From Data Science to Machine Learning

De l'analyse de données à l'apprentissage automatique

Jean-Yves Ramel

ramel@univ-tours.fr

Laboratoire d'Informatique Fondamentale
et Appliquée de Tours

LABORATOIRE D'INFORMATIQUE FONDAMENTALE ET APPLIQUÉE DE TOURS

Science / Analyse des données

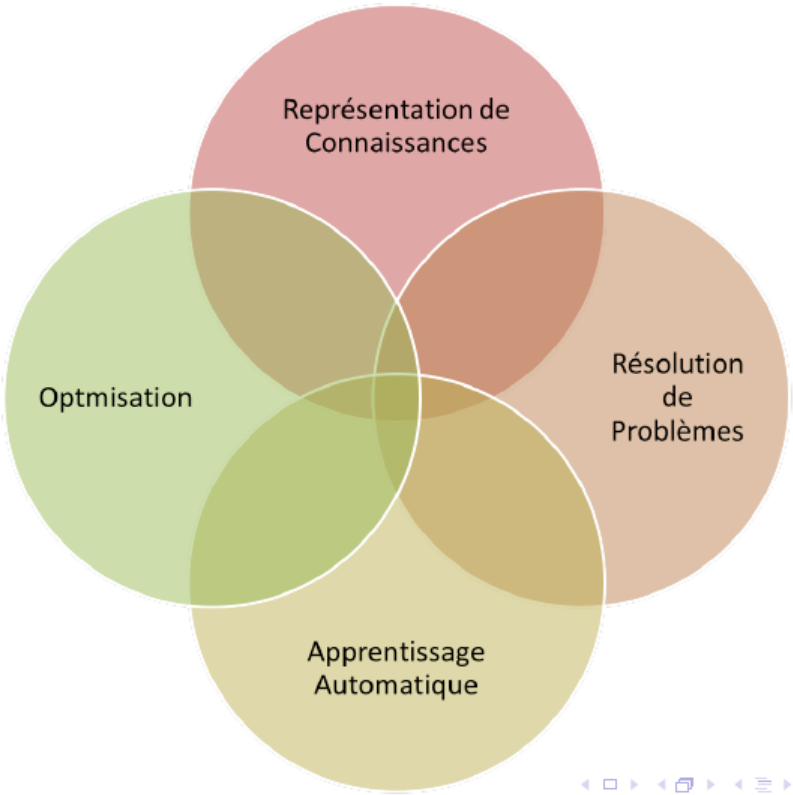
Data driven science : le 4e paradigme (Jim Gray - Prix Turing)

Extrait : "A l'heure actuelle, la science vit une révolution qui conduit à nouveau paradigme selon lequel 'la science est dans les données', autrement dit la connaissance émerge du traitement des données [...] **Le traitement de données et la gestion de connaissances représentent ainsi le quatrième pilier de la science après la théorie, l'expérimentation et la simulation.** L'extraction de connaissances à partir de grands volumes de données (en particulier quand le nombre de données est bien plus grand que la taille de l'échantillon) , l'apprentissage statistique, l'agrégation de données hétérogènes, la visualisation et la navigation dans de grands espaces de données et de connaissances sont autant d'instruments qui permettent d'observer des phénomènes, de valider des hypothèses, d'élaborer de nouveaux modèles ou de prendre des décisions en situation critique"

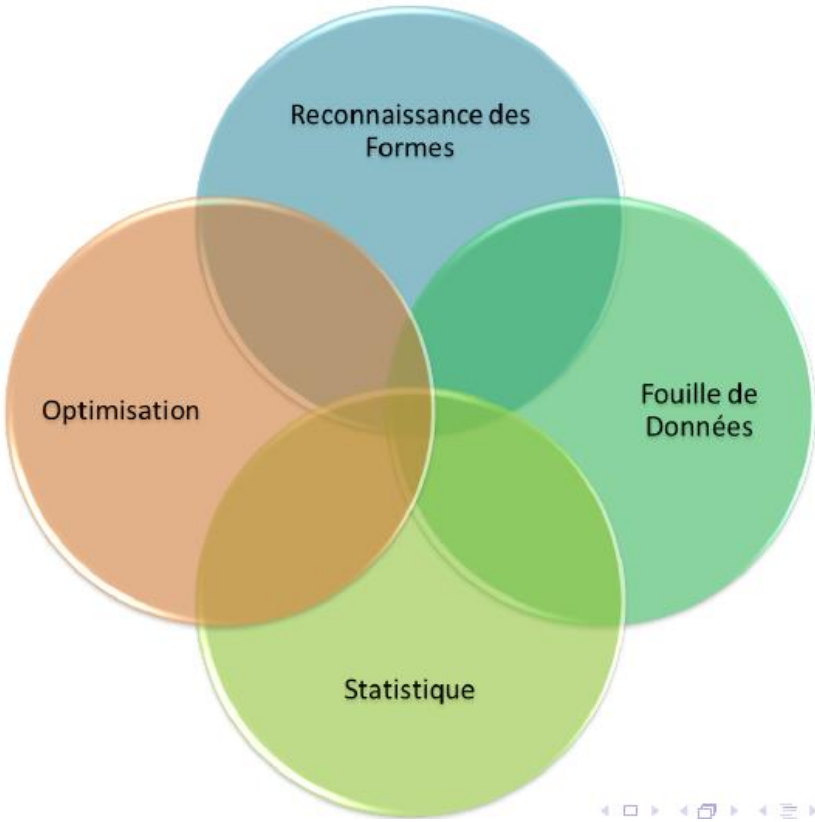
Science / Analyse des données

- L'apprentissage automatique fait largement appel à des outils et des concepts de la statistique, et fait partir de disciplines plus vastes appelées science des données et IA.

Intelligence Artificielle



Apprentissage Automatique



Apprentissage automatique

Définition initiale

- Learning is making useful changes in mind - [Marvin Minsky, 1985]
- Learning is any change in a system that allows it to perform better the second time on repetition of the same task or another task drawn from the same population - [Herbert Simon, 1983]
- Learning is the organization of experience - [Scott, 1983]
- Learning is constructing or modifying representations of what is being experienced - [Riszard Michalski, 1986]

Problématique

- Nous souhaitons avoir des ordinateurs / **agents intelligents**, adaptatifs, avec un comportement robuste
- Programmer de tels comportement est souvent impossible

Solution

- Faire un ordinateur capable de se programmer lui-meme
  - a partir d'exemples (apprentissage classique / par imitation)
  - a partir de son "expérience" (apprentissage par renforcement)

Apprentissage automatique

Autre définition

- Branche de l'IA qui concerne le développement d'algorithmes permettant de rendre une machine (un agent) capable d'accomplir des tâches complexes sans avoir été explicitement programmée dans ce but.

- Exemple : comment écrire un programme qui reconnaisse les caractères manuscrits ?
  - Entrer des règles manuellement (difficile et peu fiable)
  - Meilleur méthode : écrire un algorithme (générique) qui produit automatiquement un programme de reconnaissance de caractères à partir d'un grand nombre d'exemples.

1 – Des données aux bases d’apprentissage

1 – Des données aux bases d'apprentissage

Individus, objets VS variables, descripteurs

- **Population**
  groupe ou ensemble d'individus que l'on analyse.

- **Recensement**
  étude de tous les individus d'une population donnée.

- **Sondage**
  étude d'une partie seulement d'une population appelée échantillon.

- **Variables**
  ensemble de caractéristiques d'une population.
  - **quantitatives**: nombres sur lesquels les opérations usuelles (somme, moyenne,...) ont un sens ; elles peuvent ^être discrètes (ex : nombre d'éléments dans un ensemble) ou continues (ex: prix, taille) ;
  - **qualitatives**: appartenance a une catégorie donnée ; elles peuvent être nominales (ex : sexe, CSP) ou ordinales quand les catégories sont ordonnées (ex : très résistant, assez résistant, peu résistant).

1 – Des données aux bases d'apprentissage

Description de données quantitatives

- **Définition**
  On appelle variable un vecteur x de taille p.
  Chaque coordonnée x_i correspond a un individu.
  On s'intéresse ici a des valeurs numériques.

- **Poids**
  Chaque individu a éventuellement un poids p_i, tel que $p_1 + \ + p_n = 1$.
  On a souvent p = 1 / n.

- **Représentations**
  histogramme en découpant les valeurs de la variable en classes.

- **Indicateurs statistiques**
- on dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, étendue, 1er quartile, 3eme quartile, ...
- Ces indicateurs mesurent principalement la tendance centrale et la dispersion.

- **Les méthodes d'analyse de données cherchent à proposer des alternatives plus performantes de fouille et visualisation des données**
- On utilisera principalement la moyenne, la variance et l'écart type.

1 – Des données aux bases d'apprentissage

Rappel Moyenne arithmétique

- **Définition**
  On note

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  ou pour des données pondérées

$$\bar{x} = \sum_{i=1}^{n} p_i x_i$$

- **Propriétés**
  la moyenne arithmétique est une mesure de tendance centrale qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée a cause de ses bonnes propriétés mathématiques.

1 – Des données aux bases d'apprentissage

Rappel Variance et ecart-type

- **Définition**
  la variance de x est définie par

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad ou \qquad s_x^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$

  L'écart type s_x est la racine carrée de la variance.
- **Propriétés**
  La variance satisfait la formule suivante

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} p_i x_i^2 - (\bar{x})^2$$

  La variance est « la moyenne des carres moins le carre de la moyenne ». L'ecart-type, qui a la même unité que x, est une mesure de dispersion.

Rappel Mesure de liaison entre deux variables

- Définitions la covariance observée entre deux variables x et y est

$$s_{xy} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \overline{xy}$$

et le coefficient de corrélation est donnée par

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} p_i (y_i - \bar{y})^2}}$$

Propriétés du coefficient de corrélation

- **Borne**
  On a toujours (inégalité de Cauchy Schwarz)

$$-1 \leq r_{xy} \leq 1$$

- **Variables liées**

$$\left| r_{xy} \right| = 1 \Leftrightarrow ax_i + by_i) = c \qquad \forall 1 \leq i \leq n$$

  $|r_{xy}| = 1$ si et seulement si x et y sont linéairement liées
  En particulier, $r_{xx} = 1$.

- **Variables décorrélées**
  si $r_{xy} = 0$, on dit que les variables sont decorrelees.
  Cela ne veut pas dire qu'elles sont indépendantes !

1 – Tableaux, vecteurs, nuages de points

Notation matricielle

- **Matrice**
  tableau de données carre ou rectangulaire.

- **Vecteur**
  matrice a une seule colonne.

- **Cas particuliers**

$$I = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \qquad 1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

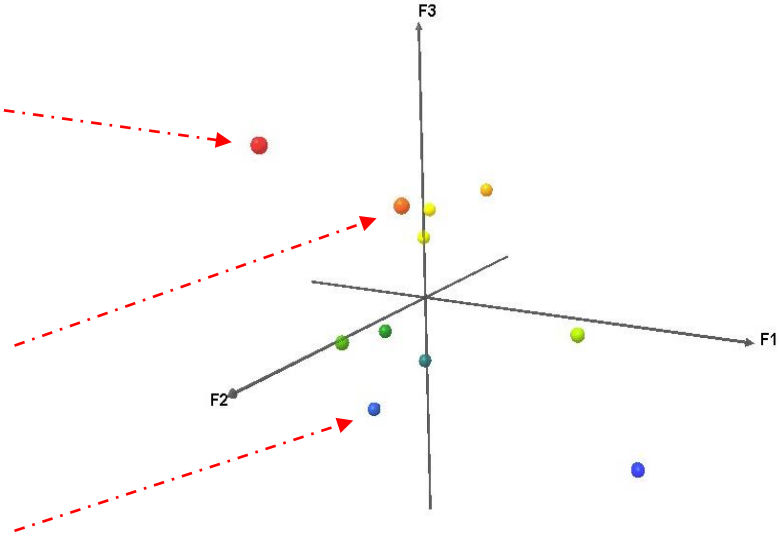
- **Transposition de matrice**
  échange des lignes et des colonnes d'une matrice ; on note M’ la transposée de M.

1 – Tableaux, vecteurs, nuages de points

Tableau de données

- Pour n individus et p variables, on a le tableau X
- X est une matrice rectangulaire a n lignes et p colonnes
- X est aussi un nuage de n points dans R^p

$$X = (x^1, \ldots, x^p) = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & & \\ & & \ddots & \\ \vdots & & x_i^j & \vdots \\ & & & \ddots \\ x_n^1 & & \cdots & x_n^p \end{bmatrix}$$

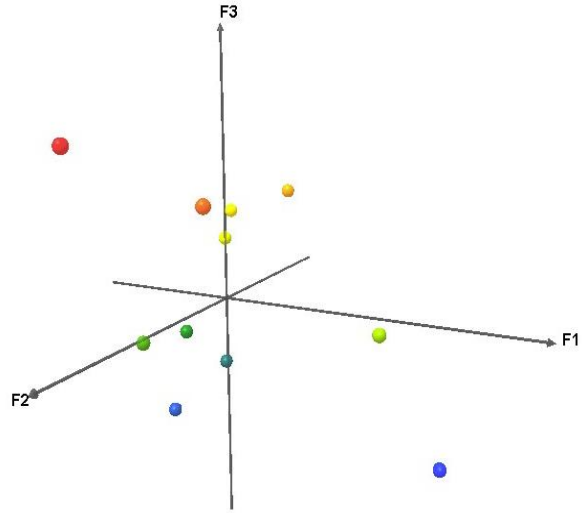


1 – Tableaux, vecteurs, nuages de points

Vecteurs variable et individu

- **Variable**
  Une colonne du tableau

$$x^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \\ \\ x_n^j \end{bmatrix}$$



- **Individu**
  Une ligne du tableau

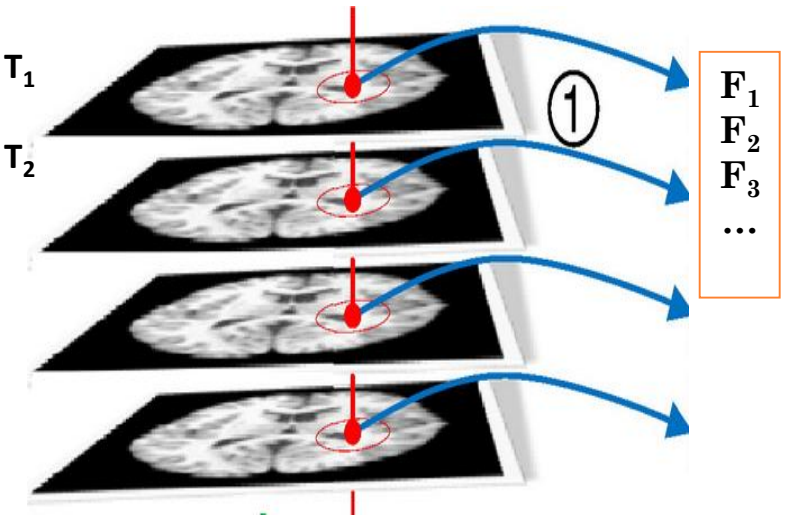
$$e_i' = (x_i^1 \quad x_i^2 \quad \quad x_i^p)$$

1 – Tableaux, vecteurs, nuages de points

Principle 1: 1 object ⇨ p features ⇨ 1 Vector = 1 Point ∈ R^p

- 1 table = n objects (individuals)
- 1 object ⇨ p features ⇨ 1 Vector = 1 Point ∈ R^p
- Statistics, Data analysis and Machine Learning tools become usable

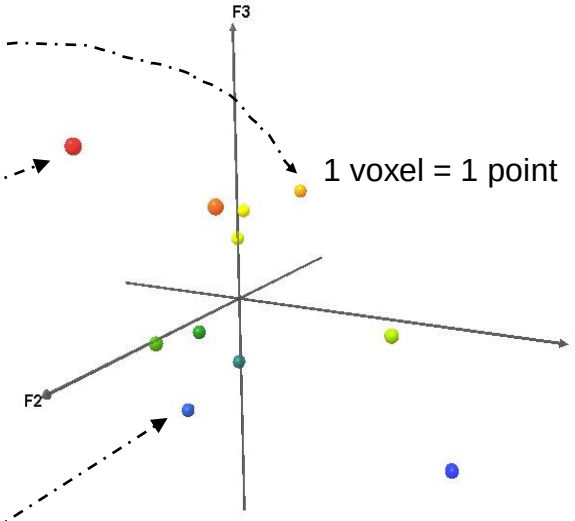
1 pixel / Voxel with several modalities



F_1
F_2
F_3
…

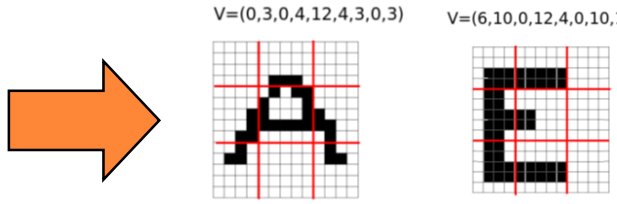
$Voxel_1 = [F_1\ F_2\ F_3\ \ldots]$
$Voxel_2 = [F_1\ F_2\ F_3\ \ldots]$
$Voxel_3 = [F_1\ F_2\ F_3\ \ldots]$
$Voxel_4 = [F_1\ F_2\ F_3\ \ldots]$
$Voxel_5 = [F_1\ F_2\ F_3\ \ldots]$
…..
$Voxel_N = [F_1\ F_2\ F_3\ \ldots]$

Table of Voxels



1 table = 1 cloud of points

- An other examples (OCR)

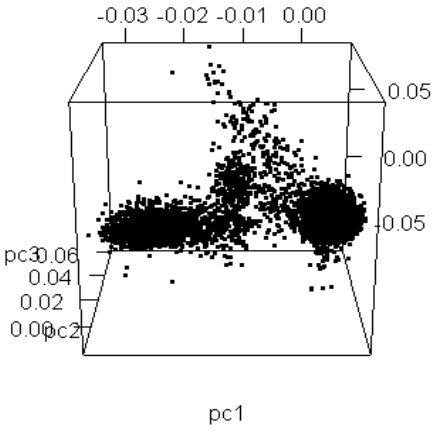


1 – Tableaux, vecteurs, nuages de points

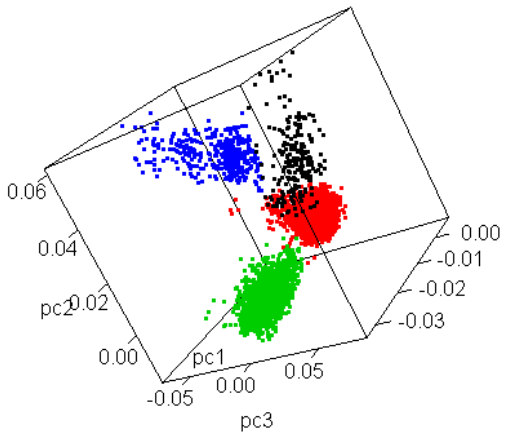
Principle 1: 1 object ⇨ p features ⇨ 1 Vector = 1 Point ∈ R^p

- Each object has to be described by features with eventually **additional labels**
- 1 object ⇨ p features ⇨ 1 Vector = 1 Point ∈ R^p **+ additional labels**
- Statistics, Data analysis and Machine Learning tools become usable

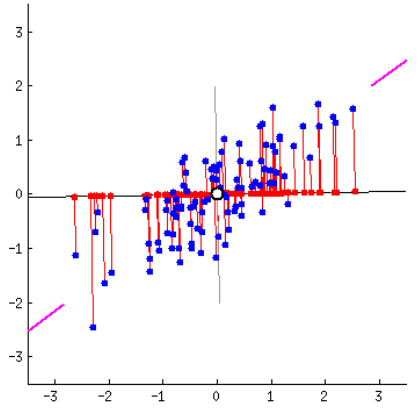
$\mathbf{Voxel_1} = [F_1\ F_2\ F_3\ \ldots]$
$\mathbf{Voxel_2} = [F_1\ F_2\ F_3\ \ldots]$
$\mathbf{Voxel_3} = [F_1\ F_2\ F_3\ \ldots]$
$\mathbf{Voxel_4} = [F_1\ F_2\ F_3\ \ldots]$
$\mathbf{Voxel_5} = [F_1\ F_2\ F_3\ \ldots]$
.....
$\mathbf{G} = [F_1\ F_2\ F_3\ \ldots]$



$\mathbf{Voxel_1} = [F_1\ F_2\ F_3\ \ldots]$ [Tumor]
$\mathbf{Voxel_2} = [F_1\ F_2\ F_3\ \ldots]$ [Normal]
$\mathbf{Voxel_3} = [F_1\ F_2\ F_3\ \ldots]$ [Tumor]
$\mathbf{Voxel_4} = [F_1\ F_2\ F_3\ \ldots]$ [Tumor]
$\mathbf{Voxel_5} = [F_1\ F_2\ F_3\ \ldots]$ [Normal]
$G_G = [F_1\ F_2\ F_3\ \ldots]$ ↑
$\mathbf{G_R} = [F_1\ F_2\ F_3\ \ldots]$ **Label**



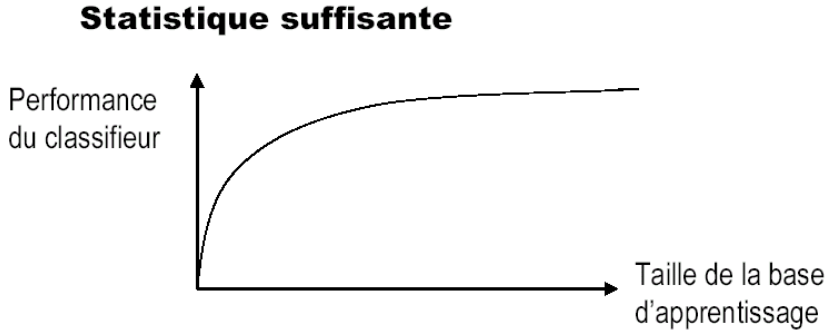
$\mathbf{Voxel_1} = [F_1]$
$\mathbf{Voxel_2} = [F_1]$
$\mathbf{Voxel_3} = [F_1]$
$\mathbf{Voxel_4} = [F_1]$
$\mathbf{Voxel_5} = [F_1]$
.....
$\mathbf{G} = [F_1]$

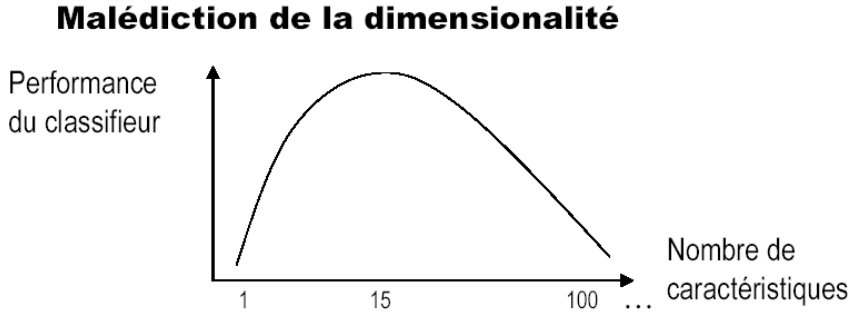


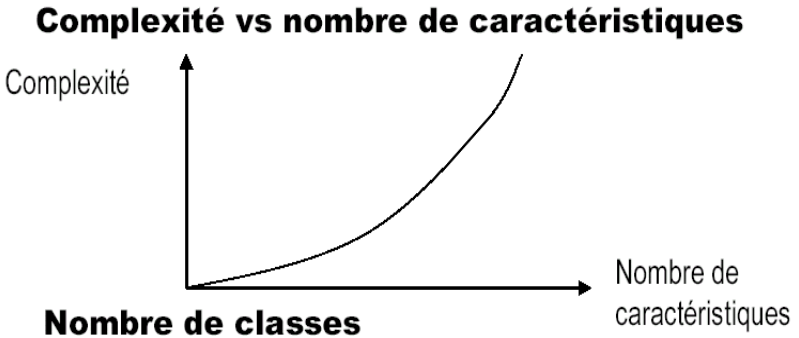
Bases d'apprentissage

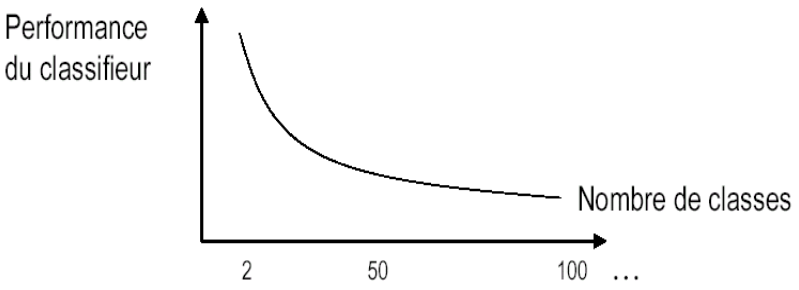
Principle 2: Data (labeled) becomes the Graal ➔ Big data / data science

- 1 object ⇨ p features ⇨ 1 Vector = 1 Point $\in R^p$
- **Dataset** = Table of Vectors with or without a label (class)
- **Selected Features are crucial** (expertise, …)
  - Stable and discriminative
  - And…
- If a label is available for each object, we speak of Learning / Training sets





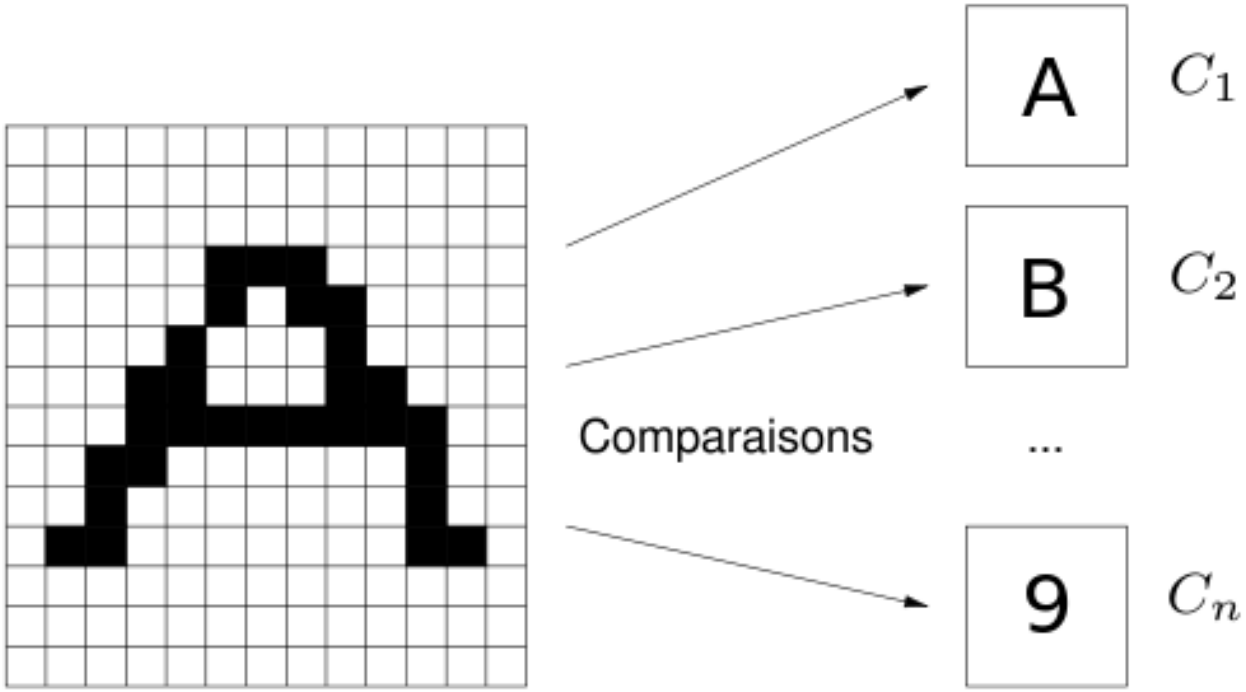




Feature selection (simplified)

Pixels can be considzered as features

- distance(C; C_i) = $\sum$ij | P (i; j) - P_i(i; j) |



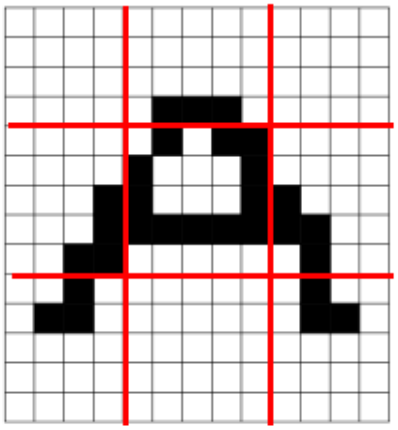
C : Unknow object

Training set (tagged models of object)

Feature selection (simplified)

Zoning

- The image is splitted in n blocks
- For each block, some features are computed (number of black pixels)
- A new feature vector is obtained : V = (Nb_1; Nb_2; :::; Nb_n)



V=(0,3,0,4,12,4,3,0,3)

Pondération, Inertie et métrique

La matrice des poids

- **Pourquoi**
  utile quand les individus n'ont pas la même importance

- **Comment**
  on associe aux individus un poids pi tel que

$$p_1 + p_2 + ... + p_n = 1$$

  et on représente ces poids dans la matrice diagonale de taille n

$$D = \begin{bmatrix} p_1 & & \dots & 0 \\ & p_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & p_n \end{bmatrix}$$

- **Cas uniforme**
  tous les individus ont le même poids $p_i = 1 / n$ et $D = I / n$

Pondération, Inertie et métrique

Point moyen et tableau centré

- **Point moyen** (centre de gravité)
  c'est le vecteur g des moyennes arithmétiques de chaque variable :

$$g' = (\bar{x}^1 \quad \dots \quad \bar{x}^p)$$

ou

$$\bar{x}^j = \sum_{i=1}^{n} p_i x_i^j$$

- On peut aussi écrire $g = X' D 1$

Tableau centré

il est obtenu en centrant les variables autour de leur moyenne

$$y_i^j = x_i^j - \bar{x}^j$$

ou, en notation matricielle,

$$Y = X - 1g' = (I - 11' D) X$$

Pondération, Inertie et métrique

Matrice de variance covariance

- **Définition**
  c'est une matrice carrée de dimension p

$$V = \begin{bmatrix} s_1^1 & s_1^2 & \dots & s_1^p \\ s_2^1 & s_2^2 & & \\ \vdots & & \ddots & \vdots \\ s_p^1 & & \dots & s_p^p \end{bmatrix}$$

  ou s_{kl} est la covariance des variables x^k et x^l et s^2_j est la variance de la variable x^j

- **Formule matricielle**

$$V = X'DX - gg' = Y'DY$$

Pondération, Inertie et métrique

Matrice de corrélation

- **Définition**
  Si l'on note

$$r_{kl} = \frac{s_{kl}}{s_k s_l} \qquad R = \begin{bmatrix} 1 & r_1^2 & \dots & r_1^p \\ r_2^1 & 1 & & \\ \vdots & & \ddots & \vdots \\ s_p^1 & & \dots & 1 \end{bmatrix}$$

- **Formule matricielle**

$$R = D_{1/s} V D_{1/s} \qquad D_{1/s} = \begin{bmatrix} \frac{1}{s_1} & & & 0 \\ \vdots & \frac{1}{s_2} & & \vdots \\ & & \ddots & \\ 0 & & & \frac{1}{s_p} \end{bmatrix}$$

Pondération, Inertie et métrique

Inertie d'un nuage de points

- **Définition**
  l'inertie en un point a du nuage de points est

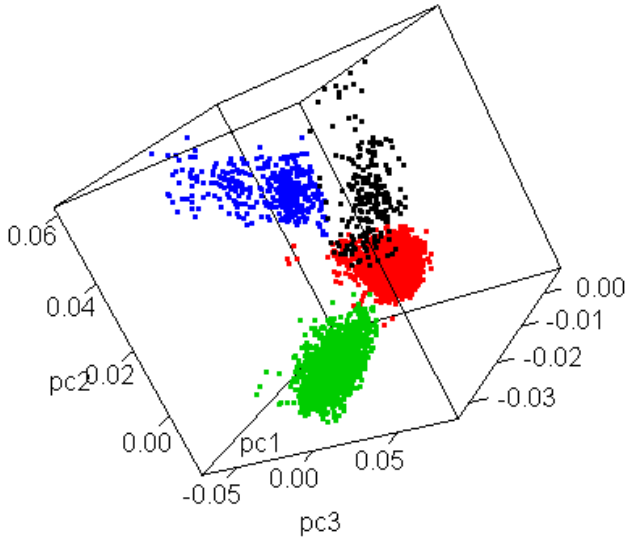
$$I_a = \sum_{i=1}^{n} p_i \left\| e_i - a \right\|_M^2 = \sum_{i=1}^{n} p_i (e_i - a)' M (e_i - a)$$

- **Autres relations**
  l'inertie totale Ig est la moitie de la moyenne des carres des distances entre les individus

$$2I_g = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left\| e_i - e_j \right\|_M^2$$

- L'inertie totale est aussi donnée par la trace de la matrice MV (la trace d'une matrice étant la somme de ses éléments diagonaux).

$$I_g = Tr(MV)$$



L'inertie est une mesure fondamentale pour évaluer la qualité d'un espace de représentation (inertie totale, inter-classe, intra-classe)

Pondération, Inertie et métrique

Distance entre individus

- **Motivation**
  afin de pouvoir considérer la structure du nuage des individus, il faut définir une distance, qui induira une géométrie.

- **Distance euclidienne classique**
  la distance la plus simple entre deux points de Rp est définie par

$$d^2(u,v) = \sum_{j=0}^{p} (u_j - v_j)^2 = \|u - v\|$$

- **Généralisation simple**
  on multiplie la variable j par $\sqrt{a_j}$

$$d^2(u,v) = \sum_{j=0}^{p} a_j (u_j - v_j)^2$$

Pondération, Inertie et métrique

Métrique

- **Matrice définie positive** *M*
  c'est une matrice symétrique telle que, pour tout u non nul, u'Mu > 0.

- **Définition**
  soit M = (m_{jk}) définie positive de dimension p. On pose

$$\|u\|_M^2 = u'Mu = \sum_{j=0}^{p}\sum_{k=1}^{p} m_{jk}u_j u_k \qquad et \qquad d_M^2(u,v) = \|u-v\|_M^2$$

- **Espace métrique**
  il est défini par le produit scalaire

$$\langle u,v\rangle_M = u'Mu = \sum_{j=0}^{p}\sum_{k=1}^{p} m_{jk}u_j u_k$$

  On dit que u et v sont orthogonaux si $\langle u,v\rangle_M = 0$

Pondération, Inertie et métrique

Métriques particulières

- **Métrique usuelle**
  M = I correspond au produit scalaire usuel et

$$I_g = Tr(V) = \sum_{j=1}^{p} s_i^2$$

- **Problèmes**
  - la distance entre individus dépend de l'unité de mesure.
  - la distance privilégie les variables les plus dispersées.
- **Métrique réduite**
  c'est la plus courante ;
  on prend la matrice
  diagonale des inverses
  des variances

$$M = D_{1/s^2} = \begin{bmatrix} \frac{1}{s_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{s_p^2} \end{bmatrix}$$

$$I_g = Tr(D_{1/s^2} V) = Tr(D_{1/s} V D_{1/s}) = Tr(R) = p$$

Pondération, Inertie et métrique

Métriques et tableaux transformes

- Utiliser la métrique M = T'T sur le tableau X est équivalent a travailler avec la métrique classique I sur le tableau transforme XT'.

- **Tableau transformé**
  Si on travaille sur le tableau transforme XT' (changement de variables) au lieu de X, alors les nouveaux individus seront de la forme Te_i et

$$\left\langle Te_{i_1}, Te_{i_2} \right\rangle = (Te_{i_1})'(Te_{i_2}) = e_{i_1}'T'Te_{i_2} = e_{i_1}'Me_{i_2} = \left\langle e_{i_1}, e_{i_2} \right\rangle_M$$

- **Réciproque**
  pour toute matrice symétrique positive M, il existe une matrice T (racine carrée de M) telle que

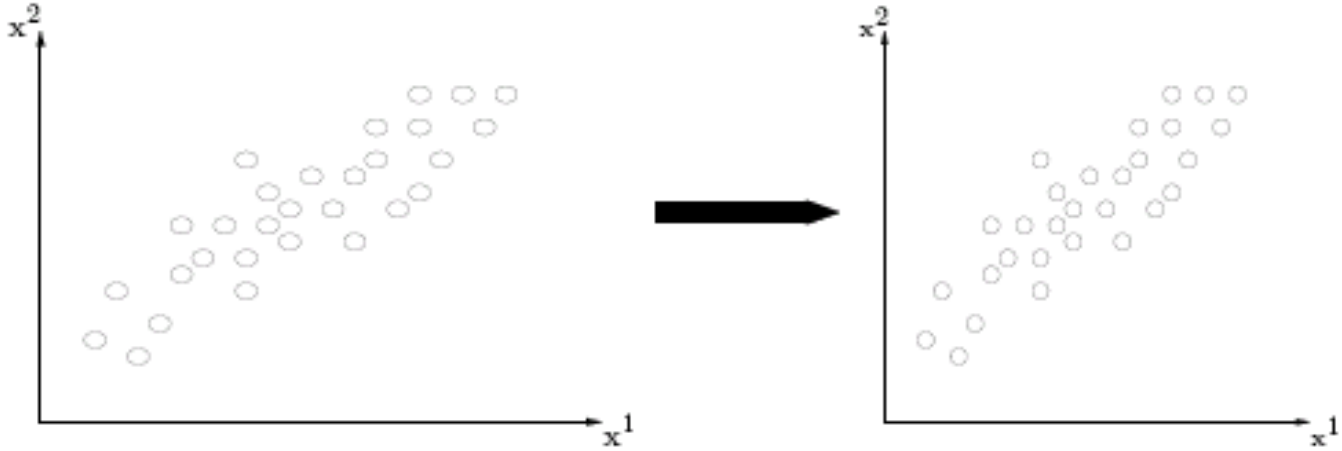
$$M = T'T$$

  et donc on peut ramener l'utilisation de la métrique a un changement de variables.

Pondération, Inertie et métrique

Métriques et tableaux transformes (suite)

- Utiliser une métrique est donc équivalent à tordre les données pour les rendre comparables

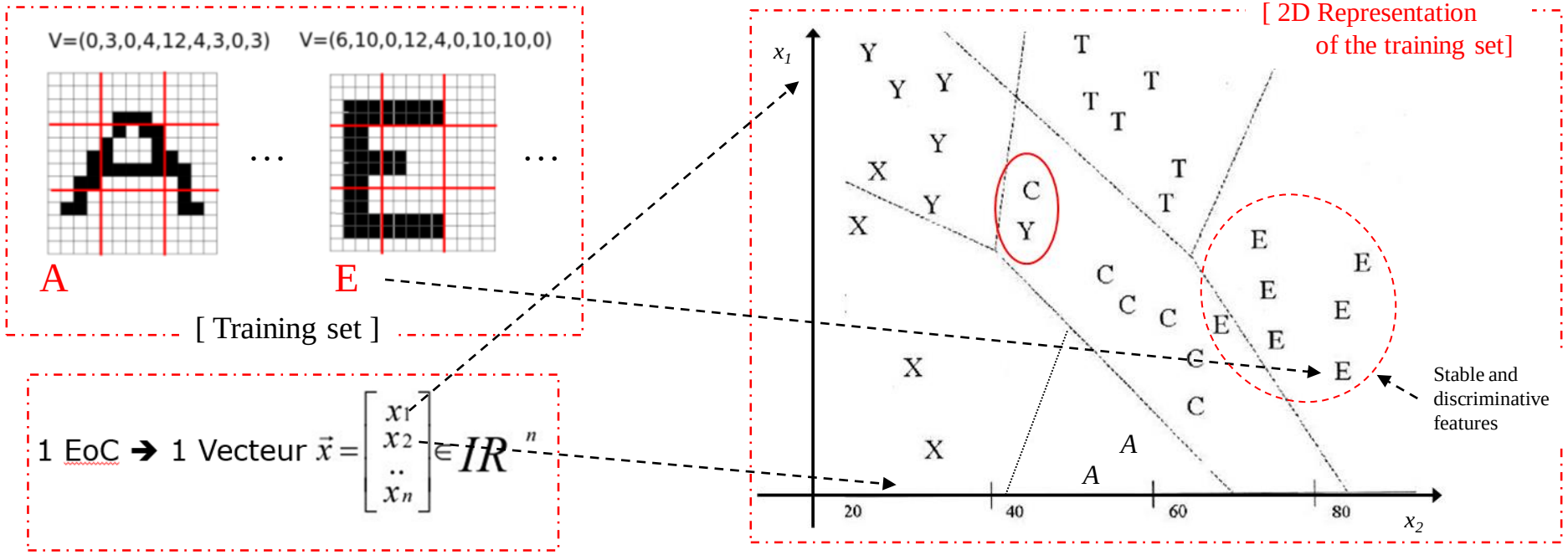


- Exemple utiliser la métrique réduite est équivalent a travailler sur les données centrées réduites $Z = YD_{1/s}$.

Similarités, dissimilarités, distances

How computers can recognize objects?

- We need a **large set of (labelled) examples** similar to the patterns to be recognized ➔ a training set
- We need a list of **stable and discriminative** features (shape, color, size,…) used to describe the patterns (labelled ones and unknown one)

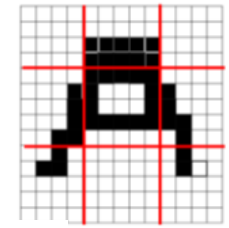


Similarités, dissimilarités, distances

How computers can recognize objects?

- We need a powerful **similarity measure** (dissimilarity, distance, metric) to compare objects together

Unknown object



V=(0,10,0,5,14,5,3,0,2)

V=(0,3,0,4,12,4,3,0,3)

V=(6,10,0,12,4,0,10,10,0)

A

Forme inconnue

?

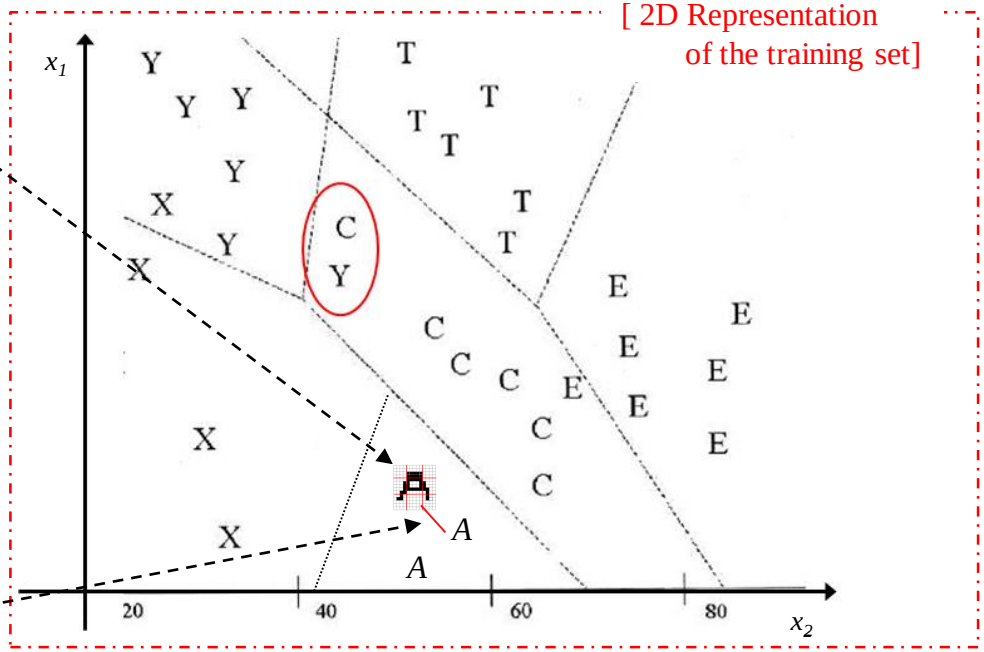
?

V=(0,10,0,5,14,5,3,0,2)

$$D(A,?) = \sqrt{(0-0)^2 + (3-10)^2 + (0-0)^2 + (4-5)^2 + \ldots + (3-2)^2}$$

$$D(A,?) = 7,48 \text{ et } D(B,?) = 19,05$$

[2D Representation of the training set]



TP…

That All for today…

MERCI…

Support de cours :

- TP : http://www.rfai.lifat.univ-tours.fr/PagesPerso/jyramel/PDF/TP1cesr_eleves.ipynb
- Slides : http://www.rfai.lifat.univ-tours.fr/PagesPerso/jyramel/PDF/HNcesr_ADML2019part1.pdf