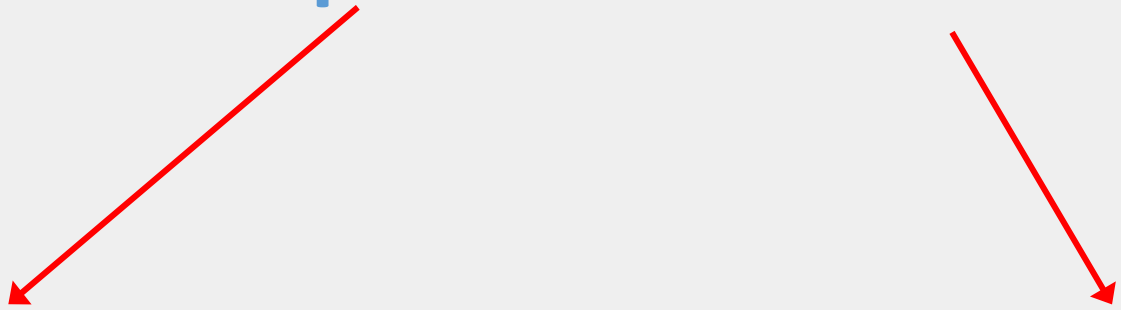
Chapitre 5

La classification supervisée:
Les K-plus proches voisins/arbres de décision

La classification supervisée

Apprentissage automatique
Apprendre = Optimiser & Généraliser



Optimisation :Méthodes « classiques » (programmation mathématique) et moins classiques (algorithmes génétiques …)	Théorie de la généralisation Principe de minimisation de l'erreur en généralisation estimée

La classification supervisée

Que veut dire bien classifier un ensemble D de données ?

Soit L une fonction de coût :

Par exemple L(f(x),y) = 1 si f(x) <> y et 0 sinon

Alors Erreur moyenne que commet f sur D :

$$Err_D(f) = 1/n \cdot \sum_{i=0}^{n} L(f(xi),yi)$$

Si $Err_D(f)$ est faible , bonne classification des données de D

La classification supervisée

Le problème de la généralisation

On veut déterminer f telle que f se comporte bien sur de nouvelles données du même type (même distribution)

$$L(f(x_i), y_i)$$

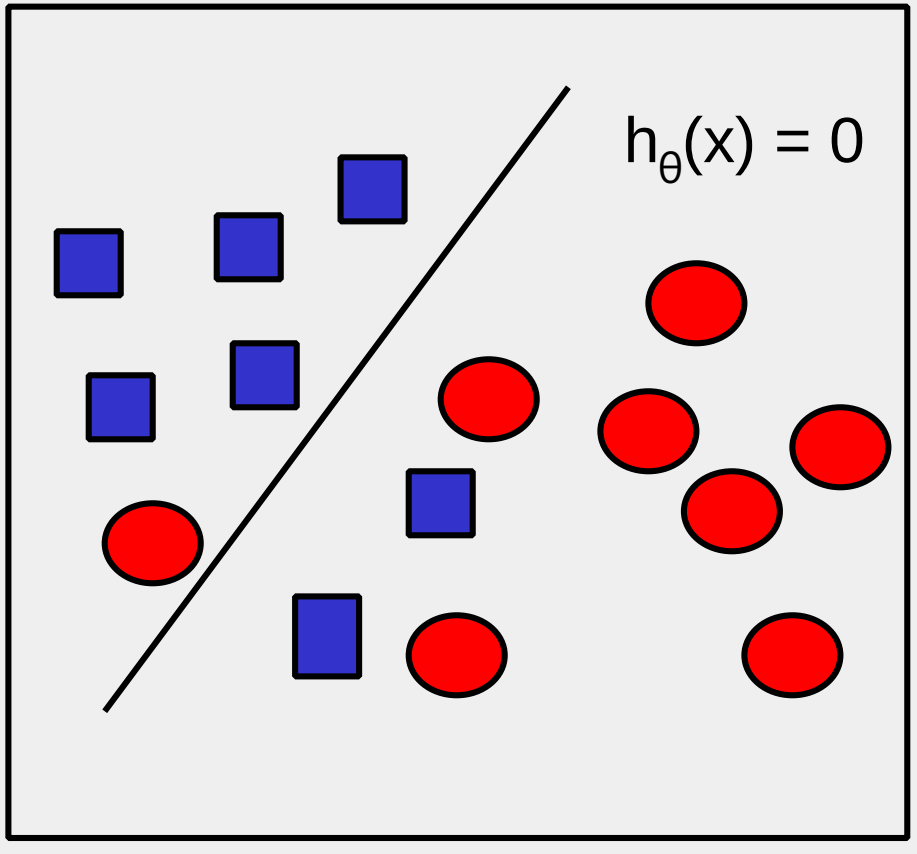
Comment savoir si f généralise bien ?

1° On peut calculer $Err_T(f)$, l'erreur que commet f en moyenne sur un ensemble de données de test T

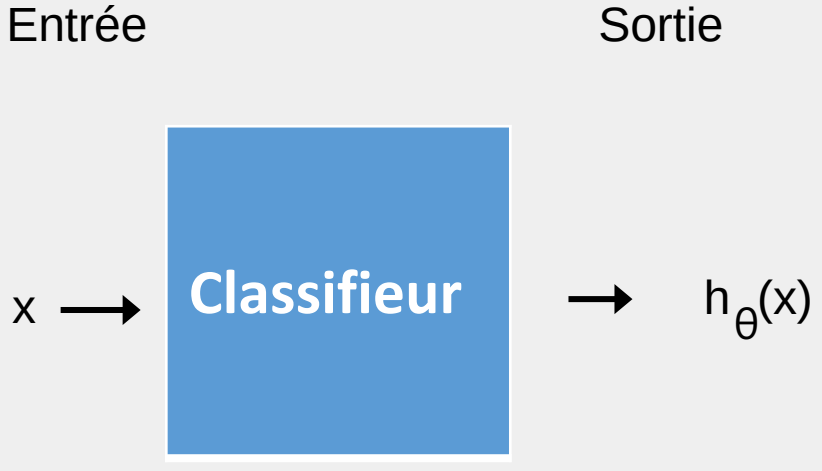
2° On peut procéder à de la validation croisée (cross validation)

La classification supervisée

Apprentissage supervisé



$h_\alpha \epsilon H$: famille de classifieurs



La classification supervisée

Méthodologie de l'apprentissage

Représentation

Comment représenter les données ? Quelle famille de modèles choisir (linéaires, non linéaires …)?

Optimisation

Quel est l' objectif à atteindre ?

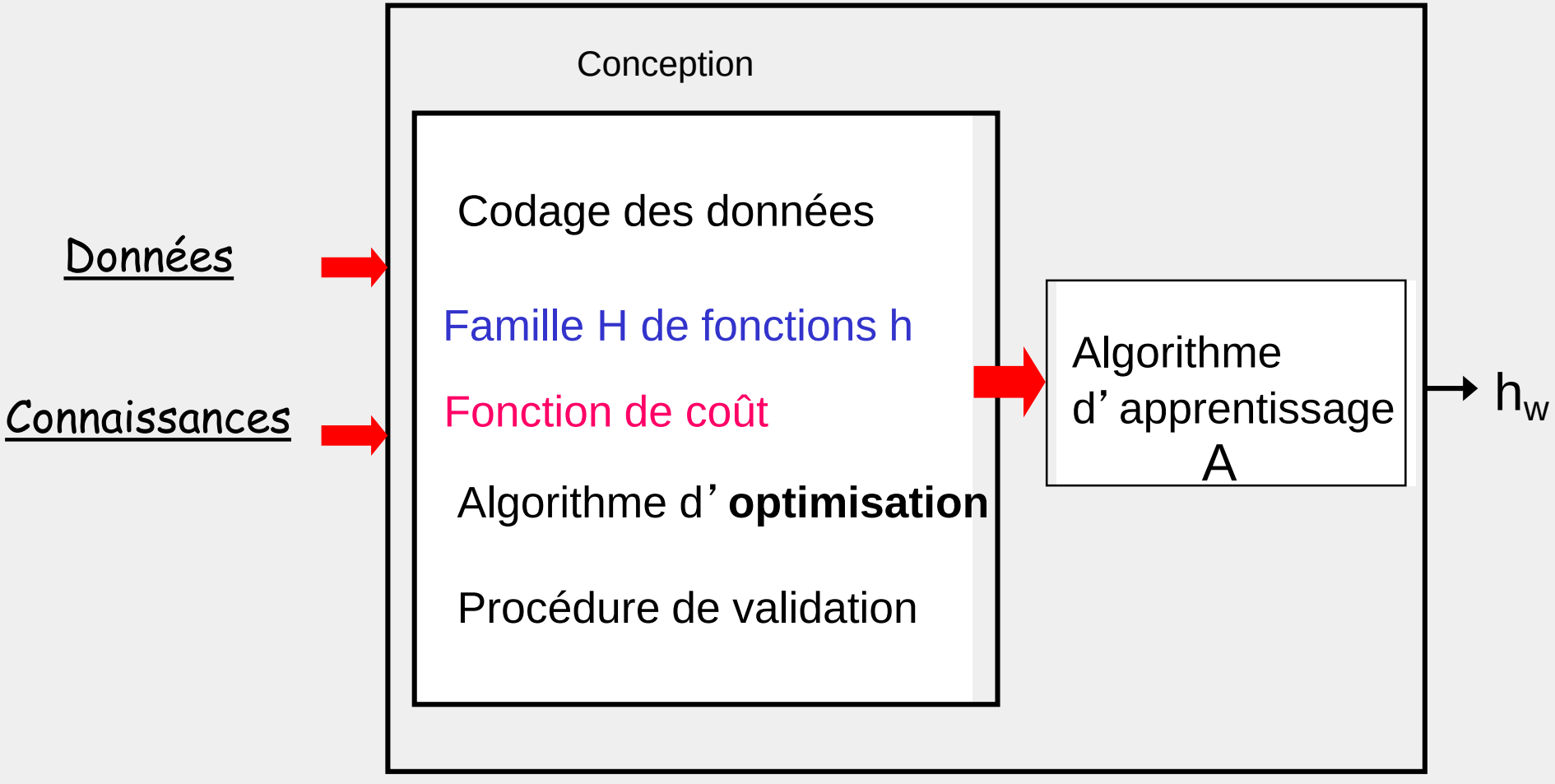
Comment le traduire par un problème d' optimisation ?

Quelle méthode utiliser ?

Validation interne

Sensibilité de la méthode au bruit, à la taille de l' échantillon Comportement en généralisation

La classification supervisée



Algorithme des K-plus proches voisins

Généralités

- Apprendre par analogie
  - Recherchant d' un ou des cas similaires déjà résolus
- Classifier ou estimer
- "Dis moi qui sont tes amis, et je te dirais qui tu es"
- Pas de construction de modèle
  - C'est l'échantillon d'apprentissage, associé à une fonction de distance et d'une fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constitue le modèle

Algorithme des K-plus proches voisins

Algorithme

- Paramètre : le nombre k de voisins
- Donnée : un échantillon de m exemples et leurs classes

– La classe d'un exemple X est c(X)

- Entrée : un nouvel exemple Y

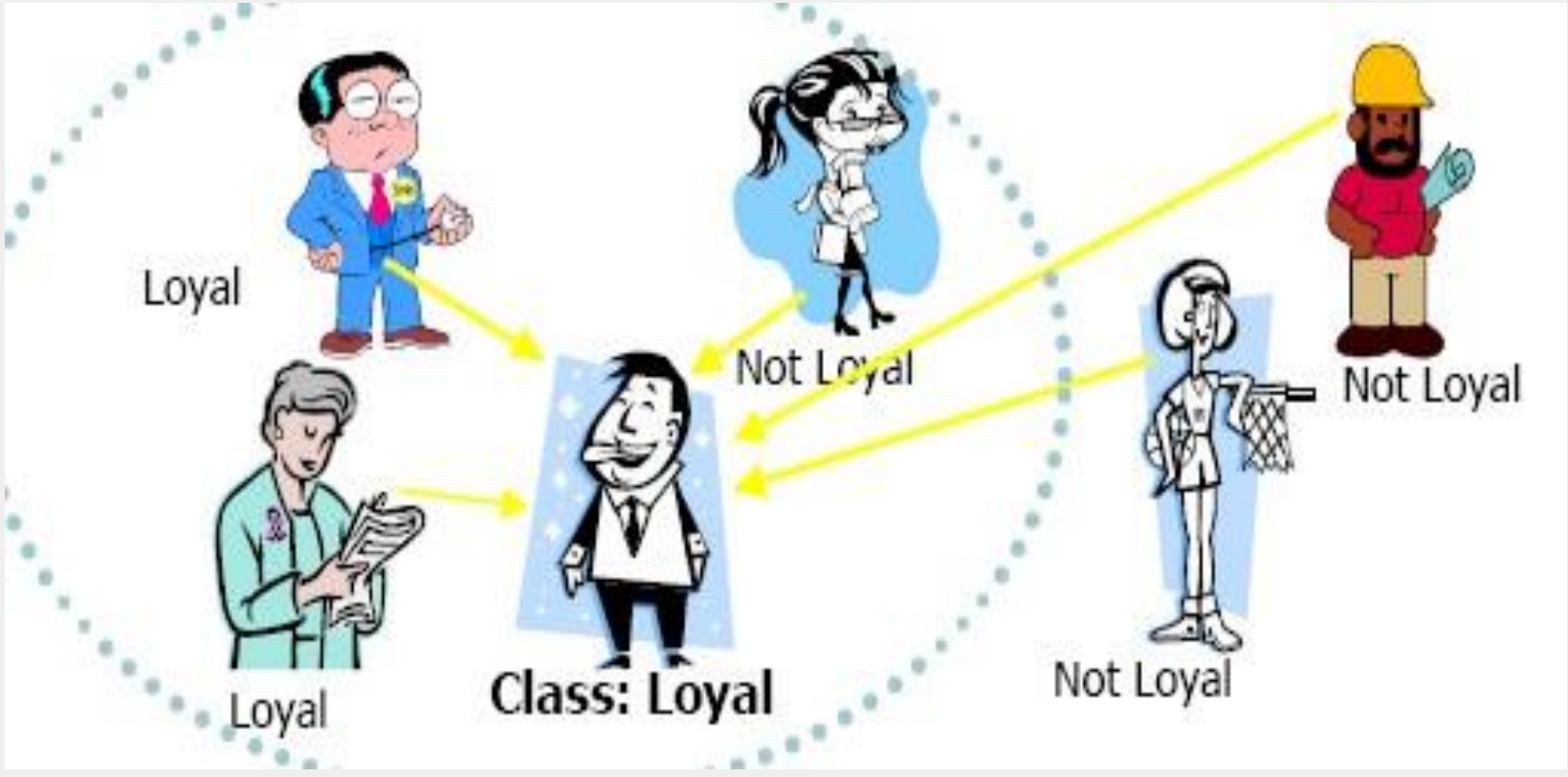
1. Déterminer les k plus proches exemples de Y en calculant les distances
2. Combiner les classes de ces k exemples en une classe c

- Sortie : la classe de Y est c(Y)=c

Algorithme des K-plus proches voisins

Exemple: Client loyal ou non

K = 3



Algorithme des K-plus proches voisins

Distance

- Le choix de la distance est primordial au bon fonctionnement de la méthode

- Les distances les plus simples permettent d'obtenir des résultats satisfaisants (lorsque c'est possible)

Algorithme des K-plus proches voisins

Choix de la classe

- Choix de la classe majoritaire
- Choix de la classe majoritaire pondérée
  - Chaque classe d'un des k voisins sélectionnés est pondéré
  - Soit V le voisin considéré. Le poids de w(V) est inversement proportionnel à la distance entre l'objet V' à classer et V
- Calculs d'erreur

Algorithme des K-plus proches voisins

Exemple

Customer	Age	Income	No. credit cards	Loyal
John 	35	35K	3	No
Rachel 	22	50K	2	Yes
Hannah 	63	200K	1	No
Tom 	59	170K	1	No
Nellie 	25	40K	4	Yes
David 	37	50K	2	?

Algorithme des K-plus proches voisins

Exemple

K = 3

Customer	Age	Income	No. credit cards	Loyal	Distance from David
John 	35	35K	3	No	sqrt [(35-37)²+(35-50)² +(3-2)²]=**15.16**
Rachel 	22	50K	2	Yes	sqrt [(22-37)²+(50-50)² +(2-2)²]=**15**
Hannah 	63	200K	1	No	sqrt [(63-37)²+(200-50)² +(1-2)²]=**152.23**
Tom 	59	170K	1	No	sqrt [(59-37)²+(170-50)² +(1-2)²]=**122**
Nellie 	25	40K	4	Yes	sqrt [(25-37)²+(40-50)² +(4-2)²]=**15.74**
David 	37	50K	2	**Yes**	

Algorithme des K-plus proches voisins

Mise en œuvre de la méthode

- Choisir les attributs pertinents pour la tâche de classification considérée et les données
- Choix de la distance en fonction du type des attributs et des connaissances préalables du problème
- Choix du nombre k de voisins déterminé par utilisation d'un ensemble test ou par validation croisée

– Une heuristique fréquemment utilisée est de prendre k égal au nombre d'attributs plus 1

Algorithme des K-plus proches voisins

Discussion

- Interprétations: La classe attribuée à un exemple peut être expliquée en exhibant les plus proches voisins qui ont amené à ce choix
- La méthode peut s'appliquer dès qu'il est possible de définir une distance sur les attributs
- La méthode permet de traiter des problèmes avec un grand nombre d'attributs.
  – Mais, plus le nombre d'attributs est important, plus le nombre d'exemples doit être grand.

Algorithme des K-plus proches voisins

Discussion

- Tous les calculs doivent être effectués lors de la classification (pas de construction de modèle)
- Le modèle est l'échantillon

  Espace mémoire important nécessaire pour stocker les données, et méthodes d'accès rapides nécessaires pour accélérer les calculs
- Les performances de la méthode dépendent du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins.

  En règle générale, les distances simples fonctionnent bien.