



Introduction to clustering

Philippe Lenca et Romain Billot

Institut Mines Telecom, Telecom Bretagne
UMR CNRS 3192 Lab-STICC
[prenom.nom]@telecom-bretagne.eu





Plan du cours

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

- 1 Introduction
- 2 Organisation du cours
- 3 Méthodes de base
 - La classification non hiérarchique
 - La classification hiérarchique
 - Mesures de validation d'une classification
- 4 Différencier apprentissage non supervisé et apprentissage supervisé
 - Principe et vocabulaire
 - Méthodologie d'apprentissage
 - Un algorithme simple : les k plus proches voisins
- 5 Bibliographie



Plan

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

1 Introduction

2 Organisation du cours

3 Méthodes de base

- La classification non hiérarchique
- La classification hiérarchique
- Mesures de validation d'une classification

4 Différencier apprentissage non supervisé et apprentissage supervisé

- Principe et vocabulaire
- Méthodologie d'apprentissage
- Un algorithme simple : les k plus proches voisins

5 Bibliographie



- Dans beaucoup de pays, le mot anglais *clustering* est utilisé,
- En France, on parlera de classification automatique ou classification non supervisée,
- Attention à ne pas confondre classification et classement, ni à utiliser le mot *classification* en anglais qui veut dire autre chose.

La terminologie dépend bien sûr du domaine :

- Medecine : la nosologie est la classification des maladies,
- Marketing, Enquêtes : typologie,
- Sciences naturelles : taxinomie ou taxonomie.

Les méthodes de classification non supervisée ou algorithmes de regroupement permettent la construction automatique de telles classifications



Exemple : classification des espèces

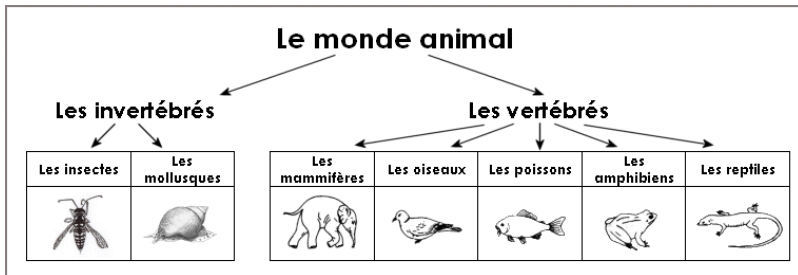
Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.



<http://soutien67.free.fr/svt/animaux/classification/classification01.htm>



Exemple : classification des espèces

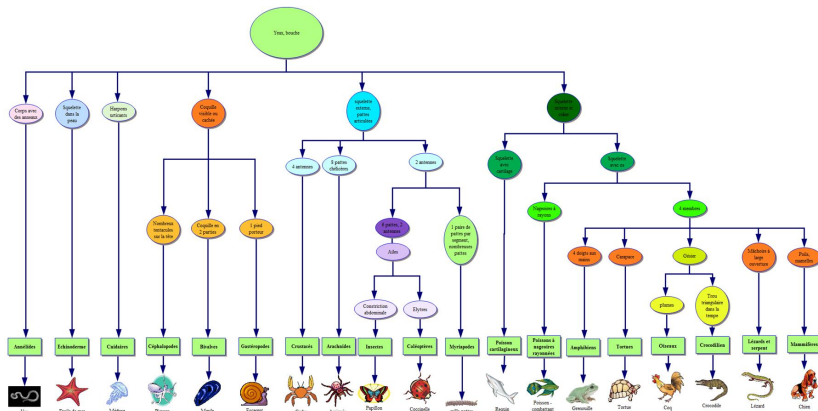
Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.



<http://www.cours-svt.fr/sixieme/fiche-chapitre-0>



Exemple : classification de l'escargot de Bourgogne

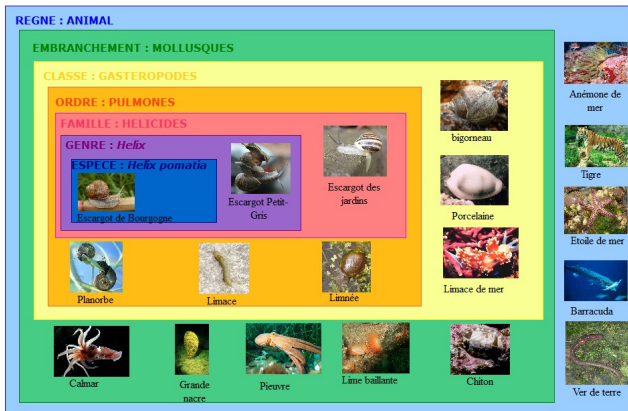
Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.



https://fr.wikidia.org/wiki/Classification_classique

Exemple : classification d'individus dans un rassemblement

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.



<http://www.halalbook.fr/>



Exemple : classification des étoiles








Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Étoiles de la séquence principale						
						
Type spectral : O	B	A	F	G	K	M
Température : 40 000K	20 000K	8500K	6500K	5700K	4500K	3200K
Rayon (Soleil=1) : 10	5	1.7	1.3	1.0	0.8	0.3
Masse (Soleil=1) : 50	10	2.0	1.5	1.0	0.7	0.2
Luminosité : 100 000	1000	20	4	1.0	0.2	0.01
Durée (Millions ans) : 10	100	1000	5000	10 000	50 000	100 000
Abondance : 0.00001%	0.05%	0.3%	1.5%	4%	9%	80%
Étoiles Géantes		Naines blanches		Étoiles Supergéantes		
Étoile de faible masse à la fin de sa vie		Résidu mourant d'une étoile explosée		Étoile massive à la fin de sa vie		
Type spectral : F, G, K or M		Type spectral : D		Type Spectral : O, B, A, F, G, K or M		
Température : 3000 to 10 000		Température : Under 50 000K		Température : 4000 to 40 000K		
Rayon (Soleil=1) : 10 to 50		Rayon (Soleil=1) : Under 0.01		Rayon (Soleil=1) : 30 to 500		
Masse (Soleil=1) : 1 to 5		Masse (Soleil=1) : Under 1.4		Masse (Soleil=1) : 10 to 70		
Luminosité : 50 to 1000		Luminosité : Under 0.01		Luminosité : 30 000 to 1000 000		
Durée (Millions ans) : 1000		Durée (Millions ans) : -		Durée (Millions ans) : 10		
Abondance : 0.5%		Abondance : 5%		Abondance : 0.0001%		

<http://atunivers.free.fr/2501ys/startype.html>



Plan

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

1 Introduction

2 Organisation du cours

3 Méthodes de base

- La classification non hiérarchique
- La classification hiérarchique
- Mesures de validation d'une classification

4 Différencier apprentissage non supervisé et apprentissage supervisé

- Principe et vocabulaire
- Méthodologie d'apprentissage
- Un algorithme simple : les k plus proches voisins

5 Bibliographie



Principe de la classification non supervisée

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Analyse de données

Ensembles de méthodes de réduction de données,
représentation simplifiée des données initiales

Principe de la classification automatique

Organisation d'un ensemble en classes homogènes, naturelles,
sans hypothèse *a priori* sur la structure des données. Permet
de mieux comprendre les données.

Point de départ : un tableau individus/variables

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Les données à traiter se présentent sous la forme d'un tableau numérique de taille (N, P) correspondant à un ensemble Ω de N individus, observations ou exemples pour lesquels nous connaissons la valeur de P variables ou attributs.

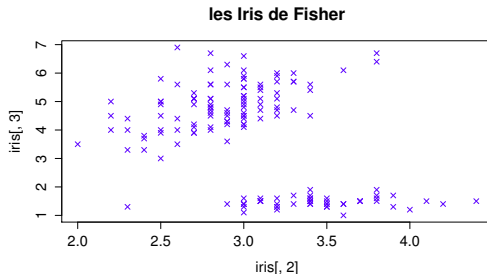


FIGURE : Combien de classes naturelles ?



Proximité entre individus

Toute méthode de classification non supervisée se fonde sur une mesure de proximité plus ou moins complète entre individus, par exemple :

- Une distance : euclidienne, Manhattan, Mahalanobis, χ^2 pour un tableau de fréquence, Jaccard pour données binaires etc.
- Une mesure de similarité ou dissimilarité,
- Une ultramétrie (distance plus stricte) pour la classification hiérarchique.

Choix de la métrique

Le bon choix de la métrique est un pré-requis à toute méthode d'analyse de données !

But de la classification non supervisée : former une partition

Une partition de Ω est un ensemble de parties non vides de Ω , P_1, P_2, \dots, P_k vérifiant :

$$\forall i \neq j \ P_i \cap P_j = \emptyset$$

$$\bigcup_{i=1}^k P_i = \Omega.$$

Explosion combinatoire

Le nombre de partitions d'un ensemble de n points en k classes est équivalent à $\frac{k^n}{k!}$ si $n \rightarrow \infty$. Les quelques méthodes présentées aujourd'hui recherchent par conséquent des heuristiques, qui optimisent localement certains critères bien choisis.



Plan

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

1 Introduction

2 Organisation du cours

3 Méthodes de base

- La classification non hiérarchique
- La classification hiérarchique
- Mesures de validation d'une classification

4 Différencier apprentissage non supervisé et apprentissage supervisé

- Principe et vocabulaire
- Méthodologie d'apprentissage
- Un algorithme simple : les k plus proches voisins

5 Bibliographie



Partition floue et partition dure

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

La structure de partition classique (*crisp clustering*) impose une appartenance stricte à une classe ($c_{ik} \in \{0, 1\}$). La notion de partition floue (méthodes de *fuzzy clustering*) généralise en quelque sorte la notion de partition classique en associant à chaque individu un vecteur d'appartenance U aux différentes classes vérifiant

$$\forall i, k \ c_{ik} \in [0, 1] \text{ et } \forall i \sum_{k=1}^q c_{ik} = 1.$$



La méthode des centres mobiles (*k-means*)

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

- La méthode des *k-means* est aussi appelée centres-mobiles ou réallocation-centrage,
- Objectif : partitionner un ensemble $X = x_1, \dots, x_N$ en K classes, le nombre K de classes (ou clusters) étant fixé par avance. L'ensemble X appartient (ici) à l'espace vectoriel \mathbb{R}^p muni de la distance euclidienne.
- Tous les individus ont généralement des pondérations égales mais il est tout à fait possible d'introduire une pondération des différents exemples.



Détail de l'algorithme des *k-means*

Entrées : le vecteur des N instances $(x_1, \dots, x_N) \forall i = 1, \dots, N$ avec $x_i \in \mathbb{R}^d$; le nombre K initial de clusters (partitions, groupes) souhaité ;

Sorties : Une partition en K groupes

début

Sélectionner aléatoirement K centres initiaux parmi les instances de départ.

pour i allant de 1 à N **faire**

 Affecter chaque point à la classe du centre le plus proche

fin

pour k allant de 1 à K **faire**

 Recalculer les nouveaux centres des K classes

fin

Répéter les deux étapes précédentes jusqu'à ce que les centres des classes ne varient plus. (Convergence)

fin



Illustration

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

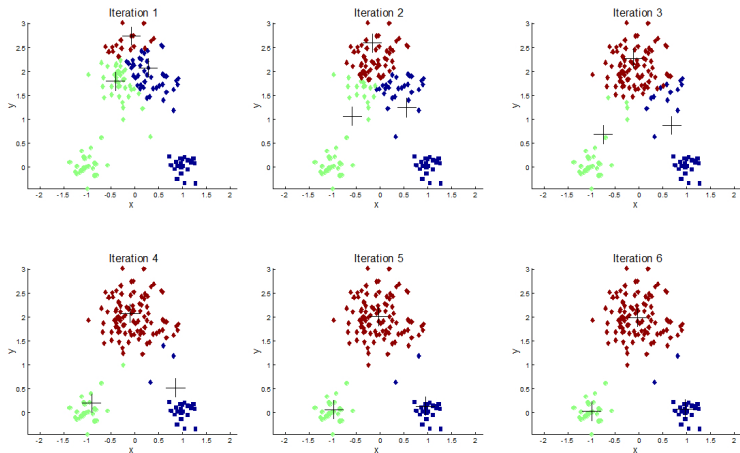


FIGURE : Visualisation des différentes étapes.



Critère de qualité et convergence

Le critère de qualité d'un couple partition-centres se définit comme la somme des inerties des classes par rapport à leur centre, à savoir

$$C(P, L) = \sum_{k=1}^K \sum_{x \in P_k} d^2(x, \lambda_k)$$

où $P = (P_1, P_2, \dots, P_K)$ est une partition de Ω en K classes et $L = (\lambda_1, \lambda_2, \dots, \lambda_K)$ représente un K -uple de \mathbb{R}^p . L'algorithme des centres mobiles va optimiser localement ce critère pour former assez rapidement (une dizaine d'itérations *a maxima*) un couple (P, L) . La méthode des centres-mobiles construit donc des partitions successives en diminuant l'inertie intra-classe.



Faiblesses de la méthode

- Inhérente à toute méthode de regroupement : choix *a priori* du nombre de classe K (voir plus tard),
- Optimisation locale : on fournit une suite de couples $C(P, L)$ dont la valeur du critère va en décroissant,
- Sensibilité aux données aberrantes et valeurs extrêmes.

Une amélioration : *Partitioning Around Medoids (PAM)*

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Entrées : le vecteur des N instances $(x_1, \dots, x_N) \forall i = 1, \dots, N$ avec $x_i \in \mathbb{R}^d$; le nombre K initial de clusters ;

Sorties : Une partition en K groupes

début

Sélectionner aléatoirement K médoïdes initiaux parmi les instances de départ.

1. **pour** i allant de 1 à N **faire**

 Affecter chaque point à la classe du centre le plus proche

fin

2. **pour** k allant de 1 à K **faire**

 3. **pour** chaque point non médoïde b **faire**

 Echanger b et k et recalculer le coût total

fin

fin

4. Sélectionner la configuration avec le coût minimum.

Répéter les étapes 2, 3, 4 jusqu'à ce que les centres des classes ne varient plus (Convergence).

fin

Algorithme 2: Procédure de l'algorithme *PAM*



Exemple graphique (1)

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

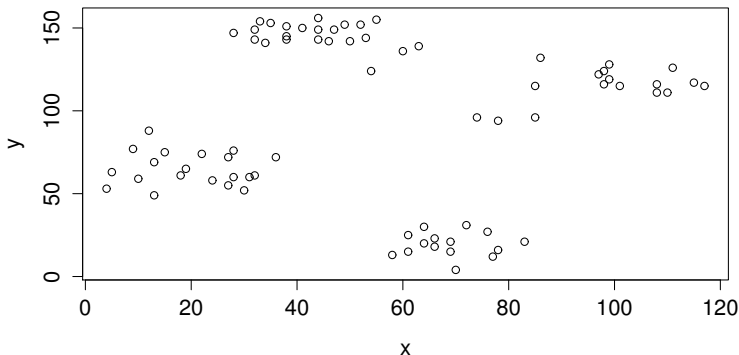


FIGURE : Visualisation de la classification du jeu de données Ruspini



Exemple graphique (1)

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

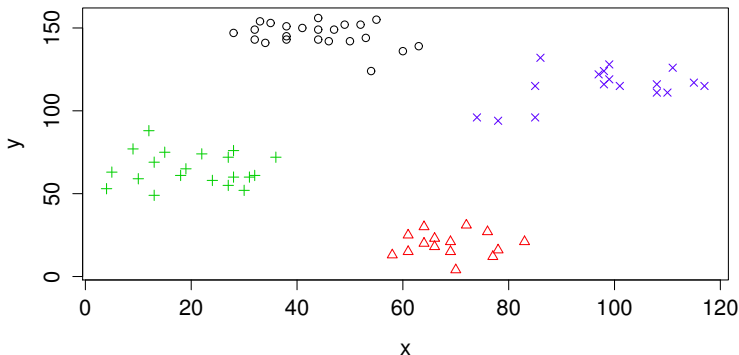


FIGURE : Visualisation de la classification du jeu de données Ruspini



Exemple graphique (2)

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

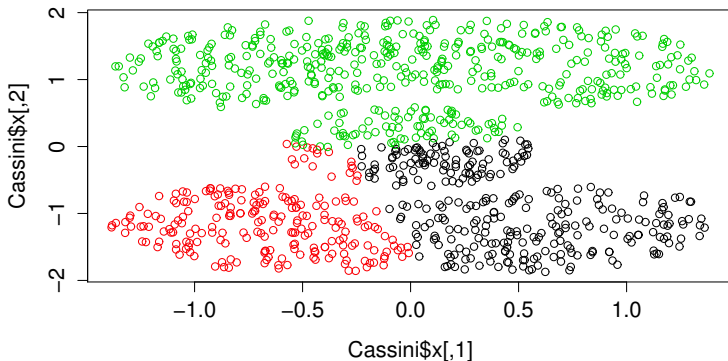


FIGURE : *K-means* : visualisation de la classification du jeu de



Exemple graphique (2)

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

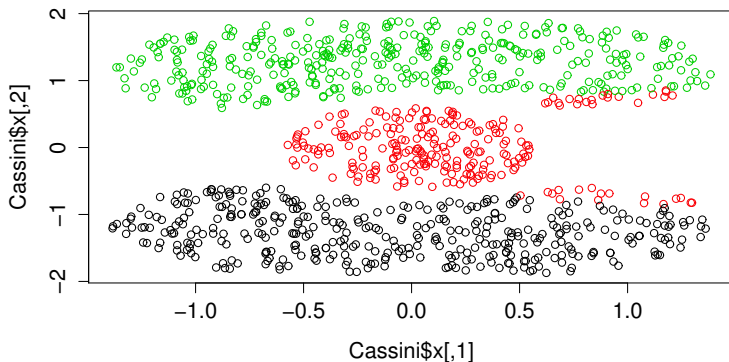


FIGURE : *PAM* : visualisation de la classification du jeu de données



La classification hiérarchique

- Il s'agit de former une hiérarchie indicée, *i.e* une suite de partitions emboîtées,
- Construction graduelle d'un arbre hiérarchique, appelé dendrogramme, par optimisation d'un critère à chaque itération.

Deux approches :

1. **Méthode dite divisive ou descendante** : consiste à partir d'une unique partition comportant tous les exemples pour la diviser ensuite itérativement en sous-groupes, jusqu'à obtenir des singletons,
2. **Méthode ascendante**, ou *agglomerative* en anglais : va quant à elle partir des singletons pour fusionner progressivement les classes les plus similaires par regroupements successifs et terminer la procédure à l'obtention de la partition "racine" comportant tous les exemples,

Classification Ascendante Hiérarchique (CAH)

Introduction

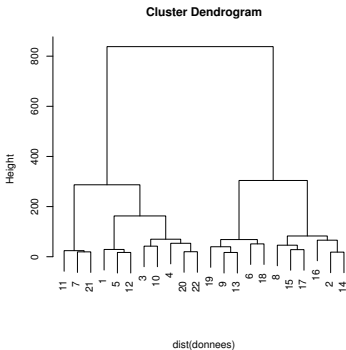
Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

A chaque itération, les classes en présence, en partant des singletons, sont fusionnées selon l'optimisation d'un critère se rapportant à la matrice de dissimilarité de départ, de taille $N \times N$.





Critère d'agrégation

A chaque niveau de la hiérarchie, les deux classes les plus proches au sens d'un certain critère sont regroupées :

- Le critère du lien minimum (single link)

Les classes de plus petite dissimilarité entre elles sont regroupées, soit pour deux classes A et B

$$D(A, B) = \min d(x, y), \quad x \in A, y \in B;$$

- Le critère du lien maximum

$$D(A, B) = \max d(x, y), \quad x \in A, y \in B;$$

- Le critère de la distance moyenne (average link)

$$D(A, B) = \frac{\sum_{x \in A} \sum_{y \in B} d(x, y)}{|A| |B|}$$



Le critère de Ward

A chaque étape d'une classification ascendante hiérarchique (CAH), la fusion entre deux classes ira forcément de pair avec une augmentation de l'inertie intra-classe et une diminution de l'inertie inter-classe. Il s'agit pour le critère de Ward de minimiser l'inertie intra-classe

$$D(A, B) = \frac{n_A n_B}{n_A + n_B} d^2(g(A), g(B)),$$

où n_J représente le nombre d'instances de la classe J si les pondérations sont unitaires. Sinon n_J représente la somme des pondérations des éléments d'une classe J . g_J désigne le centre de gravité de la classe J .



Comment valider un *clustering* ?

Le problème de la classification non supervisée réside dans le choix du nombre K optimum puis la validation de la partition proposée. Deux cas possibles :

1. On fait appel à un expert pour valider un nombre K de classes et une partition formée,
2. On recherche la meilleure partition pour plusieurs nombres de classes, plusieurs méthodes, au sens d'un certain critère (méthode du coude).



Critères de qualité d'une partition

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Champ de recherches à part entière, citons quelques approches :

1. Isolation et connectivité (Pauwels et Frederix, 1999),
2. Stabilisation des partitions par ré-échantillonnage Bootstrap,
3. Critères de qualités internes fondés sur l'inertie intra- et inter-classe : *Dunn*, *Davies* – *Bouldin*, *Silhouette*, ratio *WB*, *gap* etc.



Exemple de la silhouette

La méthode de la silhouette moyenne se fonde tout d'abord sur une mesure de dissimilarité d'un point i à un cluster C , représentée par la distance moyenne du point i au cluster C :

$$d_{i,C} = \frac{1}{|C|} \sum_{x_t \in C} d(x_i, x_t).$$

Soient ensuite $b_i = \min_{C \neq C(i)} d_{i,C}$ et $a_i = d_{i,C(i)}$. La silhouette d'une instance i s'exprime

$$Sil_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

On cherche ensuite à maximiser la silhouette moyenne d'une partition :

$$S = \frac{1}{N} \sum^N Sil_i.$$



Plan

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

- 1 Introduction
- 2 Organisation du cours
- 3 Méthodes de base
 - La classification non hiérarchique
 - La classification hiérarchique
 - Mesures de validation d'une classification
- 4 Différencier apprentissage non supervisé et apprentissage supervisé
 - Principe et vocabulaire
 - Méthodologie d'apprentissage
 - Un algorithme simple : les k plus proches voisins
- 5 Bibliographie



Principe et vocabulaire

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

La dichotomie entre apprentissage supervisé et non supervisé se fonde sur la connaissance *a priori* ou non d'information sur les classes relatives aux individus.

- L'apprentissage supervisé désigne un cadre où les exemples sont reliés à une information relative à leur classe, à un concept,
- Les méthodes supervisées produisent par la suite, à partir d'une base d'exemples d'apprentissage pour lesquels la classe est connue, une règle de décision visant à prédire la classe de nouvelles observations.
- Cette règle de décision, appelée aussi classifieur, modèle ou hypothèse, peut être considérée géométriquement comme une hypersurface séparant les exemples représentés dans un espace multidimensionnel.



Apprentissage supervisé (2)

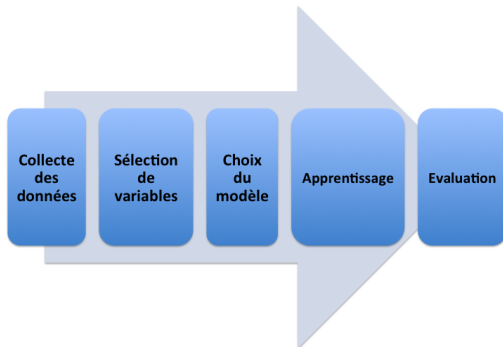
Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.



La difficulté dépend de la nature de la classe Y à prédire :

1. $Y = \mathbb{R}^q$: regression multiple de y en x ,
2. $Y = -1, 1$: **discrimination en deux classes**,
3. $Y = 1, \dots, q$: discrimination en q classes.



Exemple : ho, la belle prise...

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.



- Une entreprise de conditionnement de poissons décide de mettre en place un processus de tri des poissons selon l'espèce, à la chaîne, à l'aide de caméras,
- Etude pilote sur deux espèces : saumons et bars,
- Analyse d'un échantillon d'images.



Comment différencier un bar d'un saumon ?

Introduction

Organisation du cours

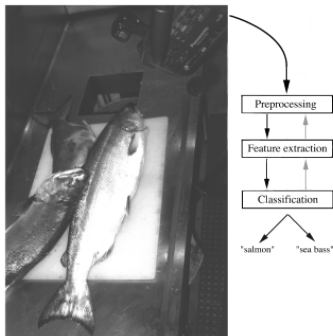
Méthodes de base

Non supervisé et supervisé

Biblio.

Sélection de variables

Les premières images permettent de sélectionner des variables discriminantes :





Comment différencier un bar d'un saumon ?

Introduction

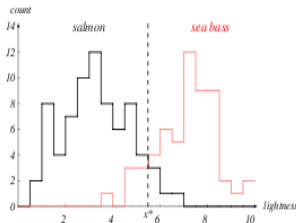
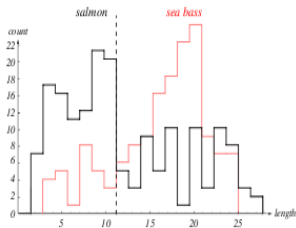
Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Il semble que la longueur et la luminosité de la peau peuvent être des facteurs discriminants



Ensemble d'apprentissage et règle de décision

Introduction

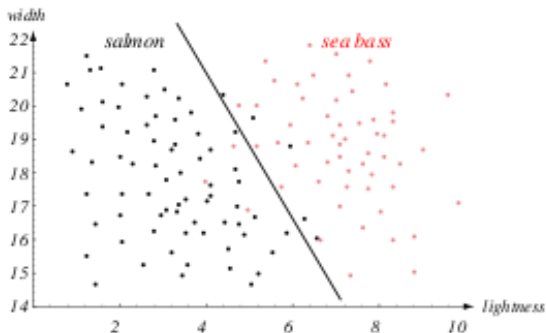
Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

On dispose d'un ensemble d'apprentissage sur lequel nous allons construire une règle de décision :



Ensemble d'apprentissage et règle de décision

Introduction

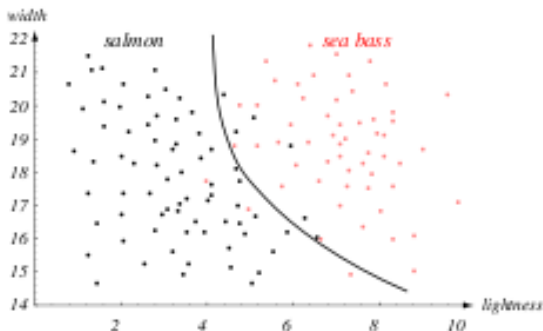
Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

On dispose d'un ensemble d'apprentissage sur lequel nous allons construire une règle de décision :



Ensemble d'apprentissage et règle de décision

Introduction

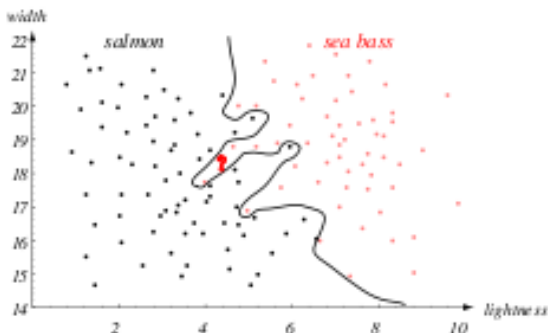
Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Cette règle de décision dépend de la méthode et sera plus ou moins raffinée :





Généralisation

Introduction

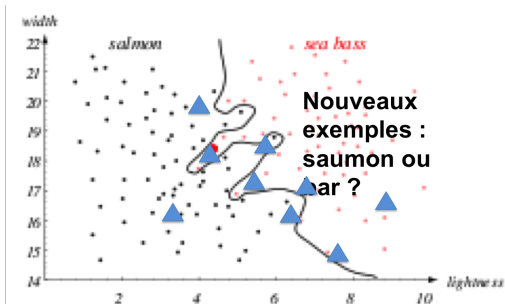
Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

La règle de décision sera ensuite testée sur de nouveaux exemples :



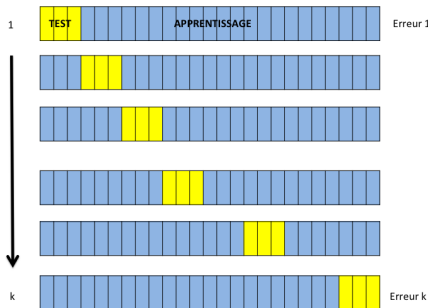
Evaluation

On évalue les performances d'une méthode à travers le taux d'erreur (mauvaise classification) sur de nouvelles données.



On divise le jeu de données selon le triptyque suivant :

1. Ensemble d'apprentissage,
2. Ensemble de test,
3. Ensemble de validation,





On recherche une certaine parcimonie dans la construction du modèle (principe du rasoir d'Ockham)

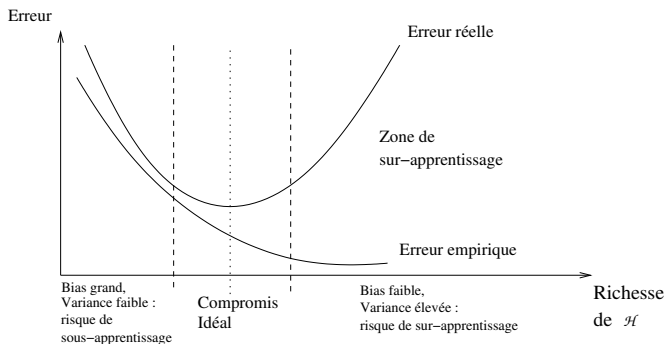


FIGURE : Attention au surapprentissage !

Un algorithme simple : les k plus proches voisins

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

Entrées : N instances (x, y) , $y_i \in -1, 1$; une distance d entre les instances ; un paramètre k entier positif ;

Sorties : Une étiquette pour chaque nouvelle observation

début

pour *toute nouvelle observation x* **faire**

 1. Calcul des $d(x, x_i)$

 2. Tri par ordre croissant,

 3. Prédiction y pour x : valeur majoritaire parmi les k plus proches voisins.

fin

fin

Algorithme 3: Procédure de l'algorithme kNN



Plan

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.

- 1 Introduction
- 2 Organisation du cours
- 3 Méthodes de base
 - La classification non hiérarchique
 - La classification hiérarchique
 - Mesures de validation d'une classification
- 4 Différencier apprentissage non supervisé et apprentissage supervisé
 - Principe et vocabulaire
 - Méthodologie d'apprentissage
 - Un algorithme simple : les k plus proches voisins
- 5 Bibliographie



Références I

Introduction

Organisation du cours

Méthodes de base

Non supervisé et supervisé

Biblio.



MIRKIN, BORIS. *Clustering : a data recovery approach*. CRC Press, 2012.



JAIN, ANIL K. ET DUBES, RICHARD C. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.



GOVAERT, GÉRARD (ed.). *Data analysis*. John Wiley and Sons, 2013.



JAIN, ANIL K., MURTY, M. NARASIMHA, ET FLYNN, PATRICK J. Data clustering : a review. *ACM computing surveys (CSUR)*, 1999, vol. 31, no 3, p. 264-323.



JAIN, ANIL K. Data clustering : 50 years beyond K-means. *Pattern recognition letters*, 2010, vol. 31, no 8, p. 651-666.



HALKIDI, MARIA, BATISTAKIS, YANNIS, ET VAZIRGIANNIS, MICHALIS. On clustering validation techniques. *Journal of intelligent information systems*, 2001, vol. 17, no 2, p. 107-145.