Université de Montréal

Département d'informatique et de recherche opérationnelle

IFT3390/6390
Fondements de l'apprentissage machine

http://www.iro.umontreal.ca/~vincentp/ift3390

Deuxième cours:

Terminologie de l'apprentissage supervisé

Méthodes à base de voisinage (formalisation)

Professeur: Pascal Vincent

LISA Laboratoire d'Informatique des Systèmes d'Apprentissage

Au programme aujourd'hui

- Bref rappel de points importants du premier cours
- Terminologie de l'apprentissage supervisé
- Formalisation mathématique des méthodes à base de voisinage

Apprendre à partir d'exemples !

"cheval"

"cheval"

"cheval"

Principe beaucoup plus général que d'écrire à la main, en partant de zéro, un algorithme pour reconnaître un cheval...

Les catégories de problèmes (tâches) standards de l'apprentissage automatique

Apprentissage supervisé

- Classification
- Régression

Apprentissage non supervisé

- Estimation de densité
- Partitionnement (*clustering*)
- Réduction de dimensionalité

Apprentissage par renforcement

Terminologie de l'apprentissage supervisé

Ensemble de données d'entrainement (*training set*)

Une entrée est généralement représentée par un vecteur de dimension d.

$x \in \mathbb{R}^d$

dimensionalité de l'entrée

1	entrée, observation, *input*, x_1	cible, *target*, sortie désirée, y_1
2	entrée, observation, *input*, x_2	cible, *target*, sortie désirée, y_2
3	entrée, observation, *input*, x_3	cible, *target*, sortie désirée, y_3
⋮	etc...	⋮
n	entrée, observation, *input*, x_n	cible, *target*, sortie désirée, y_n

taille de l'ensemble, nombre d'exemples, d'échantilons.

point de test → ? ? ?

On cherche un algorithme qui produit une sortie (*output*) qui est une bonne prédiction de la cible.
Cet algorithme trouve une bonne fonction x -> y

Terminologie de l'apprentissage supervisé

- Lorsque la cible est une étiquette de classe, une variable catégorique (indiquant à quelle classe ou catégorie l'entrée appartient, parmi plusieurs) on dit qu'on a affaire à un problème de CLASSIFICATION. (on utilise souvent un entier comme étiquette).
- Lorsque la cible est une (ou plusieurs) valeur réelle à prédire, on parle de problème de RÉGRESSION.

Quand on n'a pas de cible explicite, on est dans le cadre de l'apprentissage non supervisé.

Ex. de problème de classification

Ensemble de données d'entrainement (*training set*)

étiquette *label* y_i

entrée x_i

(représentation vectorielle)

$x \in \mathbb{R}^d$

$x = (0, 0, \ldots, 54, 120, \ldots, 0, 0)$

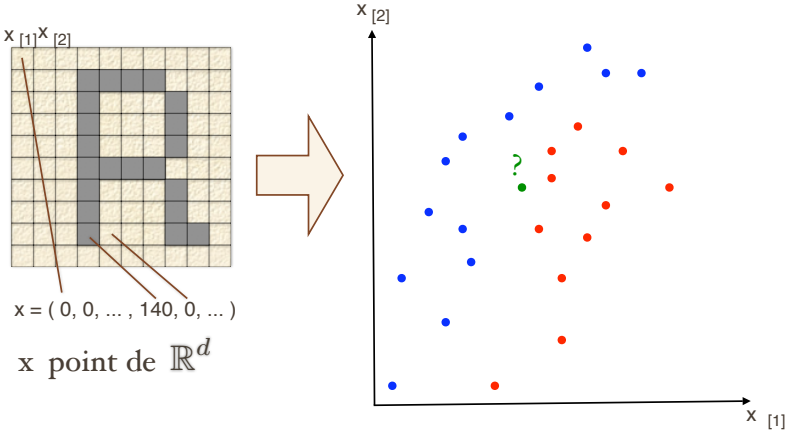
Apprendre n'est pas simplement mémoriser...

C'est être capable de généraliser à de nouveaux cas!

Point de test: (nouveau x)

→ 2 ou 3?

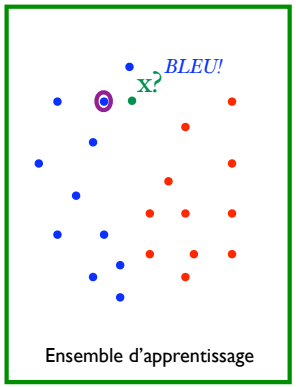
Représentation des données



L'algorithme du plus proches voisin

Pour un point test x:

- On trouve **le plus proche voisin** de x parmi l'ensemble d'apprentissage selon une certaine mesure de distance (ex: distance Euclidienne).
- On associe à x la classe de ce plus proche voisin.



L'algorithme classique des k plus proches voisins (kNN)

Pour un point test x:

- On trouve les k plus proches voisins de x parmi l'ensemble d'apprentissage (typiquement selon la distance Euclidienne).
- On associe à x la classe majoritaire parmi ses k voisins