Analyse des données

Apprentissage supervisé et non supervisé [Introduction]

3ème année ENSEIRB-MATMECA - CISD

Quelles données ?

Un **tableau de données** peut contenir des données :

- ***quantitatives***
- qualitatives
- mixtes

	amère	sucrée	acide	salée	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Badoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3
Arvie	3.1	3.2	3.0	5.2	5.0
Chateaune	3.7	2.8	3.0	5.2	4.6
Salvetat	4.0	2.8	3.0	4.1	4.5
Perrier	4.4	2.2	4.0	4.9	3.9

Quelles données ?

Un **tableau de données** peut contenir des données :

- quantitatives
- ***qualitatives***
- mixtes

	taille	poids	velocite	intellig	affect	agress
beauceron	T++	P+	V++	I+	Af+	Ag+
basset	T-	P-	V-	I-	Af-	Ag+
ber_allem	T++	P+	V++	I++	Af+	Ag+
boxer	T+	P+	V+	I+	Af+	Ag+
bull-dog	T-	P-	V-	I+	Af+	Ag-
bull-mass	T++	P++	V-	I++	Af-	Ag+
caniche	T-	P-	V+	I++	Af+	Ag-

Quelles données ?

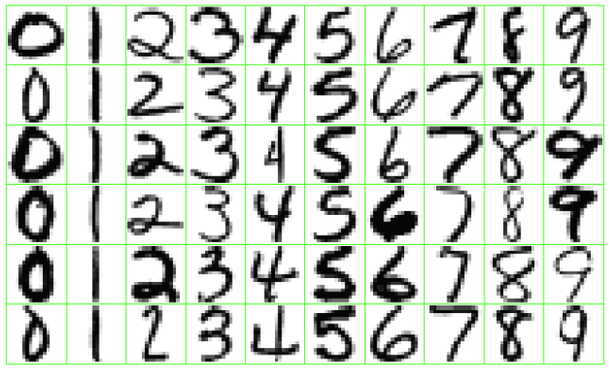
Un **tableau de données** peut contenir des données :

- quantitatives
- qualitatives
- ***mixtes***

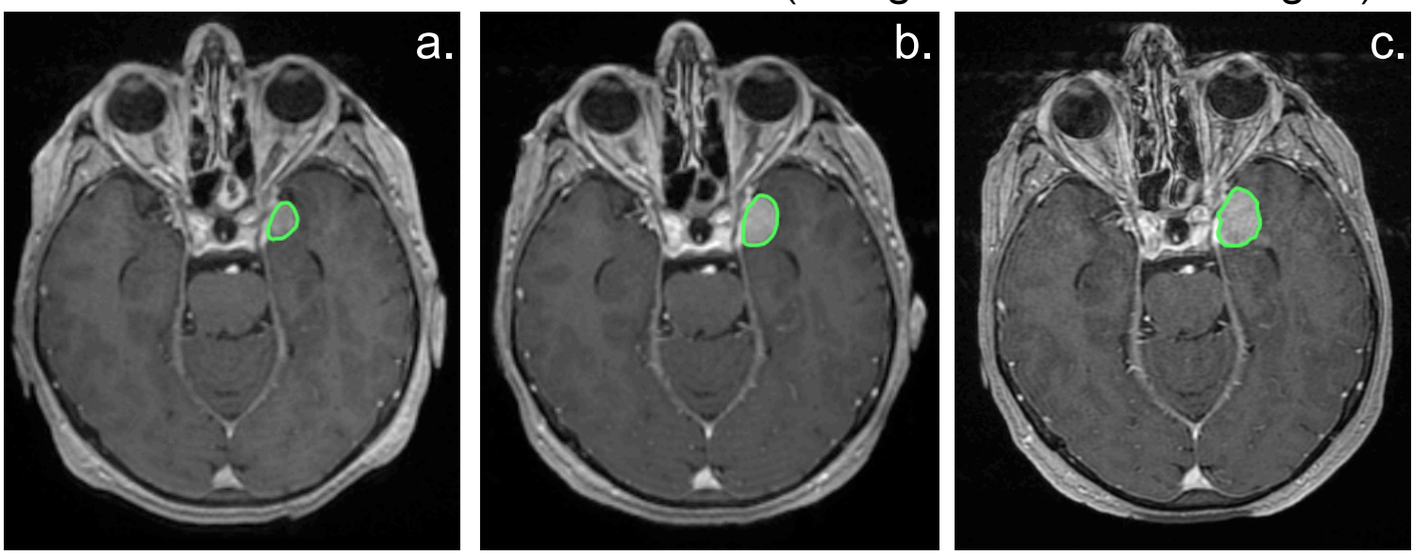
	Label	Bitterness	Smooth	Harmony
2EL	Saumur	1.9	2.7	3.1
1CHA	Saumur	1.9	2.5	3.0
1FIN	Bourgueuil	2.0	2.7	3.1
1VAU	Chinon	2.0	1.7	2.0
1DAM	Saumur	2.1	3.0	3.6

Quelles données ?

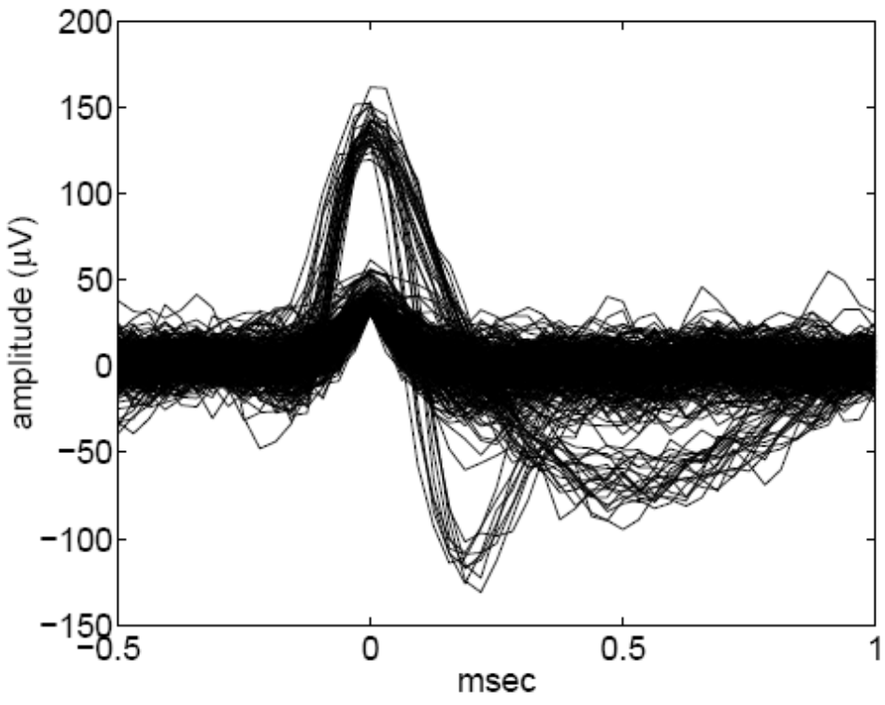
Autres type de données …



CHU Bordeaux MRI T1 (with gadolinium contrast agent)



$t_0 = 0$ $t_1 = 169$ days $t_2 = 330$ days



Spam

```
WINNING NOTIFICATION
We are pleased to inform you of the result of the Lottery Winners
International programs held on the 30th january 2005. [...] You
have been approved for a lump sum pay out of 175,000.00 euros.
CONGRATULATIONS!!!
```

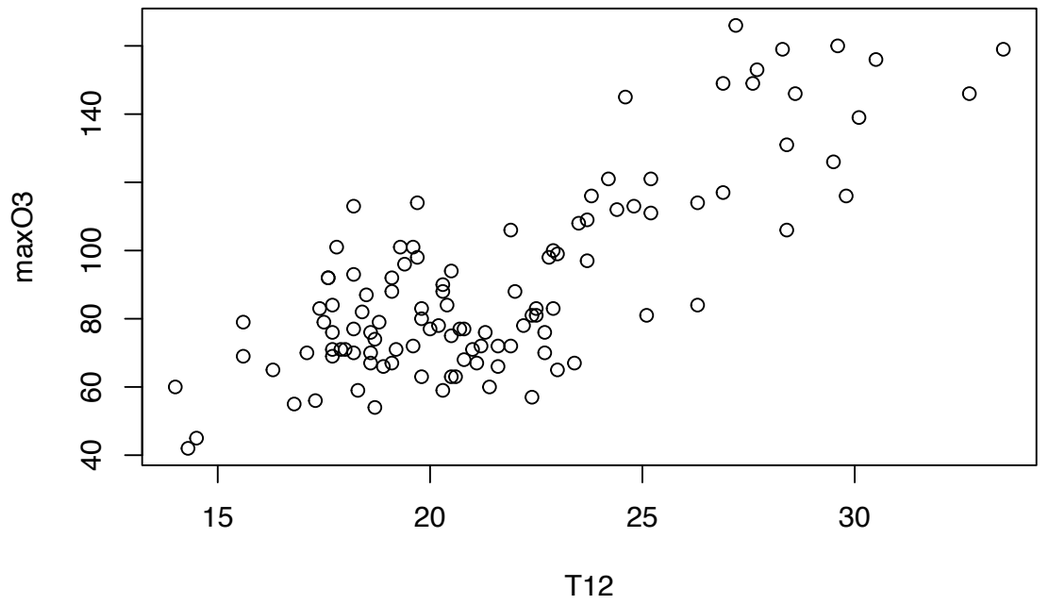
No Spam

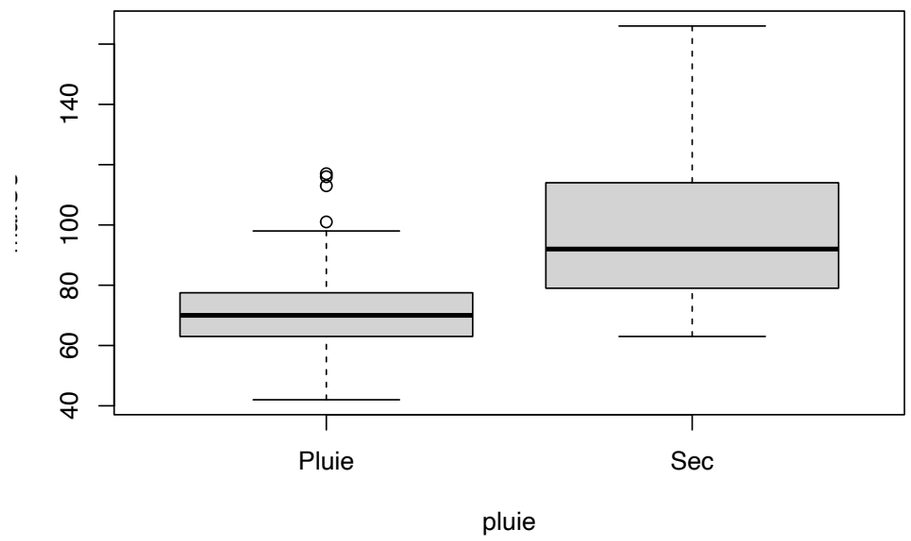
```
Dear George,
Could you please send me the report #1248 on the project
advancement?
Thanks in advance.

Regards,
Cathia
```

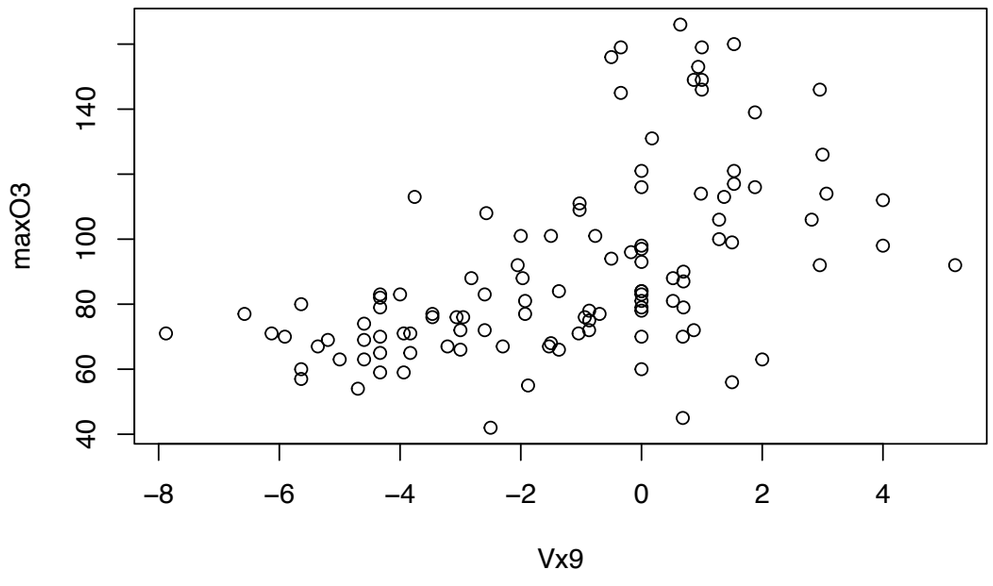
Premier exemple d'application

Prédire le maximum de concentration d'ozone

Prédire une variable quantitative : problème de régression







Données de surveillance de la qualité de l'air à Rennes pendant l'été 2001 pendant 112 jours ;

max03 : concentration maximum (variable à expliquer quantitative)

T12 : température à 12h (variable explicative quantitative)

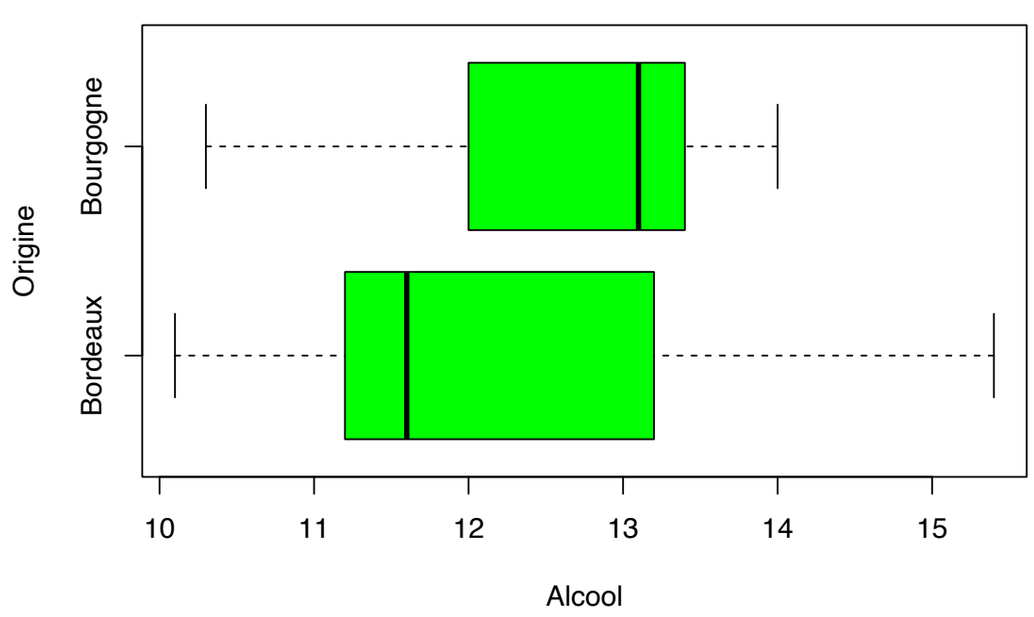
Vx9 : composante E-0 du vent à 9h (variable explicative quantitative)

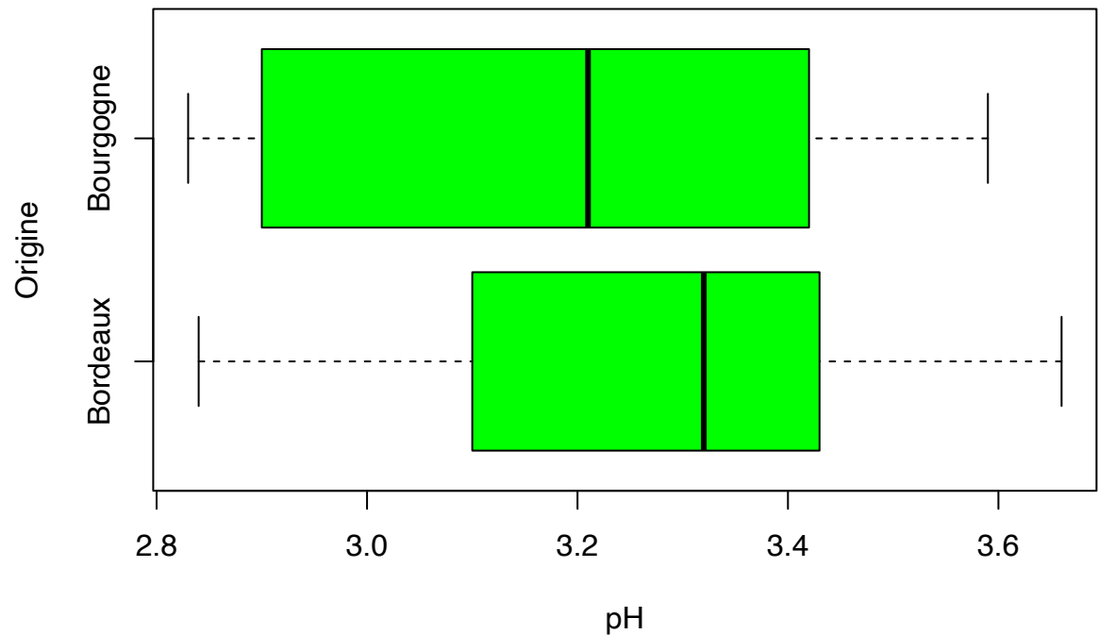
Pluie :variable explicative qualitative à deux modalités (sec, pluie)

Deuxième exemple d'application

Prédire l'origine d'un vin

Prédire une variable qualitative : problème de classification





Données qui décrivent 31 vins

Origine : variable à expliquer qualitative à 2 modalités (Bordeaux, Bourgogne)

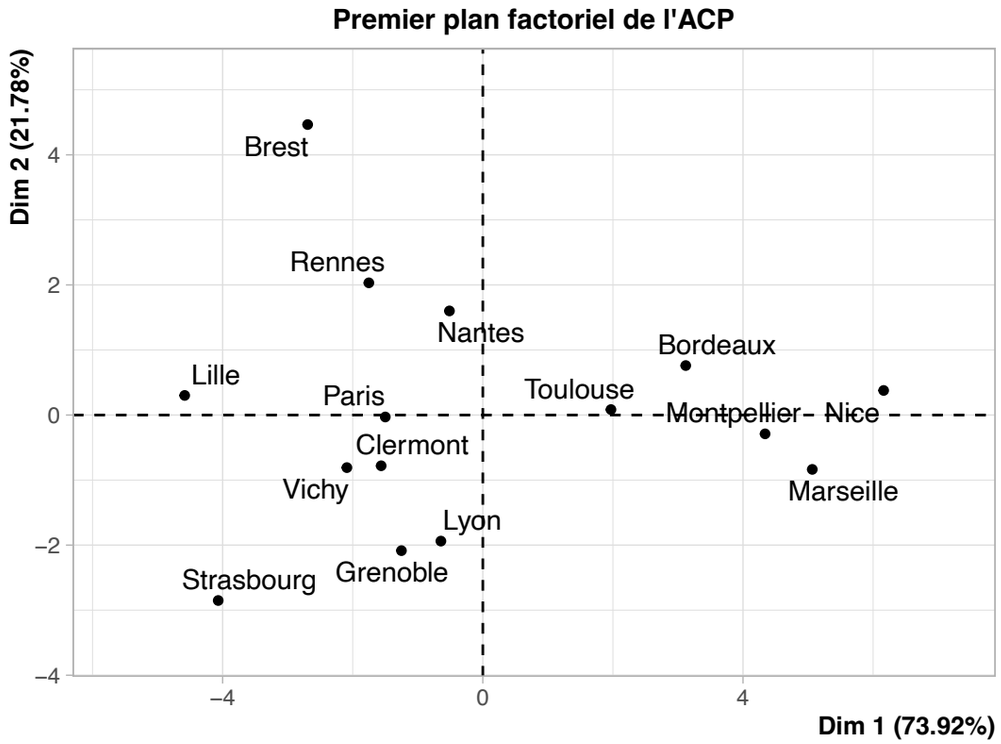
Alcool : variable explicative quantitative

pH : variable explicative quantitative

Troisième exemple d'application

Visualiser des villes en fonction de leurs températures moyennes mensuelles.

Pas de variable à prédire : réduire la dimension (ACP)



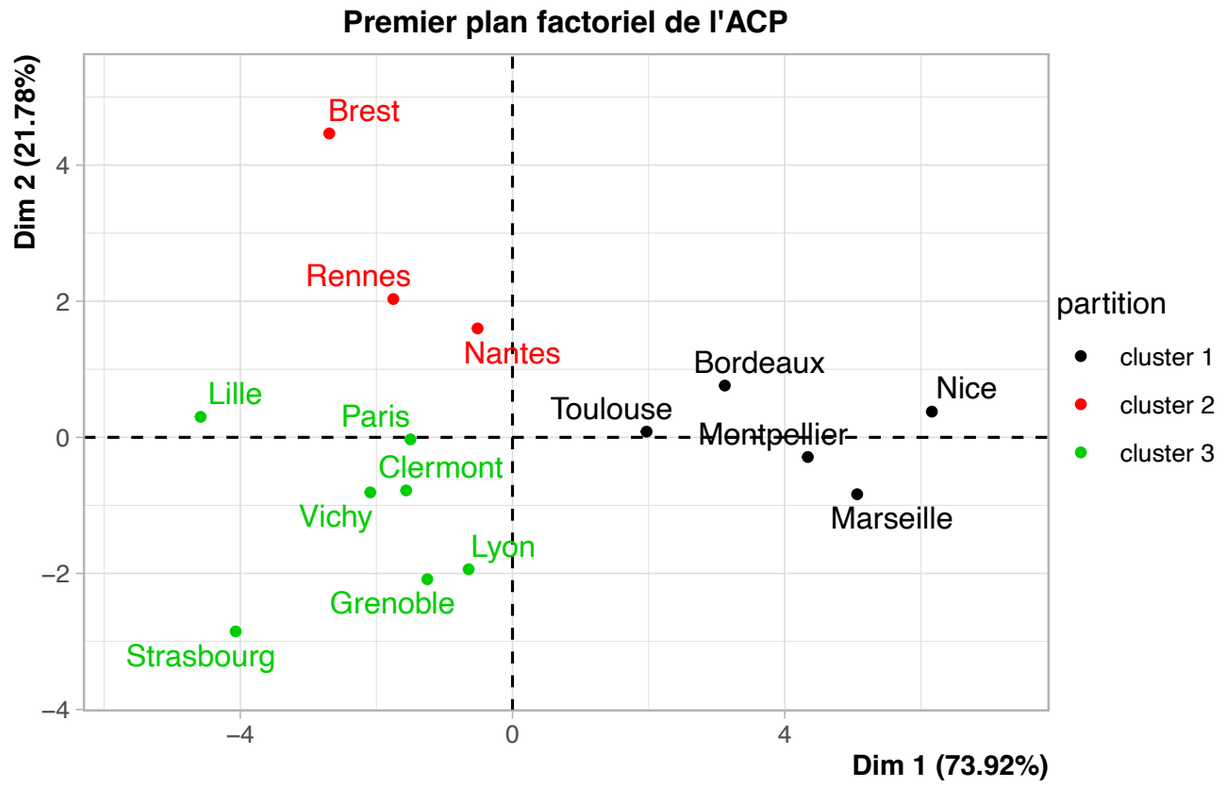
Données qui décrivent 15 villes en fonction de leurs températures moyennes sur les 12 mois de l'année

12 variables quantitatives corrélées résumées ici par 2 nouvelles variables non corrélées

Quatrième exemple d'application

Regrouper les villes en fonction de leurs températures

Pas de variables à prédire : trouver une partition (clustering)



Les 15 villes sont partitionnées en 3 classes (clusters) : création d'une nouvelle variable qualitative

Histoire de l'analyse de données

Bien que le terme **apprentissage statistique** soit relativement nouveau, la plus part des concepts utilisés ont été développés il y a longtemps.

~1800
Méthode des moindres carrés
(forme ancienne de la régression linéaire)
Legendre et Gauss
[Prédire des valeurs quantitatives]

1936
Analyse discriminante linéaire
Fisher
[Prédire des valeurs qualitatives]

1940 ++
Régression logistique
Divers auteurs
[Prédire des valeurs qualitatives]

1972
Modèles linéaires généralisés
Nelder et Wedderbur
[Généralisation des précédents concepts]

1984
Arbres de classification et de régression
Validation croisée
Breiman, Friedman, Olshen et Stone
[Passage au non linéaire]

Choisir une méthode

Méthodes de Machine Learning
(Apprentissage Automatique)

Non supervisées

$X^1, \ldots, X^p$: variables (quantitatives ou qualitatives)

Clustering

Création d'une nouvelle variable qualitative

Exemple : k-means, CAH, GMM,…

Réduction de dimension

Création de nouvelles variables quantitatives qui résument $X^1, \ldots, X^p$.

Exemples de méthodes linéaires :

- ACP si les données sont quantitatives
- ACM si les données sont qualitatives
- ACPmixte si les données sont mixtes

Exemple de méthodes non linéaire : AutoEncoders,…

Supervisées

$X^1, \ldots, X^p$: variables d'entrées

Y : variable de sortie (quantitative ou qualitative

Régression : Y quantitatif

Exemples de méthodes linéaires :

- Régression linéaire simple et multiple si entrées quantitatives
- ANOVA si entrées qualitatives
- ANCOVA si les entrées sont mixtes

Exemples de méthodes non linéaires :

- Arbre de décision et forêts aléatoires (entrées mixtes)
- SVM, réseaux de neurones

Classification : Y qualitatif

Exemples de méthodes linéaires :

- Régression logistique (Y binaire)
- LDA et QDA (entrées quantitatives)

Exemples de méthodes non linéaires :

- KNN, réseaux de neurone (entrées quantitatives)
- Bayésien naif, arbres et forêts aléatoires (entrées mixtes)

Choix de R pour le code

R est un langage de programmation et un logiciel libre destiné aux statistiques et à la science des données soutenu par la R Foundation for Statistical Computing. R fait partie de la liste des paquets GNU et est écrit en C (langage), Fortran et R. GNU R est un logiciel libre distribué selon les termes de la licence GNU GPL et disponible sous GNU/Linux, FreeBSD, NetBSD, OpenBSD, Mac OS X et Windows.
Le langage R est largement utilisé par les statisticiens, les data miners, data scientists pour le développement de logiciels statistiques et l'analyse des données.
En Janvier 2023, R est classé 18e dans l'index TIOBE qui mesure la popularité des langages de programmation.

Toutes les méthodes vues dans ce cours sont déjà implémentées dans R ce qui permettra en TP de se concentrer sur le choix de la méthode, sa validation et l'interprétation statistique qu'il est possible de faire.

Ce dont nous ne parlerons pas …

Apprentissage profond (Deep learning)
Ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaires