Classification

Farah@labged.net

Plan du cours

- 1- Introduction

  définitions, notion de système, applications
- 2- Classification

  pré-traitements (ACP), extraction de caractéristiques, représentation, classification, combinaison
- 3- Classification « statistique » des formes

  - théorie bayésienne de la décision
  - méthodes paramétriques ou non / bayésiennes ou non
  - arbres de décision
  - réseaux de neurones

Nouveau

Ancien

Connues

Connues

Définition 1

L'objectif principal des méthodes de classification automatique est de répartir les éléments d'un ensemble en groupes, c'est-à-dire d'établir une partition de cet ensemble. Différentes contraintes sont bien sûr imposées, chaque groupe devant être le plus homogène possible, et les groupes devant être les plus différents possibles entre eux.

Définition II

- Taxinomie, taxonomie
  - Étude théorique des bases, lois, règle, principes d'une classification
    - Classification des plantes, animaux, microbes, science fondatrice de la biologie
    - Livre : « L'analyse des données, La taxinomie », J.B. Benzécri, 1973, Dunod
    - Taxinomie des syntagmes !!!
- Catégorisation (plus spécifique que classe)
  - Classement par catégories, notamment en linguistique, en psychologie sociale

Classification supervisée: classes des documents

- articles scientifiques à regrouper en paquets homogènes
  - thème général (mathématique, physique, littérature …)
  - date de publication, nom des auteurs
  - Ceux qui traitent à la fois d'informatique et de biologie
  - Ceux qui se ressemblent
- selon un certain critère
  - Des critères précis aux critères vagues

Généralités

Données = tableau n*p individus*variables

Objectif = recherche d'une typologie ou segmentation, c'est à dire d'une partition ou répartition des n individus dans des classes, sur la base de l'observation de p descripteurs

Moyen = chaque classe doit être la plus homogène possible et, entre elles, les plus distinctes possibles, au sens d'un critère à définir.

Le Web présenté par sujets et par catégories

Achats
Gastronomie, Habillement, Musique, ...

Actualité
A la Une, Presse, Télévision, ...

Arts
Cinéma, Littérature, Musique, ...

Commerce et économie
Emploi, Immobilier, ...

Formation
Formation professionnelle, Universités, ...

Informatique
Ordinateurs, Logiciels, ...

Internet
Actualités, Recherche, ...

Jeux
Jeux de rôle, Jeux vidéo, ...

Loisirs
Collections, Humour, Tourisme, ...

Maison
Bricolage, Décoration, Jardin, ...

Références
Bibliothèques, Langues, ...

Régional
France, Amérique, Europe, ...

Santé
Maladies, Médecine, ...

Sciences
Astronomie, Sciences humaines, ...

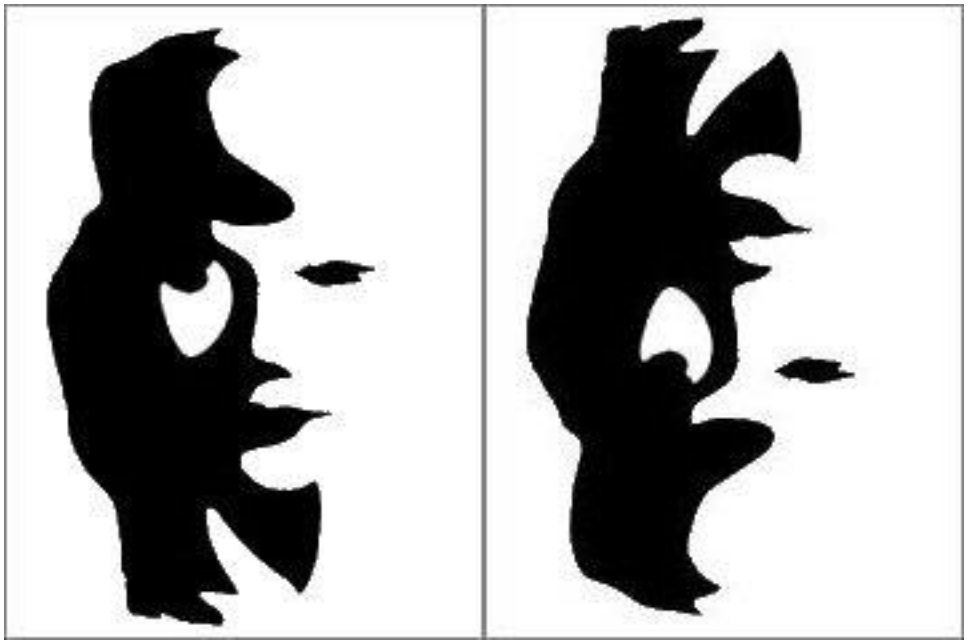
Société
Associations, Institutions, Religion, ...

Sports
Automobile, Football, Rugby, ...

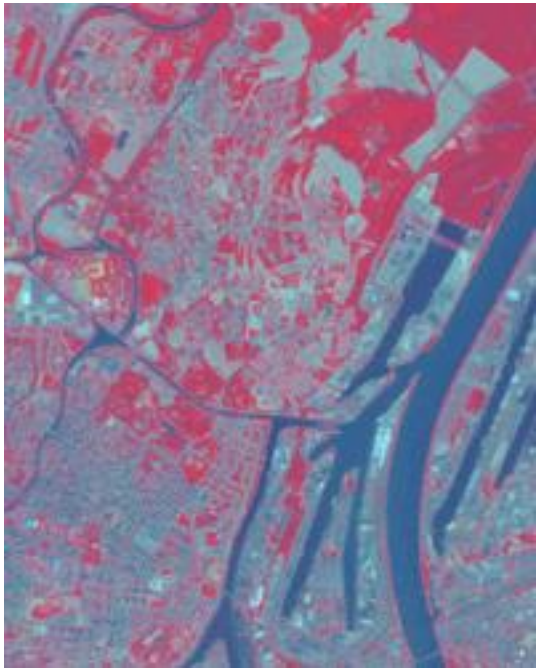
Catégorie apparentée :
Kids and Teens > International > Français (2736)

Références Biblio.

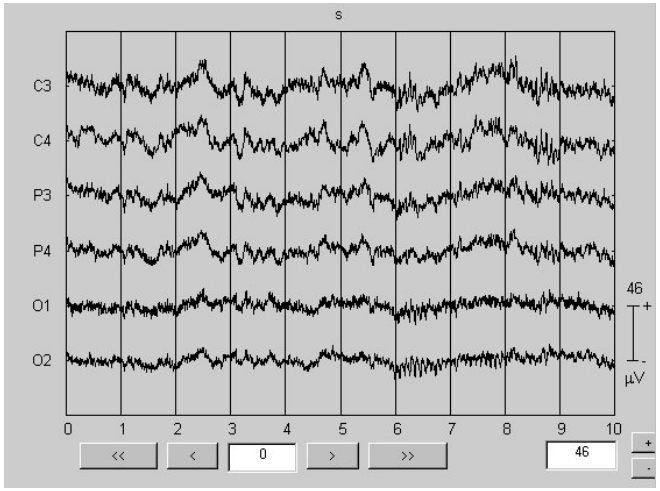
- Livres :
  - Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. (1989). Classification automatique des données. Dunod.
  - Duda R.O. and Hart P.E. (1973). Pattern classification and Scene Analysis. Wiley & sons.
  - Kaufman L., Rousseeuw P.J.(1990). Finding groups in data. Wiley & sons.
  - Lebart L., Morineau A., Piron M. (1995). Statistique exploratoire multidimensionnelle. Dunod.



Etudes de cas



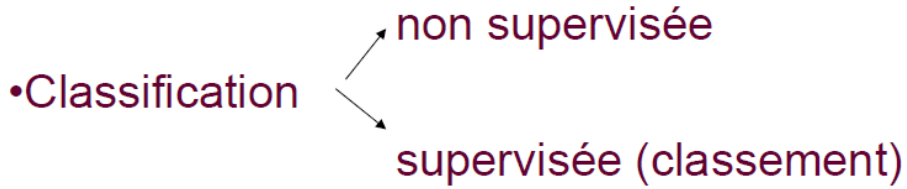
Etudes de cas



Des objectifs…. Des méthodes

- **Organiser** l'information
- Mettre ensemble dans une même classe les objets qui se **ressemblent**
- Obtenir des classes d'éléments formant une **partition** de l 'ensemble étudié.
- Associer à chaque classe un type **généralisant** les éléments de la classe.
- Opérer un **classement**

Les familles de méthodes



- On identifie des classes d'appartenance d'un objet à partir de certains traits descriptifs pour permettre une prise de décision automatique.

(en anglais *classification = classement*

cluster = classification)

•Paramétrique / non paramétrique: on suppose que les données suivent un modèle ou non

PLAN du cours

- 1- Introduction

  définitions, notion de système, applications

2- Processus de RdF

  pré-traitements, extraction de caractéristiques, représentation, classification, combinaison

3- Classification « statistique » des formes

  - théorie bayésienne de la décision

  - méthodes paramétriques ou non / bayésiennes ou non

  - arbres de décision

  - réseaux de neurones

Généralités 2

Mise en œuvre d'une classification :

Choix de la mesure d'éloignement (dissimilarité, distance) entre individus (généralement distance euclidienne)

Choix du critère d'homogénéité des classes à optimiser (généralement inertie).

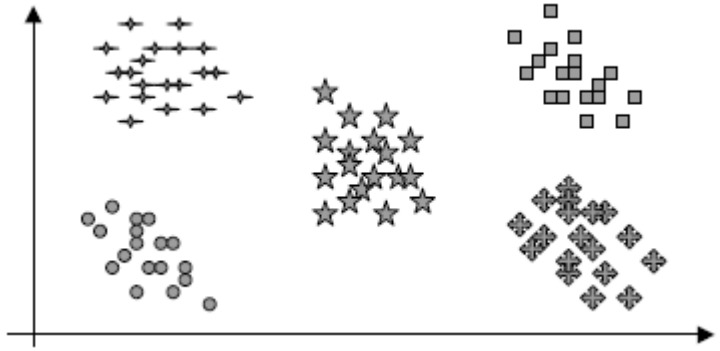
Choix de la méthode utilisée: la Classification Ascendante Hiérarchique (CAH) ou celle par réallocation dynamique sont les plus utilisées

Mesure de la qualité de la classification

Choix du nombre de classes et leur interprétation

Approche vectorielle

- Une forme est représentée par un vecteur de R^d (ensemble des *d* mesures effectuées sur la forme) ⇒ Une forme est un point dans l'espace de représentation R^d Contraintes: regroupement des classes dans des régions compactes de R^d



- Outils mathématiques utilisés pour l'approche vectorielle:

proba, stat, analyse de données, géométrie, topologie outils développés et « bien maîtrisés »

Analyses des données

OBJET DE LA STATISTIQUE

Le but de la statistique est de dégager les significations de données, numériques ou non, obtenues au cours de l'étude d'un phénomène.

Il faut distinguer les **données statistiques** qui sont les résultats d'observations recueillies lors de l'étude d'un phénomène, et la **méthode statistique** qui a pour objet l'étude rationnelle des données.

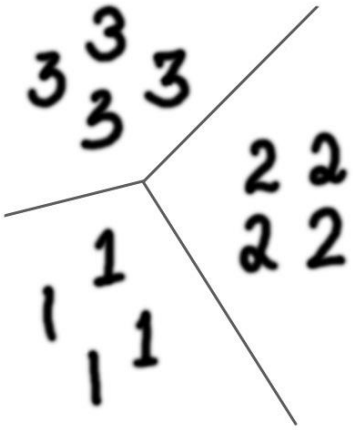
La méthode statistique comporte plusieurs étapes.

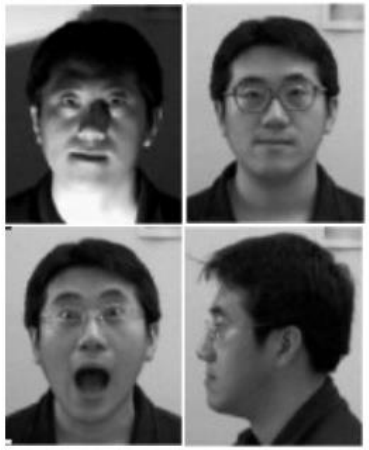
La statistique descriptive ou déductive.

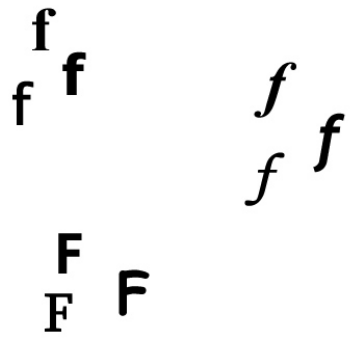
C'est l'ensemble des méthodes à partir desquelles on recueille, ordonne, réduit, et condense les données.

A cette fin, la statistique descriptive utilise des paramètres, ou synthétiseurs, des graphiques et des méthodes dites d'analyse des données (l'ordinateur a facilité le développement de ces méthodes).

exemples







Exemple

- On veut diviser la population en deux catégories selon la taille et la couleur de la peau.
- On trouvera pour chaque personne donc deux informations la taille en Cm et la couleur de la peau selon un type prédéfini et codifié (1 pour brun, 2 pour Blanc, …)
- En écriture vectorielle on aura donc

Personne (x1, x2)

Etudes sur les données

- Est-ce que ces données sont suffisantes pour représenter toute la population sujette à étude.
- Est-ce que il n'y a pas de redondance d'information, par exemple une étude statistique nous montre que tous les blanc sont de petite taille … etc
- Ne serait il pas judicieux de rajouter d'autres informations pour mieux reconnaitre les individus.

IA

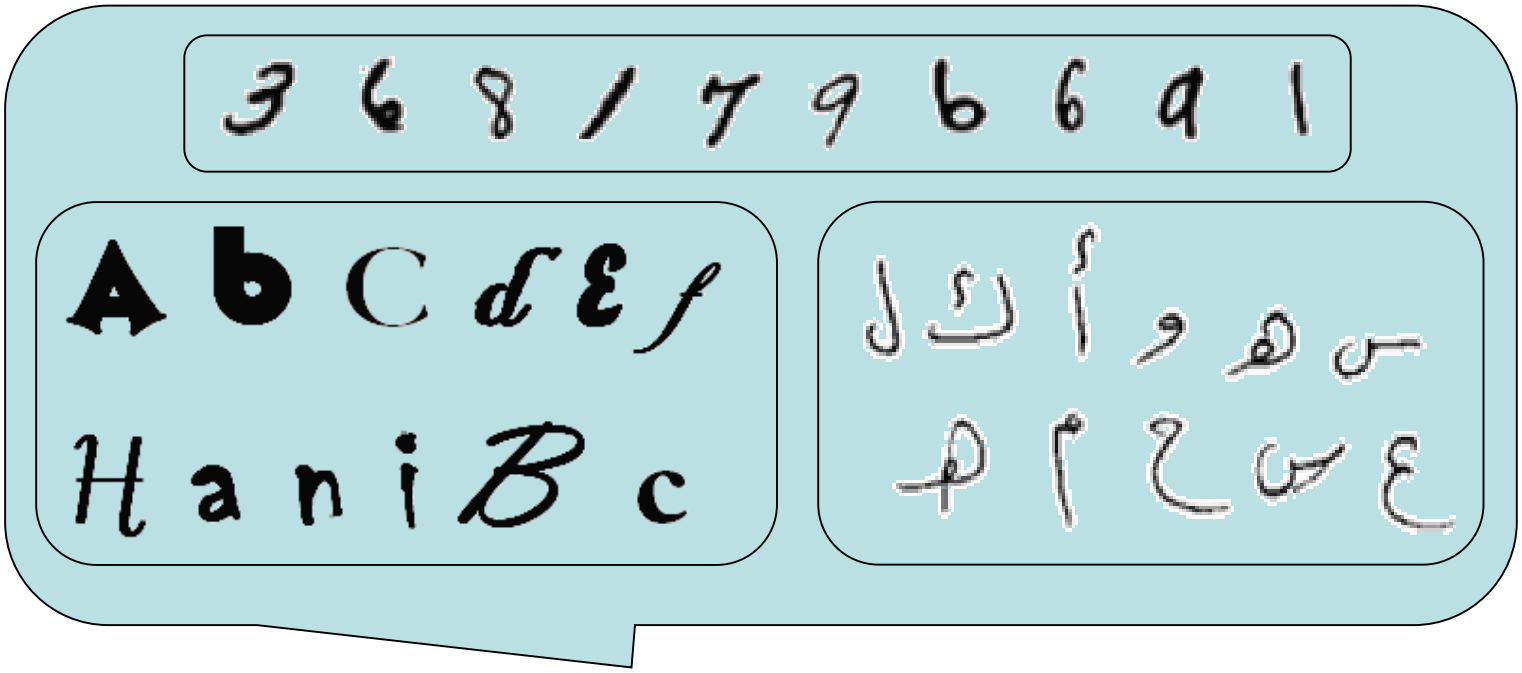
- «The Artificial Intelligence is the science of making machines do things that would require intelligence if done by humans » . *Marvin Minsky*
- Aristote et le processus de raisonnement correct, la logique
- – Ex: Socrate est un homme; tous les hommes sont mortels; donc Socrate est mortel.
- Qu'est ce que l'intelligence ?

=> Percevoir/Raisonner/Agir/Communiquer

Evaluation: capable de passer le test de Turing, jugée comme telle par l'homme.

RdF

Reconnaître les chiffres et les lettres isolés

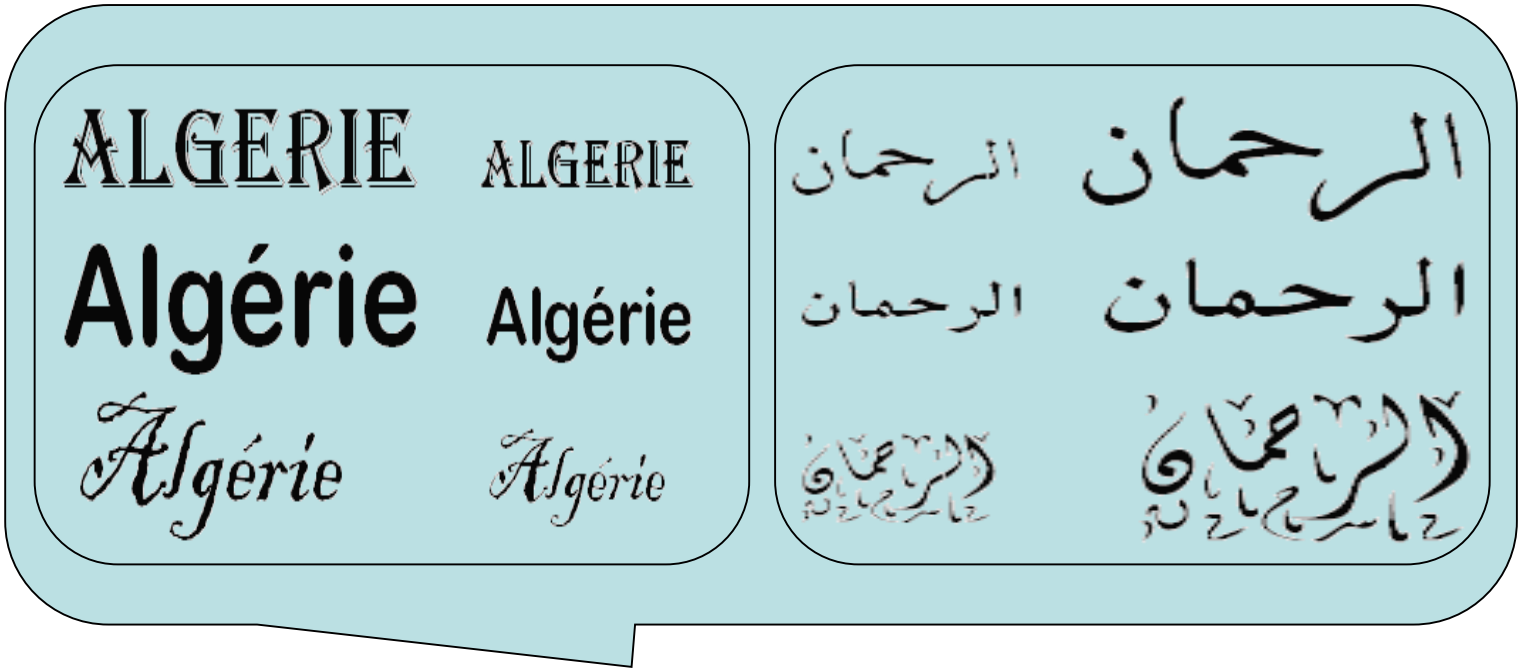


RdF

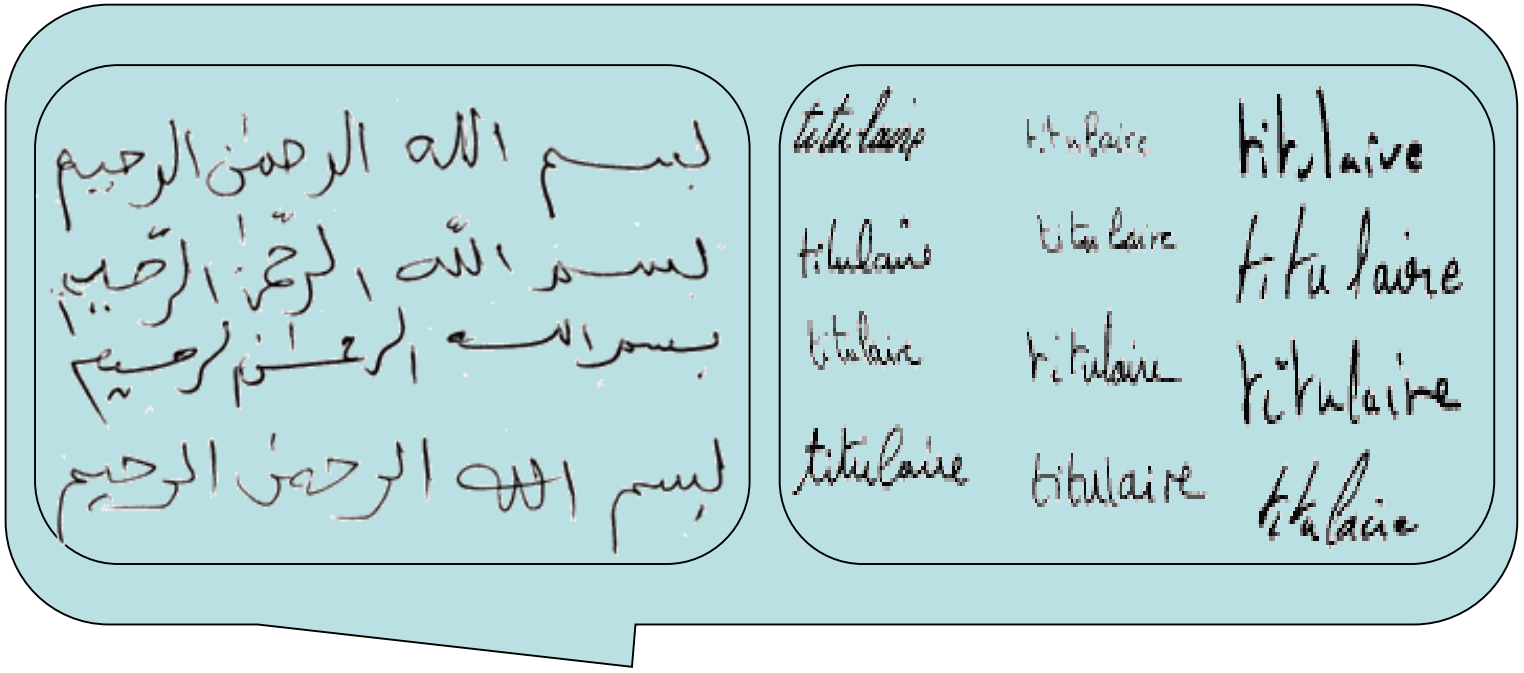
Applications

RdF

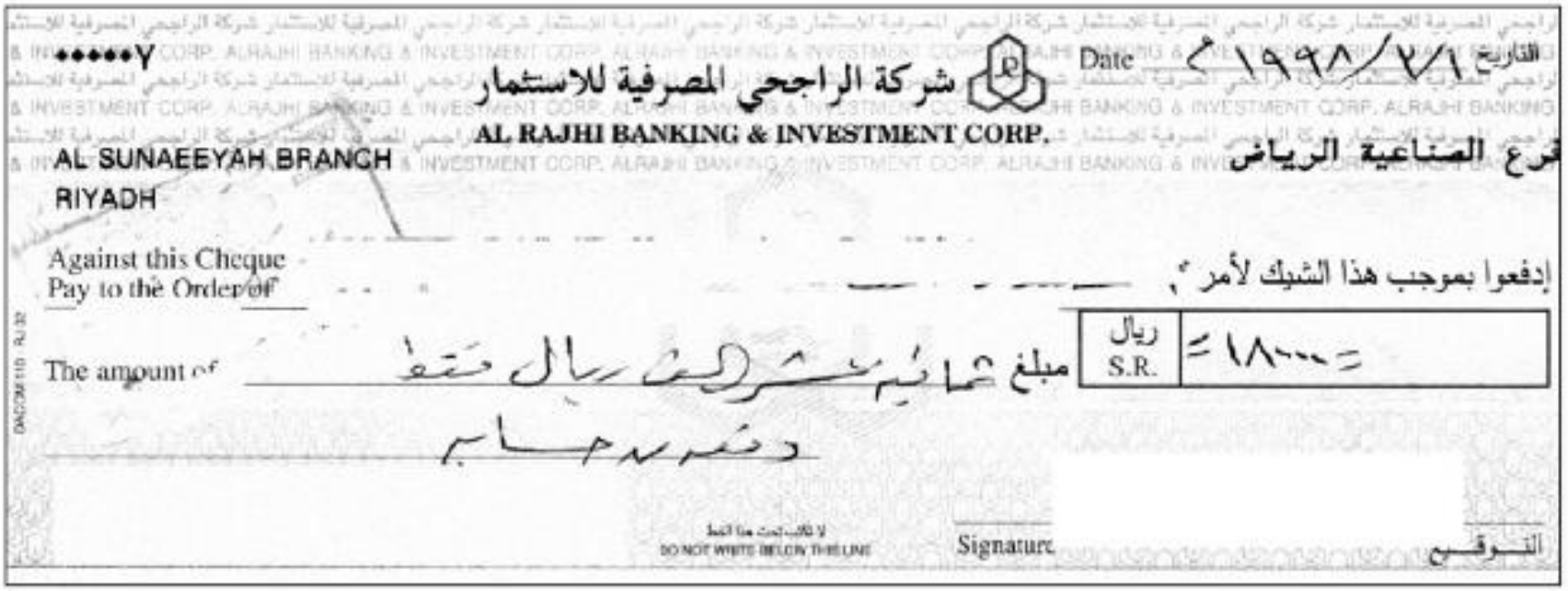
Reconnaître l'écriture imprimée (grande variation des fontes et des tailles)



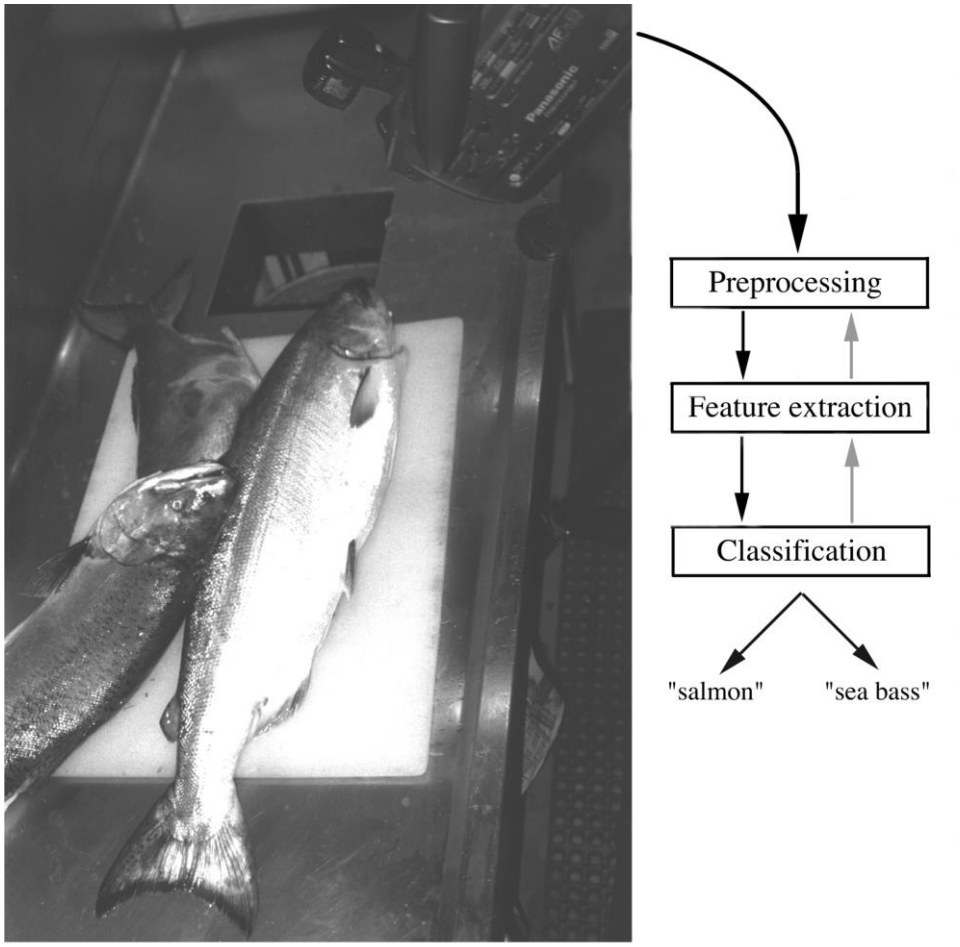
Reconnaître l’ écriture manuscrite (variation infinie des styles d’écriture)



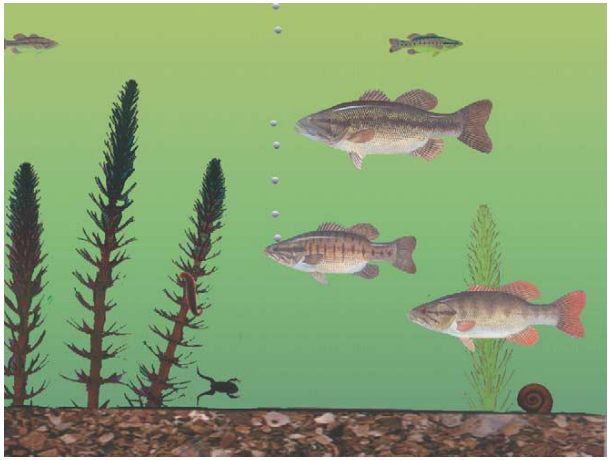
RdF

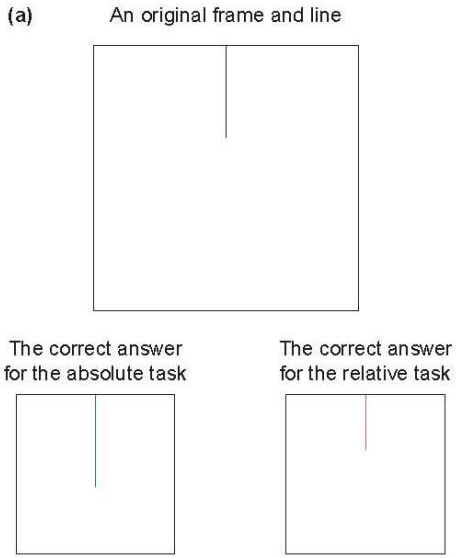

شركة الراجحي المصرفية للاستثمار
AL RAJHI BANKING & INVESTMENT CORP.
AL SUNAEEYAH BRANCH
RIYADH
فرع الصناعية الرياض
Date
التاريخ
Against this Cheque
Pay to the Order of
إدفعوا بموجب هذا الشيك لأمر
The amount of
مبلغ
ريال
S.R.
Signature
التوقيع
DO NOT WRITE BELOW THIS LINE

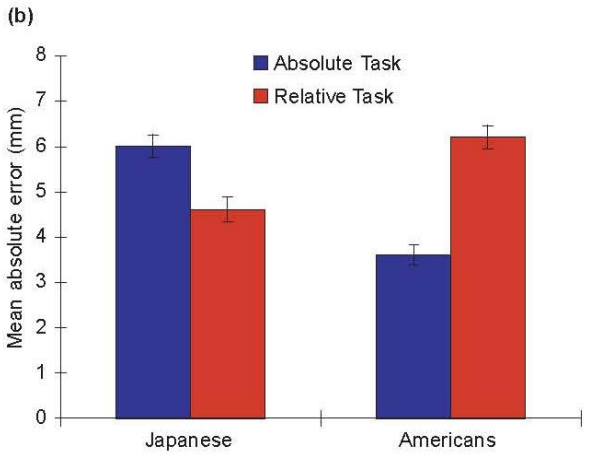
RdF

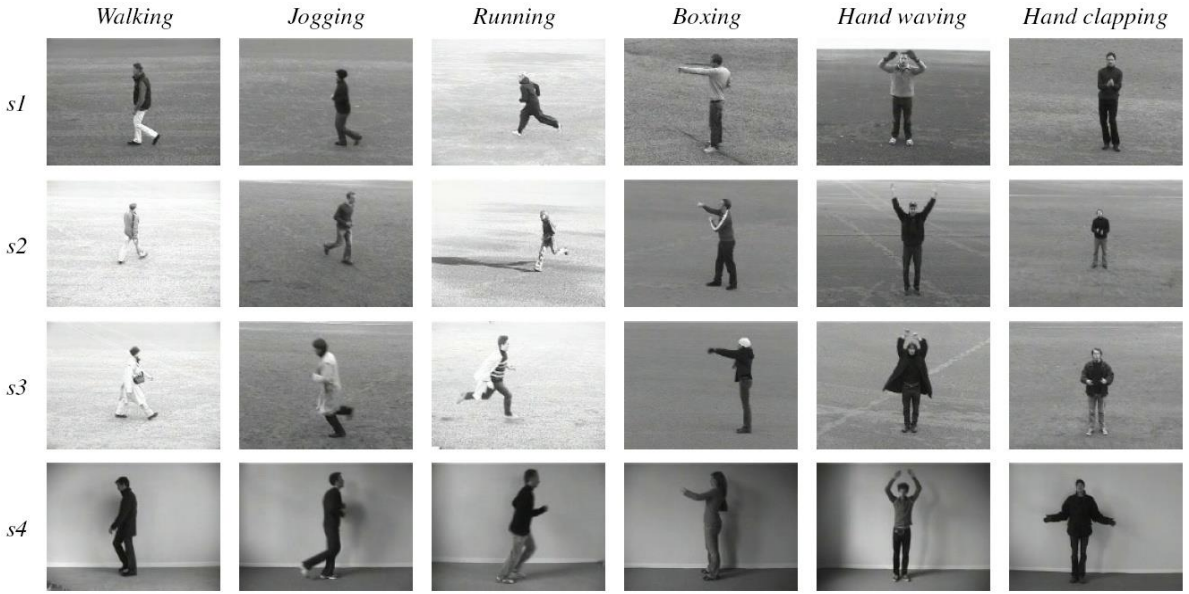

Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

Autres

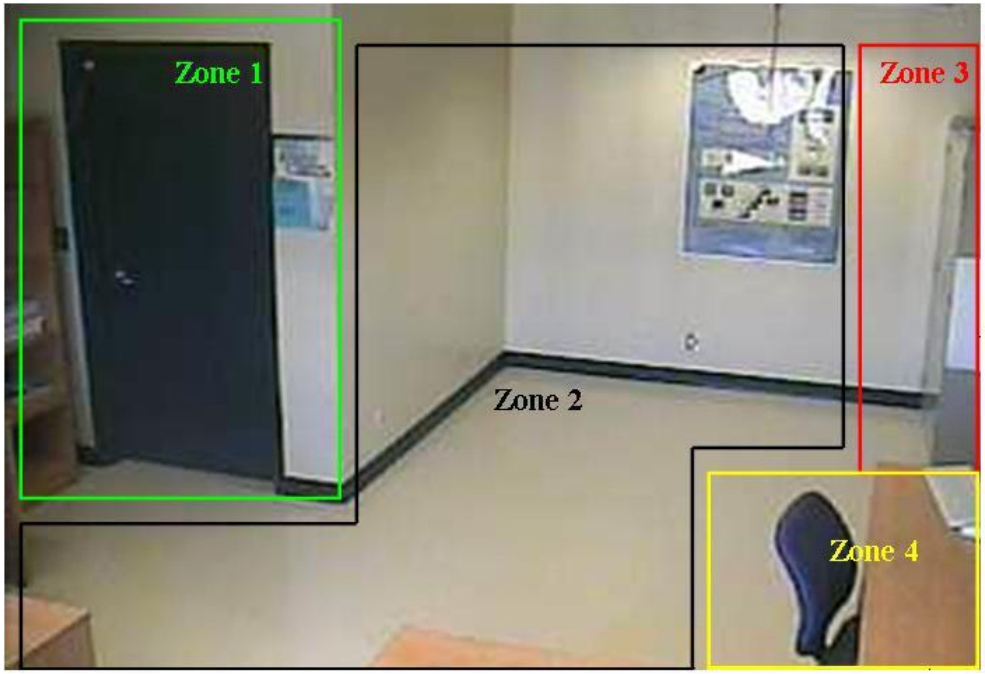



(a)
An original frame and line
The correct answer for the absolute task
The correct answer for the relative task

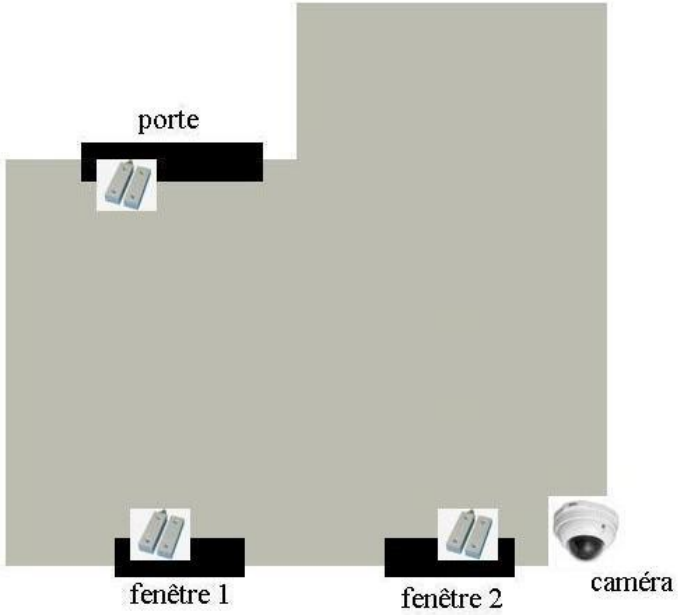

(b)
Absolute Task
Relative Task
Mean absolute error (mm)
0
1
2
3
4
5
6
7
8
Japanese
Americans


Walking
Jogging
Running
Boxing
Hand waving
Hand clapping
s1
s2
s3
s4

Autres



(a)



(b)

Apprentissage

L’apprentissage améliore les performances grâce à l'expérience.

Tous les animaux dotés d'un système nerveux central sont quasiment capables d'apprendre (même les plus simples).

Qu'est-ce que cela signifie apprendre pour un ordinateur? Pourquoi voudrions-nous lui faire apprendre?

Comment nous devons faire pour qu'ils apprennent?

Nous voulons que les ordinateurs apprennent quand il est trop difficile ou trop coûteux de programmer directement une tâche.

Laisser l'ordinateur se programmer lui-même en lui procurant des exemples d'entrées et de sorties.

En réalité: écrire un programme "paramétré", et laisser l'algorithme d'apprentissage trouver de lui-même l'ensemble des paramètres qui s'approchent le mieux de la fonction ou du comportement souhaité.

Caractéristiques

- La classification consiste à affecter un objet a un modèle.
- Il est nécessaire de choisir les meilleurs caractéristiques qui représentent le mieux nos objets.
- Ces caractéristiques peuvent être:
  - Des vecteurs de nombres réels
  - Une liste d'attributs
  - Des relations entre différentes parties

Sous Problèmes de classification

- Souvent dépendant de problèmes spécifiques

Donc :

Demande une connaissance ou une étude par le concepteur.

D’autres très réduits peuvent avoir un facteur de généralisation.

Cerveau et textes

Sleon une édtue de l'Uvinertisé de Cmabrigde, l'odrre des ltteers dnas un mtos n'a pas d'ipmrotncae, la suele coshe ipmrotnate est que la pmeirère et la drenèire soit à la bnnoe pclae. Le rsete peut êrte dnas un dsérorde ttoal et vuos puoevz tujoruos lrie snas porlblème. C'est prace que le creaveu hmauin ne lit pas chuaqe ltetre elle-mmêe, mias le mot cmome un tuot.

Extraction de caractéristiques

- Les caractéristiques idéales sont celles qui vont permettre au classifieur de faire un travail trivial.
- Les caractéristiques sont des informations dépendantes du domaines d'études (il n'y a pas de caractéristiques standards). Donc connaissance du domaine.
- Combiens de caractéristiques?
- Quelles sont les plus intéressantes?
- …..

Bruits

- Exemple des poissons
- L’intensité de la lumières
- Les ombres d’objets à proximité
- Qualité des media utilisée

- Le bruit complique la tache de classification.

Opérations Possibles

- Lissage
- Squelettisation
- Normalisation
- Corrections (autres)

Classification

- Nombre de classes inconnus
  - ❑ Clustering

- Classes connues dès le départ
  - ❑ Classification

Proximité

Comment mesurer la distance entre 2 points d(x1; x2) ?

- distance euclidienne :

  $d^2(x1; x2) = \Sigma_i (x_{1i} - x_{2i})^2 = (x_1 - x_2)(x_1 - x_2)'$

- distance de Manhattan :

  $d(x_1; x_2) = \Sigma_i |x_{1i} - x_{2i}|$

- distance de Sebestyen :

  $d^2(x_1; x_2) = (x_1 - x_2)W(x_1 - x_2)'$

  (W= matrice diagonale de pondération)

- distance de Mahalanobis :

  $d^2(x_1; x_2) = (x_1 - x_2)C^{-1}(x_1 - x_2)'$

  (C=covariance)

Des objectifs…. Des méthodes

- **Organiser** l'information
- Mettre ensemble dans une même classe les objets qui se **ressemblent**
- Obtenir des classes d'éléments formant une **partition** de l 'ensemble étudié.
- Associer à chaque classe un type **généralisant** les éléments de la classe.
- Opérer un **classement**
- …
- Autant de méthodes que de données et d'objectifs

Def

- Caractéristiques : signes ou ensembles de signes distinctifs. (En anglais : Feature).
- Objet dans le monde est représenté par une ensemble de propriétés (noumènes)

  { x1, x2 ... xn}.

Bruits

- La formation des vrais objets physiques est sujette aux influences aléatoires.
- Les propriétés d'un objet sont, donc, les valeurs aléatoires. On peut résumer ceci par une somme d'une forme "intrinsèque" X_i plus ces influences aléatoires individuelles, $\mathbf{B}_i$.
- $\mathbf{X} = X_i + \mathbf{B_i}$

Observations

- Les propriétés sont observées au travers des capteurs.
- Ceci donne une observation (un phénomène) sous forme d'une ensemble de caractéristiques : { y_1, y_2 ... y_n}.

$$\vec{\mathbf{Y}} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

Les observations sont corrompues par un bruit, $\mathbf{B_o}$

$\mathbf{Y = X + B_o}$

- Le bruit est, par définition, imprévisible. Il est aléatoire.
- Donc les caractéristiques observées sont des vecteurs aléatoires :
- La corruption des observations par un bruit aléatoire est fondamentale aux capteurs physiques.

La classification

La classification est une capacité fondamentale de l'intelligence.

Comprendre : Faire entrer dans une catégorie.

Les perceptions brutes (les phénomènes) sont comprise par l'association aux catégories mentales (les concepts).

La capacité de classer les phénomènes est caractéristique à toute espèce vivante.

Reconnaissance : Le fait de reconnaître, d'identifier un objet, un être comme tel.

Reconnaître : Saisir un objet par la pensée, en reliant entre elles, des images, des perceptions. Identifier par la mémoire, le jugement ou l'action.

1. Penser un objet présent comme ayant déjà été saisi par la pensée.
2. Juger un objet ou un concept comme compris dans une catégorie.

Identifier : Reconnaître un individu

Classer : Reconnaître un membre d'une catégorie, ou d'une classe.

Classe: n. f. Ensemble d'individus ou d'objets qui ont des caractères communs.

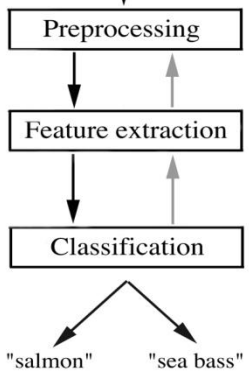
ensemble n. m. : un groupe.

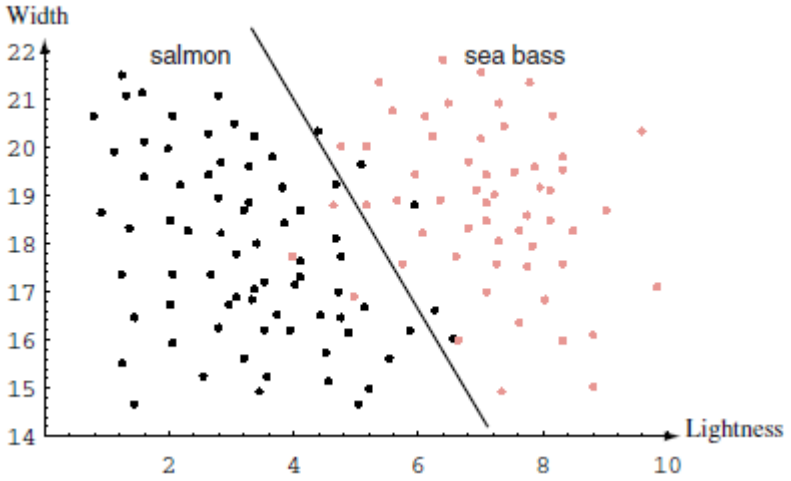
Les ensembles peuvent êtres définis par extension : une liste complète des membres

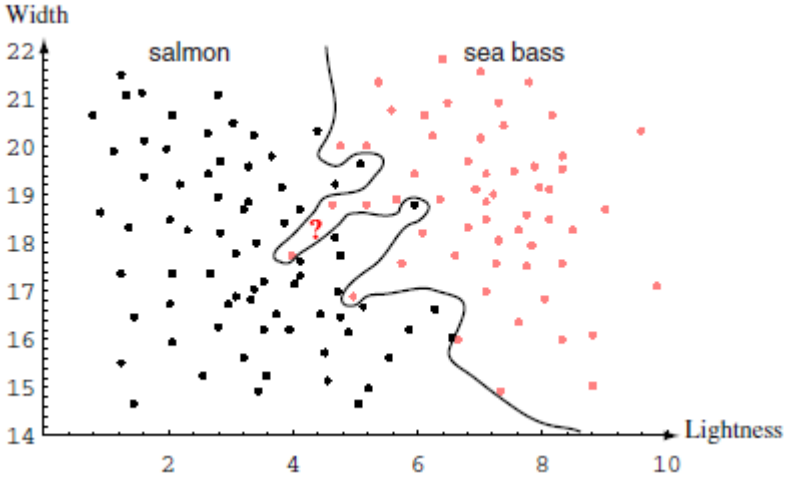
intention : une conjonction des caractéristiques.

Un ensemble est défini par un test d'appartenance.

Exemple


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"


Width
22
21
20
19
18
17
16
15
14
salmon
sea bass
2
4
6
8
10
Lightness


Width
22
21
20
19
18
17
16
15
14
salmon
sea bass
?
2
4
6
8
10
Lightness

Paramétrique /non paramétrique

Par extension : Une comparaison d'une observation avec des membres connus de l'ensemble (des prototypes).

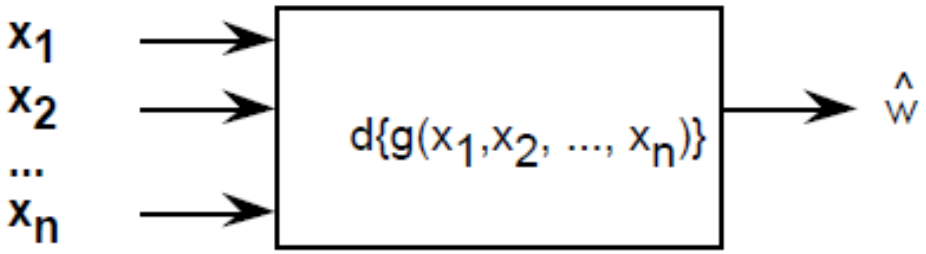
Par intention : Conjonction de prédicats définis sur les propriétés observées

Ceci correspond (grosso modo) aux deux approches de la reconnaissance statistique :

les techniques de classification paramétriques (par intention) et non-paramétriques (par extension).

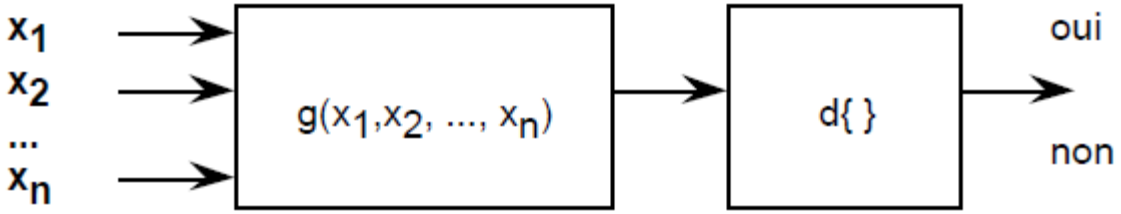
La classification est un processus d'association.

Classification



Pour un vecteur de caractéristique il sort une estimation de la classe, w.
Dans sa forme la plus réduit, on peut voir la classification comme un "test d'appartenance" d'un vecteur de caractéristique **X** à une classe w.

Processus classification



Le processus est composé de deux étapes : une fonction de discrimination, g(), est une fonction de décision, d{}.

{oui, non} ¬ d{ g(**X**) }

d{} est un prédicat. Son résultat est issu de l'ensemble {oui, non}.

g() est une fonction de discrimination évaluée sur **X.**

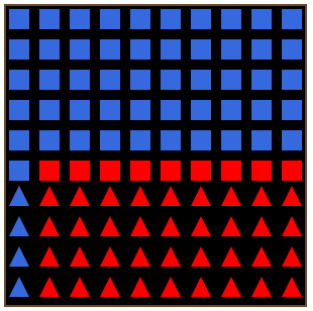
Dans tous les cas, la classification se résume à une division de l'espace de caractéristiques en partitions disjointes.

Bayes

- Cette division peut-être fait par estimation de fonctions paramétrique ou par une liste exhaustives des frontières.
- Le critère de division est la probabilité de commettre une erreur.
- Cette probabilité est fournie par la règle de Bayes.

$$p(\text{Classe}_k \mid \vec{\mathbf{X}}) = \frac{p(\vec{\mathbf{X}} \mid \text{Classe}_k)\, p(\text{Classe}_k)}{p(\vec{\mathbf{X}})}$$

Exemple 1



une urne qui contient 100 objets, pouvant avoir deux formes (carré ou triangle) et deux couleurs (bleu ou rouge). La composition détaillée de l'urne est donnée sur le dessin ci-contre.

Une main tire un objet au hasard, quel est la probabilité que ce soit un carré ? Facile ! Il y a 100 objets, 60 sont des carrés, donc la réponse est 60%.

Imaginez maintenant que la main tire un objet, et que vous parveniez à distinguer rapidement que cet objet est rouge. Quel est la probabilité que ce soit un carré ? Facile aussi, il y a 45 objets rouges, dont 9 qui sont à la fois rouges et carrés, la « probabilité d'être un carré sachant qu'il est rouge » est donc 9/45 = 20%.

Si vous comparez ces deux situations, vous constatez que la probabilité que l'objet soit un carré est fortement affectée par le fait de savoir qu'il est rouge. La « probabilité que l'objet soit carré » n'est pas la même que la « probabilité que l'objet soit carré sachant qu'il est rouge ».

Les mathématiciens parlent de probabilités « conditionnelles », et utilisent la barre verticale | pour symboliser l'expression « sachant que ». Dans les exemples précédents, on a donc

P(Carré) = 60%

P(Carré | Rouge) = 20%

faisons le calcul inverse. Vous tirez un objet les yeux bandés, vous sentez dans votre main qu'il est carré : quel est la probabilité qu'il soit rouge ? Si vous regardez attentivement la composition de l'urne, il y a 60 objets carrés, dont 9 qui sont rouges, donc

P(Rouge | Carré) = 9/60 = 15%

Une leçon importante dans cette affaire, c'est que P(Rouge | Carré) n'est pas la même chose que P(Carré | Rouge).

Suite exemple

Si vous reprenez le calcul précédent, dans le cas où l'on sait que l'objet est rouge, on calcule la probabilité d'être carré en divisant le nombre d'objets carrés et rouges par le nombre total d'objets rouges. On a donc

P(Carré | Rouge) = P(Carré et Rouge) / P(Rouge)

Mais on peut également permuter les rôles des formes et des couleurs, et écrire

P(Rouge| Carré) = P(Carré et Rouge) / P(Carré)

En regroupant les deux formules et en éliminant P(Carré et Rouge), on a

P(Carré | Rouge) = P(Rouge| Carré)P(Carré)/P(Rouge)

Thomas Bayes avait observé que cette formule est vraie en général, et pas seulement pour notre problème d'objets carrés et rouges dans une urne. La célèbre formule de Bayes s'écrit sous sa forme abstraite

Bayes 2

Imaginons deux urnes remplies de boules. La première contient dix (10) boules noires et trente (30) blanches ; la seconde en a vingt (20) de chaque. On tire sans préférence particulière des urnes au hasard et dans cette urne, on tire une boule au hasard. La boule est blanche. Quelle est la probabilité qu'on ait tiré cette boule dans la première urne sachant qu'elle est blanche ?

Bayes 3

- Soit H_1 l'hypothèse « On tire dans la première urne. » et H_2 l'hypothèse « On tire dans la seconde urne. ». Comme on tire sans préférence particulière, $P(H_1) = P(H_2)$; de plus, comme on a certainement tiré dans une des deux urnes, la somme des deux probabilités vaut 1 : chacune vaut 50 %.
- Notons D l'information donnée « On tire une boule blanche. » Comme on tire une boule au hasard dans une des urnes, la probabilité de D sachant l'hypothèse H_1 réalisée vaut :

$$P(D|H_1) = \frac{30}{40} = 75\,\%$$

Bayes 4

De même la probabilité de *D* sachant l'hypothèse H_2 réalisée vaut :

$$P(D|H_2) = \frac{20}{40} = 50\,\%$$

La formule de Bayes dans le cas discret nous donne donc.

$$\begin{aligned} P(H_1|D) &= \frac{P(H_1) \cdot P(D|H_1)}{P(H_1) \cdot P(D|H_1) + P(H_2) \cdot P(D|H_2)} \\ &= \frac{50\% \cdot 75\%}{50\% \cdot 75\% + 50\% \cdot 50\%} \\ &= 60\% \end{aligned}$$

Problèmes de classification

1. Comprendre et analyser les objectifs de l'application
2. Créer (utiliser) une base de données pour la mise au point de l'application.
3. Prétraitement et nettoyage des données
4. Analyse statistique des données (réduction de la dimension, projection, etc...)
5. Identifier le type de problèmes (discrimination, clustering, etc...) et choisir un algorithme.
6. Evaluer les performances de l'algorithme.
7. Réitérer les étapes précédentes si nécessaire.
8. Déployer l'application.

Types d’apprentissage

Sur l’ensemble des données on doit pouvoir écrire des algorithmes qui puissent classifier les informations a traiter. Pour cela on doit doter nos algorithmes d’une faculté (ou fonction) pour qu’ils apprennent à distinguer les catégories entre elles.

Types d’apprentissage

- Apprentissage supervisé
- Apprentissage non-supervisé
- Apprentissage semi-supervisé

Apprentissage supervisé

• Objectifs : à partir d'un ensemble d'observations $\{x_1, \cdot \cdot \cdot, x_n\} \in X^d$ et de mesures $\{y_i\} \in Y$, on cherche à estimer les dépendances entre l'ensemble X et Y.

Exemple : on cherche à estimer les liens entre les habitudes alimentaires et le risque d'infarctus. x_i est un patient décrit par d caractéristiques concernant son régime et y_i une catégorie (risque, pas risque).

On parle d'apprentissage *supervisé* car les y_i permettent de guider le processus d'estimation

• Exemples de méthodes : Méthode du plus proche voisin, réseaux de neurones, Séparateurs à Vastes Marges, Logique floue etc..

• Exemples d'applications : détection de fraude, marketing téléphonique, changement d'opérateurs téléphonique etc...

Apprentissage non-supervisé

• Objectifs : Comme seules les observations $\{x_1, \cdots, x_n\} \in X^d$ sont disponibles, l'objectif est de décrire comment les données sont organisées et d'en extraire des sous-ensemble homogènes.

Exemple : On cherche à étudier le panier de la ménagère dans une certaine zone démographique en fonction de certains critères sociaux. X représente un individu à travers ses caractéristiques sociales et ses habitudes lors des courses

• Exemples de méthodes : Classification hiérarchique, Carte de Kohonen, K-means, extractions de règles...

• Exemples d'applications : identification de segments de marchés, identification de document similaires.

Apprentissage semi-supervisé

- Objectifs : parmi les observations $\{x_1, \cdots, x_n\} \in X^d$, seulement un petit nombre d'entre elles ont un label $\{y_i\}$. L'objectif est le même que pour l'apprentissage supervisé mais on aimerait tirer profit des observations non labélisées.
- Exemple : pour la discrimination de pages Web, le nombre d'exemples peut être très grand mais leur associer un label est coûteux.
- Exemples de méthodes : méthodes bayesiennes, Séparateur à Vastes Marges, etc...

Théorie Bayésienne de la Décision

Notations:

$w \in \Omega$: ensemble des classes à reconnaître

$x \in R^N$: espace de représentation

$p(x)$: probabilité d'observer x

$p(w/x)$: probabilité d'appartenance à la classe w conditionnellement au fait d'être en x

$d : R^N \rightarrow \Omega$: règle de décision

$x \rightarrow d(x)=w_k$ (avec $\int_R{}^N p(x)dx = 1$ et $\forall x, \Sigma_k\ p(w_k/x) = 1$)

Règle de décision bayésienne

Règle de décision bayésienne (d1):

fonction qui à chaque objet x associe la classe la plus probable au point x:

$d1(x)=w \Rightarrow \forall w' \in \Omega, p(w/x) \geq p(w'/x)$

Remarque: si les caractéristiques sont suffisamment descriptives alors:

$p(w/x)=1$ ou $p(w/x)=0$

Théorie Bayésienne de la Décision

- Propriété: La fonction de décision bayésienne minimise la probabilité globale d'erreur du système

Démo: La probabilité d'erreur d'une règle est :

$Err(d) = \int_{RN} [1-p(d(x)/x)]p(x)dx$ et $\forall x, \forall w, p(d1(x)/x) \geq p(d(x)/x)$

ainsi : $Err(d) - Err(d1) = \int_{RN} [p(d1(x)/x)-p(d(x)/x)]p(x)dx$

d'où : $Err(d) \geq Err(d1), \forall d$

Règle: A chaque point x, associer la classe dont la densité de probabilité en x est la plus forte (maximiser la probabilité de ce qu'on observe)

Théorie Bayésienne de la Décision

- Théorème de Bayes:

Utilité: Expérience (apprentissage) $\Rightarrow$ p(x/w) ou p(x,w)

Décision $\Leftarrow$ p(w/x)

Enoncé: p(w/x) = [p(x/w)p(w)] / p(x)

car: p(x,w)=p(x/w)p(w)=p(w/x)p(x)

Exemple: n=100 observations

5 occurrences possibles pour x: x1, x2, …, x5

4 classes possibles: w1, w2, w3, w4

Suite

	x1	x2	x3	x4	x5
w1	2	20	0	1	0
w2	0	2	0	12	13
w3	6	4	8	0	2
w4	20	0	8	0	2

p(w1)=23/100; p(x3)=16/100; p(x5,w4)=2/100
p(w1/x1)=2/28; p(w2/x1)=0;
p(w3/x1)=6/28; p(w4/x1)=20/28;
⇒ si j'observe x1, je décide w4

Probabilités

- A priori

Si on a deux classes w_1 et w_2

$P(w_1)+p(w_2)=1$

Décision en utilisant que les informations a priori

On décide w_1 *si* $P(\omega_1) > P(\omega_2)$ *sinon on* décide w_2

Formule de Bayes

$$P(\omega_j \mid x) = \frac{p(x/\omega_j)P(\omega_j)}{p(x)}$$

Dans le cas de deux catégories

$$p(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j)P(\omega_j)$$

Décision de Bayes

X est une observation telle que :

Si $P(\omega_1 | x) > P(\omega_2 | x)$ vraie Alors X= ω_1

Si $P(\omega_1 | x) < P(\omega_2 | x)$ vraie Alors X = ω_2

Donc: Quand on observe un x particulier, la probabilité d'erreur est :

$P(erreur \mid x) = P(\omega_1 | x)$ *Si on décide* ω_2

$P(erreur \mid x) = P(\omega_2 | x)$ *Si on décide* ω_1

La règle de décision de Bayes minimise l'erreur

- On decide ω_1 *si* $P(\omega_1 | x) > P(\omega_2 | x)$;

  Sinon on décide ω_2

Donc !

$P(erreur \mid x) = min$ [$P(\omega_1 | x), P(\omega_2 | x)$]

(Décision de Bayes)

Perte

- Permettre des actions autres que la classification permettre surtout la possibilité de rejet
- Refuser de prendre une décision dans les cas proches ou mauvais!

Comment !

Définition fonction perte

Soit $\{\omega_1, \omega_2, \ldots, \omega_c\}$ un ensemble de c classes

Soit $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ un ensemble d' actions possible

soit $\lambda(\alpha_i/\omega_j)$ la perte quand on fait une action α_i quand la catégorie est ω_j

Risque Global

R = somme de tout $R(\alpha_i | x)$ *pour* $i = 1,\ldots,a$

Minimiser R reviend à Minimiser $R(\alpha_i|x)$ *for* $i = 1,\ldots, a$

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

Choisir l'action α_i pour laquelle $R(\alpha_i | x)$ *est minimum*
Quant R est minimum R est appelé Risque

Bayes risk = la meilleure performance qu'on peut atteindre

Classification à deux classes

α_1: *action pour décider* ω_1

α_2: *action pour décider* ω_2

$\lambda_{ij} = \lambda(\alpha_i / \omega_j)$

Perte encourue quand on décide ω_i quand la vraie catégorie est ω_j

Risque Conditionnel:

$R(\alpha_1 / x) = \lambda_{11} P(\omega_1 / x) + \lambda_{12} P(\omega_2 / x)$

$R(\alpha_2 / x) = \lambda_{21} P(\omega_1 / x) + \lambda_{22} P(\omega_2 / x)$

Risque Minimum

La règle est la suivante:

Si $R(\alpha_1|x) < R(\alpha_2|x)$

alors on prend l'action α_1: "on décide ω_1"

Formule équivalente à la règle:

On décide ω_1 *si*:

$(\lambda_{21}-\lambda_{11})\, P(x|\omega_1)\, P(\omega_1) > (\lambda_{12}-\lambda_{22})\, P(x|\omega_2)\, P(\omega_2)$

On décide ω_2 *sinon*

Vraisemblance de la règle de décision

- La règle précédente est équivalente à :

$$if \ \frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

- Alors on prend l'action α_1 *(on décide* w_1*)*
- *Sinon on prend l'action* α_2 *(on décide* w_2*)*

Méthodes Paramétriques / Non Paramétriques

Problème du modèle de base: difficile de mesurer les $p(x/w_i)$!!!

2 types d'approches suivant les hypothèses retenues:

- méthodes paramétriques:

on fait appel à une famille paramétrable de densités ou de surfaces et on optimise le(s) paramètre(s) pour minimiser la probabilité d'erreur

- méthodes non paramétriques:

on développe un algorithme qui converge vers la densité ou la surface idéale quelle qu'elle soit (moyennant certaines hypothèses)

Suite

	Méthodes bayésiennes	Méthodes non bayésiennes
Méthodes paramétriques	Estimation de gaussiennes	Séparation linéaire
Méthodes non paramétriques	Fenêtres de Parzen	K plus proches voisins

Estimation de Gaussiennes

- Méthode paramétrique bayésienne

$p(x/w_k) = (2\pi)^{-d/2}(\det\Sigma_k)^{-1/2}\exp[-0.5(x-m_k)^t \Sigma_k^{-1}(x-m_k)]$

où: d dimension de l'espace de représentation

Σ_k matrice de variance-covariance de la classe w_k

m_k centre (moyenne) de la classe w_k

Apprentissage: estimer pour chaque classe w_k : m_k et Σ_k

Estimateur du Maximum de Vraisemblance

$m_k = (1/n_k)\Sigma\, x_i$

$\Sigma_k = (1/n_k\text{-}1)\Sigma(x_i-m_k)(x_i-m_k)^t$ (non biaisé)

Décision: appliquer la règle de Bayes

Choisir w_k qui maximise $p(w_k/x)=p(x/w_k)p(w_k)/p(x)$