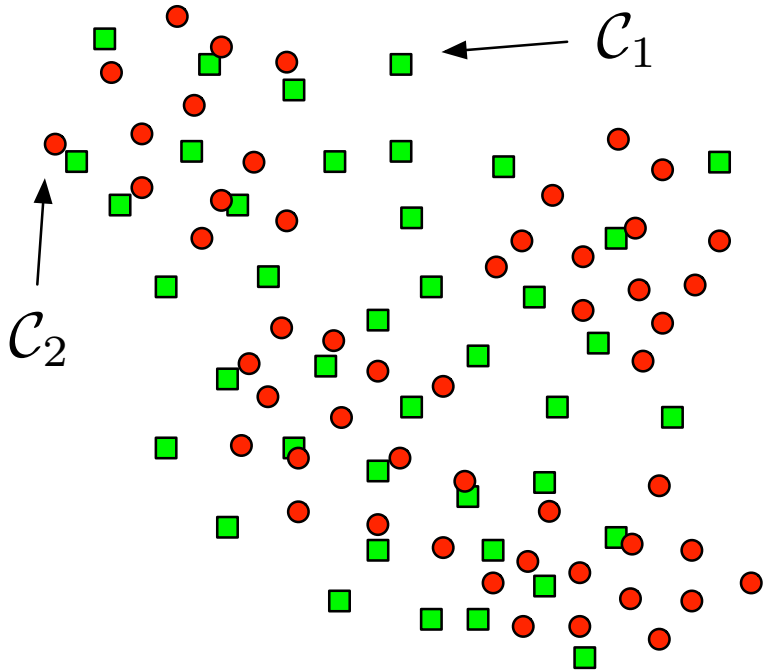
- ▷ données inadaptées
- ▷ données non représentatives
- ▷ données entachées de bruit
- ▷ données mal enregistrées
- ▷ données aberrantes
- ▷ données manquantes
- ▷ données en nombre insuffisant (malédiction de la dimensionnalité)

Conseil. Il est indispensable de se familiariser avec le phénomène étudié et les données recueillies par des études exploratoires, la représentation des données, etc.

Reconnaissance des formes

Données inadaptées

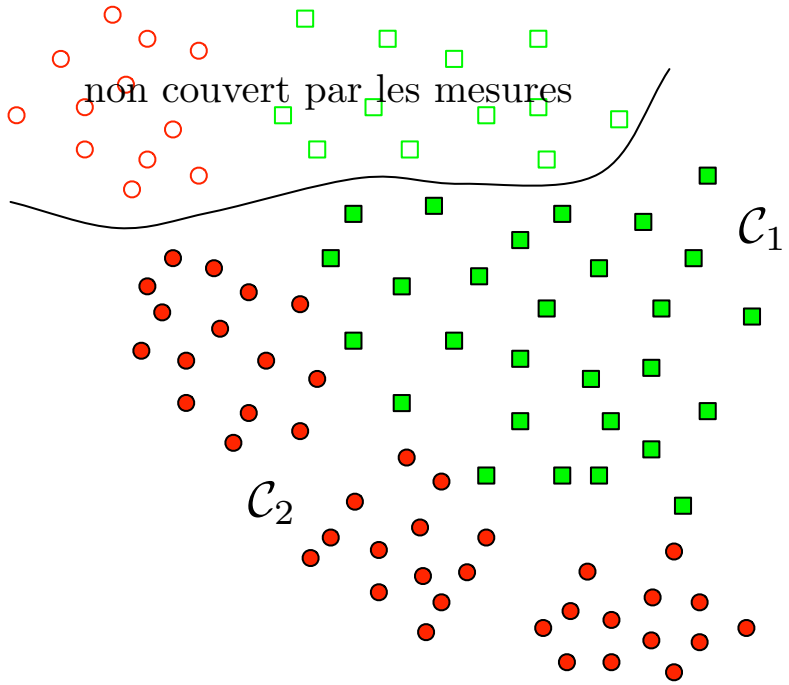
Les variables utiles n'ont pas été recueillies. On ne peut pas répondre à la question avec les mesures effectuées. Il est nécessaire de recommencer les acquisitions...



Reconnaissance des formes

Données non représentatives

Une base de données non exhaustive correspond à une couverture partielle du support des distributions (d'une statistique suffisante), par exemple due à un problème d'échantillonnage. Ceci peut être critique pour les applications sensibles.



Reconnaissance des formes

Données manquantes

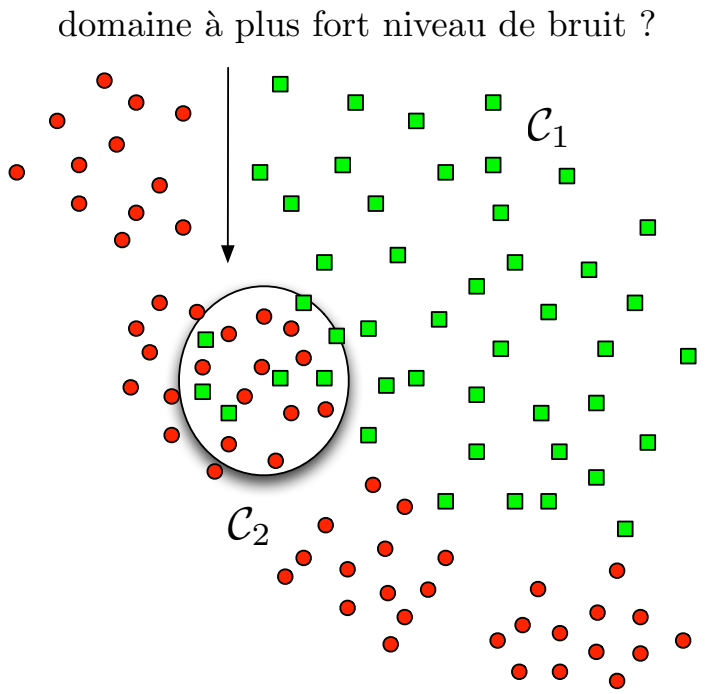
— Pour quelques observations, les valeurs de certaines variables mesurées peuvent manquer
— Les observations à valeurs manquantes peuvent être éliminées mais, parfois, il peut être opportun de combler les manques par des estimations
— Le fait qu'une donnée soit manquante peut être étroitement lié à la valeur qu'elle aurait due prendre

Conseil. Il est indispensable de comprendre pourquoi des données manquent

Reconnaissance des formes

Données entachées de bruit

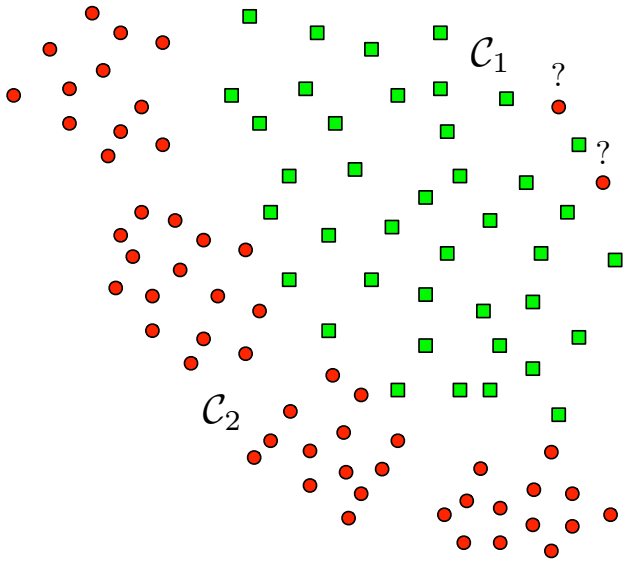
Le niveau de bruit qui affecte des données peut dépendre des valeurs prises, ce qui ajoute une difficulté supplémentaire. Il convient d'identifier les sources de bruit et leurs spécificités.



Reconnaissance des formes

Données aberrantes

Les données aberrantes peuvent être dues à des erreurs d'enregistrement, à du bruit, à des phénomènes significatifs...

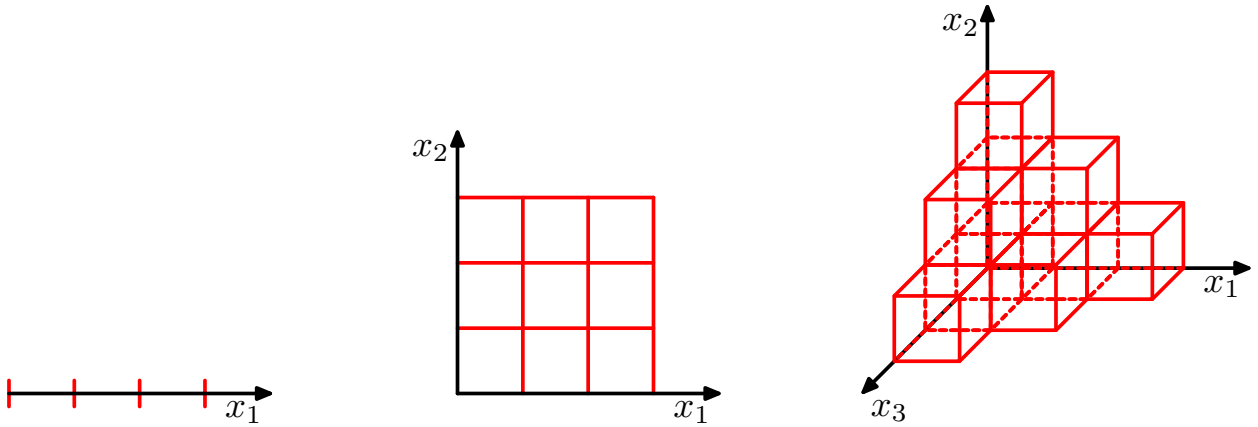


Stratégies. Il est possible d'ignorer ces données par des techniques robustes, ou chercher à les détecter et les expliquer.

Reconnaissance des formes

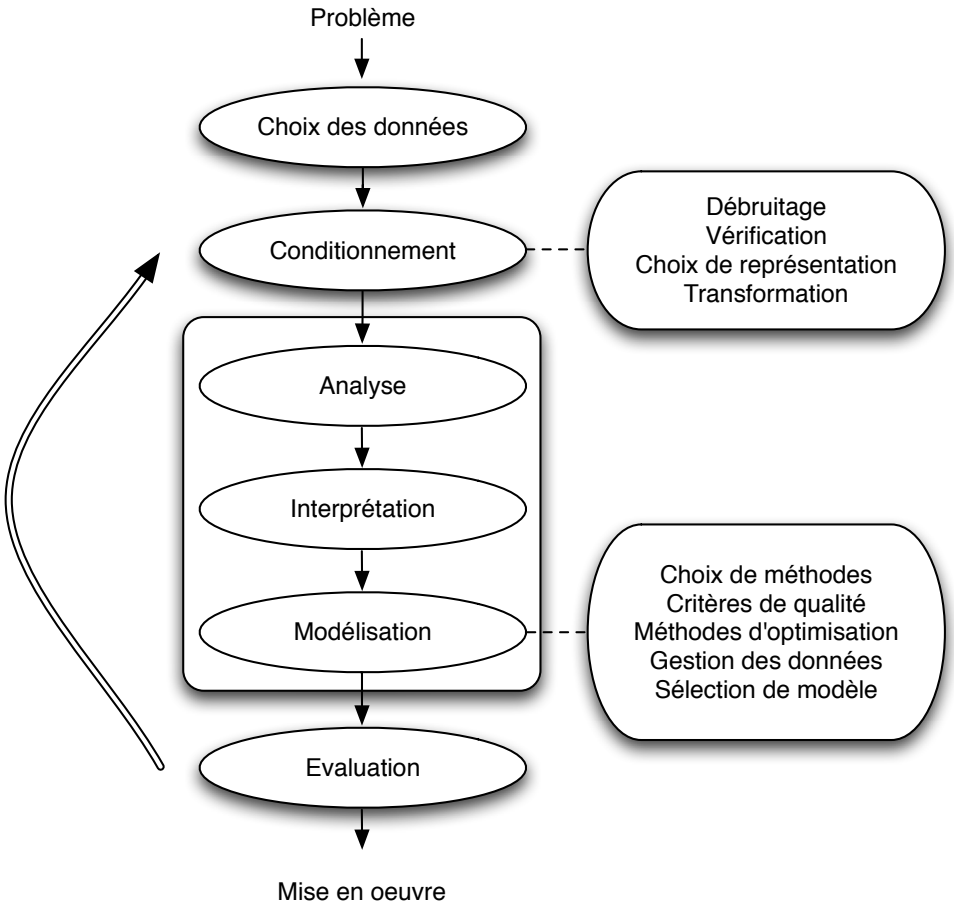
Malédiction de la dimensionnalité

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



Reconnaissance des formes

Chaine de traitement



Reconnaissance des formes

Critères pour le choix d'une méthode

Précision : Les performances de la méthode sont souvent privilégiées.Les critères de performance ci-dessous ne doivent pas être l'unique motivation du choix.

- ▷ erreur minimale ;
- ▷ taux minimal de faux positifs à taux d'erreur borné ;
- ▷ ...

Pour autant les critères ci-dessus ne doivent pas être l'unique motivation pour le choix d'une méthode.

Lisibilité : On peut souhaiter recueillir des décisions et résultats interprétables.

- ▷ pour des applications critiques, on ne peut se contenter d'une boite noire ;
- ▷ la lisibilité rend possible la vérification/validation.

Rapidité : Des contraintes de temps peuvent être décisives.

- ▷ rapidité de construction du modèle ;
- ▷ contrainte sur la prise de décision.

Contraintes liées à l'application.

Reconnaissance des formes

Critères pour le choix d'une méthode

Usability ou facilité d'emploi
Un expert est-il indispensable pour mettre au point le modèle et pour toute évolution ultérieure ?

Embedability ou facilité d'introduction dans un système global
La méthode impose-t-elle des contraintes sur l'échange de données d'entrée/sortie ?

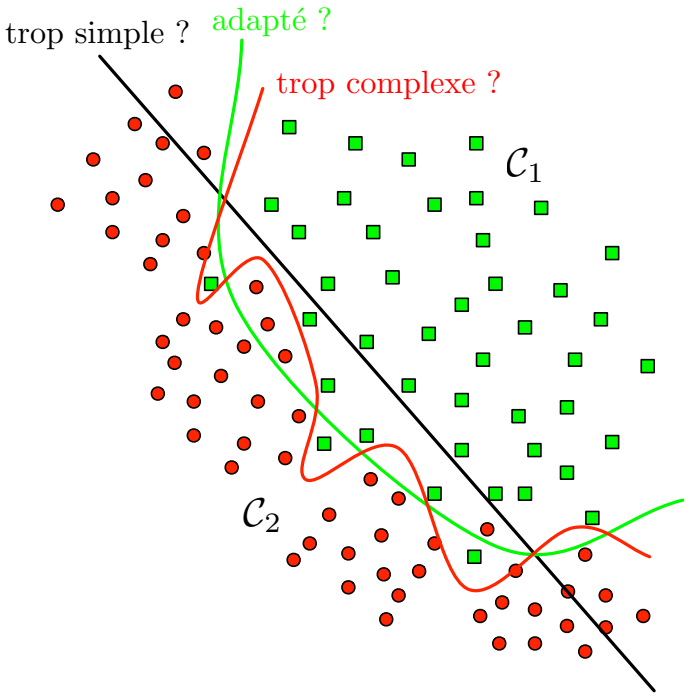
Flexibilité ou adaptation aisée au changement de spécifications
Faut-il reprendre tout le système si un capteur est remplacé par un autre, de courbe de réponse différente ?

Scalability ou passage à l'échelle
Le système s'accommode-t-il d'une augmentation du volume et débit de données ?

Reconnaissance des formes

Problèmes de sélection de modèle

Les modèles et règles élaborés doivent être dotés d'une bonne capacité de généralisation vis-à-vis de données autres que celles utilisées pour l'apprentissage.



Reconnaissance des formes

Problèmes de sélection de modèle

La capacité d'apprentissage d'un modèle est notamment gouvernée par

- le nombre de degrés de liberté ;
- les termes de régularisation ;
- ...

Elle doit être adaptée au problème traité et au nombre de données disponibles car

- capacité trop faible : inaptitude à représenter de façon satisfaisante les informations contenues dans les données disponibles ;
- capacité trop forte : problème de détermination des paramètres du modèle mal posé, conduisant à de nombreuses solutions équivalentes.

Reconnaissance des formes

Problèmes de sélection de modèle

Les performances estimées sur l'ensemble d'apprentissage ne constituent pas un bon indicateur pour l'ajustement des paramètres associés. Il convient de se munir de

- ▷ un ensemble de validation ;
- ▷ un ensemble de test éventuellement.

Un ensemble de validation de faible taille fournit une estimation imprécise des performances. S'il est de taille conséquente, l'apprentissage en souffre.