SY09 – Analyse de données et *Data Mining*

Cours n° 1 – Introduction

Sylvain Rousseau

Printemps 2019

Emploi du temps et documents

- Responsable : Benjamin Quost
- Intervenants : Benjamin Quost, Sébastien Destercke et Sylvain Rousseau
- Emploi du temps :
  - Cours : mercredi 10h15-12h15 (FA106)
  - TD :
    - lundi 14h15-16h15 (FA509)
    - mercredi 8h-10h (FB113)
    - jeudi 10h15-12h15 (FA509)
    - vendredi 14h15-16h15 (FA509)
- Documents :
  - Polycopié de cours (BUTC)
  - Transparents (Moodle)
- Projet (40%), final (60%, note ≤ 6 éliminatoire)

Terminologie

- **Analyse de données** – Traitement des données au sens large
  - Acquisition, nettoyage, modélisation
- **Fouille de données** (*Data mining*) – Extraction de connaissance
  - Toute méthode permettant d'extraire de la connaissance
- **Reconnaissance de formes** (*Pattern recognition*)
  - Découverte automatique de régularité dans les données
- **Apprentissage automatique**
  - Synonyme. Issu de la communauté informatique
- **Apprentissage statistique** (*Machine learning*)
  - Apprentissage automatique + Statistique

Exemples

- **Reconnaissance de formes** : La recherche de « formes » ou de régularités dans un ensemble de données est un problème ancien qui a souvent conduit à des découvertes importantes :
  - Lois de Kepler à partir d'observations astronomiques
  - Physique quantique à partir de l'observation de spectres de lumière
- ***Data mining*** : Rechercher des règles, corrélation
- **Apprentissage statistique**
  - classement
  - régression

Étapes du processus d'extraction d'information

OSEMN : Obtain, Scrub, Explore, Model, and iNterpret

- Récupération des données *(Obtain)*
  - Web scraping, base de données, API, expériences
- Nettoyage des données (60 % du processus) *(Scrub)*
  - Données manquantes, données atypiques (*outliers*), mise au format, désambiguïsation...
- Exploration (*Explore*)
  - Sélection des variables, des individus
  - Visualisation simple
- Modélisation (*Model*)
  - Choix des objectifs : résumé, classification, régression
  - Choix des méthodes
  - Application des méthodes
- Analyse des résultats *(iNterpret)*
  - Visualisation, Interprétation
  - Retour aux étapes précédentes

Étapes du processus d'extraction d'information

Concerne plusieurs disciplines indépendantes en apparence

- Informatique
  - Stockage (base de données, *data warehouse*)
  - Parallélisation (Hadoop, Spark,...)
  - Indexation (Elasticsearch, Lucene,...)
- Mathématiques/Statistiques
  - Modélisation statistique
  - Inférence statistique
  - Optimisation
- Compréhension des données
  - Représentation de l'information
  - Description

Avènement du *Big Data*

- ***Volume*** – Explosion de la quantité de données disponibles, nombreux gisements de données
  - Appareils de mesure : capteurs de pollution, images satellitaires, ...
  - Fichiers de logs
  - Le Web : photos, réseaux sociaux, discussions
- ***Variety*** – Diversité des données qui sont très différentes pas nature : données non structurées
  - photos
  - vidéos
  - audio
  - texte
  - réseaux
- ***Velocity*** – Les données sont générées à un rythme effréné
  - Problématique temps réel

Les étapes du processus d'extraction d'information deviennent inter–dépendantes !

Science des données

- Science des données (*data science*) : nouvelle discipline à l'intersection des mathématiques et statistiques, de l'informatique et de visualisation des données
- Des journaux scientifiques (début 2000) :
  - Journal of Data Science
  - CODATA Data Science Journal
- Une profession ? *data scientist* : expression forgée en 2008 dans la Silicon Valley par deux ingénieurs travaillant chez LinkedIn et Facebook

Analyse du jeu de données *midge*

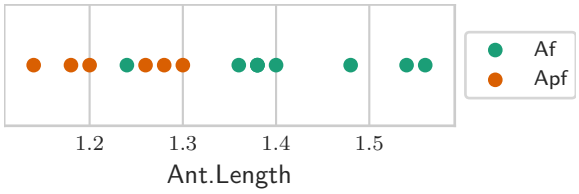
- Découverte d'un nouveau genre de moucherons (1981)
- Deux espèces en particulier difficiles à distinguer
  - *Amerohelea fasciata* (Af)
  - *Amerohelea pseudofasciata* (Apf)
- Jeu de données
  - 9 moucherons Af et 6 moucherons Apf
  - Longueurs de l'aile et de l'antenne en mm

Species	Ant.Length	Wing.Length
Af	1.38	1.64
Af	1.4	1.7
Af	1.24	1.72
Af	1.36	1.74
Af	1.38	1.82
Af	1.48	1.82
Af	1.54	1.82
Af	1.38	1.9
Af	1.56	2.08
Apf	1.14	1.78
Apf	1.2	1.86
Apf	1.18	1.96
Apf	1.3	1.96
Apf	1.26	2
Apf	1.28	2

Comment discriminer les deux espèces ?

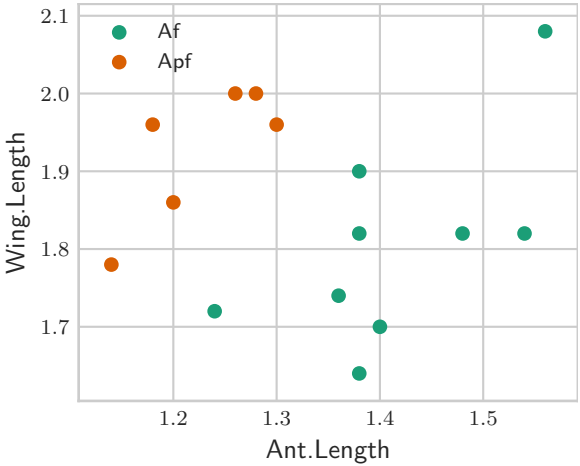
Analyse exploratoire univariée

Species	Ant.Length	Wing.Length
Af	1.38	1.64
Af	1.4	1.7
Af	1.24	1.72
Af	1.36	1.74
Af	1.38	1.82
Af	1.48	1.82
Af	1.54	1.82
Af	1.38	1.9
Af	1.56	2.08
Apf	1.14	1.78
Apf	1.2	1.86
Apf	1.18	1.96
Apf	1.3	1.96
Apf	1.26	2
Apf	1.28	2



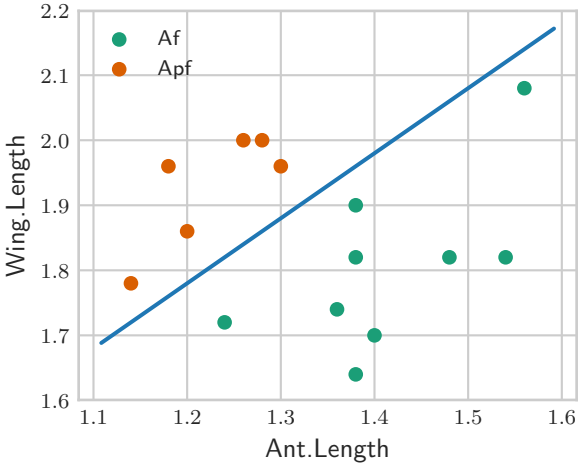
Analyse exploratoire bivariée

Species	Ant.Length	Wing.Length
Af	1.38	1.64
Af	1.4	1.7
Af	1.24	1.72
Af	1.36	1.74
Af	1.38	1.82
Af	1.48	1.82
Af	1.54	1.82
Af	1.38	1.9
Af	1.56	2.08
Apf	1.14	1.78
Apf	1.2	1.86
Apf	1.18	1.96
Apf	1.3	1.96
Apf	1.26	2
Apf	1.28	2



Frontière de décision

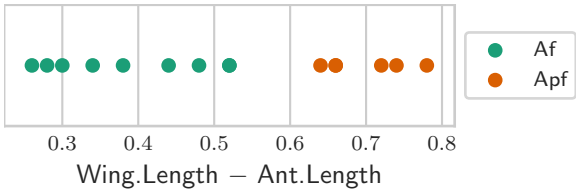
Species	Ant.Length	Wing.Length
Af	1.38	1.64
Af	1.4	1.7
Af	1.24	1.72
Af	1.36	1.74
Af	1.38	1.82
Af	1.48	1.82
Af	1.54	1.82
Af	1.38	1.9
Af	1.56	2.08
Apf	1.14	1.78
Apf	1.2	1.86
Apf	1.18	1.96
Apf	1.3	1.96
Apf	1.26	2
Apf	1.28	2



Il est facile visuellement de tracer une ligne

Remarques

- Variables discriminantes :
  - Wing.Length − Ant.Length
  - Wing.Length/Ant.Length
- Et si on a plus de deux variables ?
- Validité des résultats sur la population totale ?



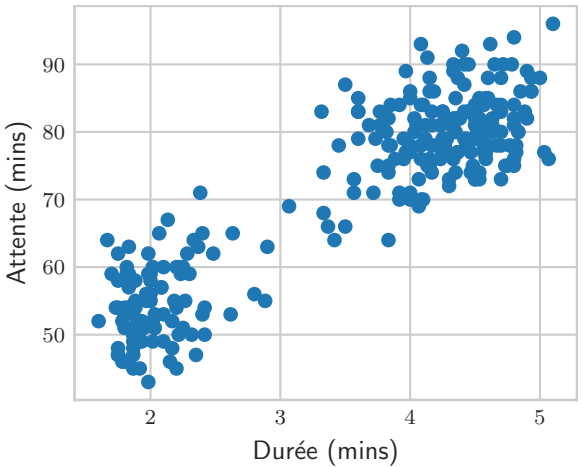
Les éruptions d'*Old faithful*

Jeu de données *Old faithful*

- Geyser situé à Yellowstone
- Chaque éruption est caractérisée par
  - sa durée
  - l'attente avant la prochaine éruption
- Jeu de données de 272 éruptions

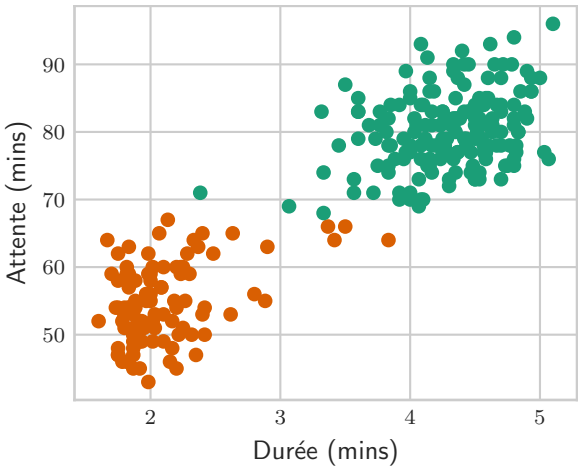
Recherche de groupes

- Deux types d'éruptions distinctes semblent se dessiner



Recherche de groupes

- Deux types d'éruptions distinctes semblent se dessiner



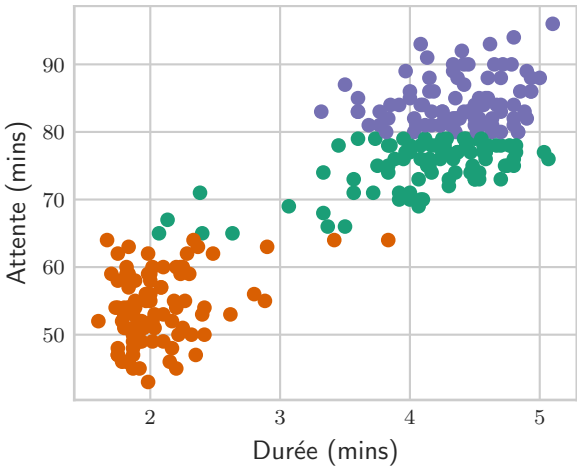
Problématiques

Problème de recherche d'une partition (ou *clustering*)

- Combien de groupes ?
- Comment savoir si la partition est bonne ?

Algorithmes possibles

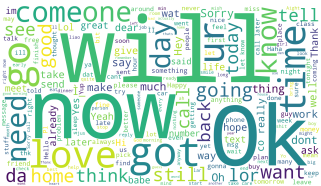
- *k*-means
- classification hiérarchique



Jeu de données *spams*

5572 SMS étiquetés *spam* ou *ham*

- Un méthode très utilisée : le filtrage bayésien
  - Repose sur la fréquence d'apparition de certains mots
- Algorithmes possibles
  - Bayésien naïf, QDA, FDA, arbre de décision



Cancer de la prostate

- Données : 97 patients
  - Niveau de gravité (log) : lcavol (difficile à calculer)
  - Poids de la prostate (log) : weight
  - Âge
  - Quantité d'hyperplasie prostatique (log) : lbph
  - Invasion de la vésicule séminale : svi
  - Pénétration capsulaire (log) : lcp
  - Score de Gleason
  - Pourcentage de score de Gleason 4 ou 5 : pgg45
- Problème : prédire *lcavol*
- Problème de régression

Quelques exemples : Marketing

- Analyse comportementale des consommateurs : ventes croisées, similarités de comportements, cartes de fidélité, ...
- Prédiction de réponse à un mailing ou à une opération de marketing
- Prédiction de la fuite des clients : quels sont les indices de comportements permettant de détecter la probabilité qu'un client a de quitter son fournisseur
- Recommandation
- ...

Quelques exemples : Détection de fraude

- Assurance, santé, banque, ...
- Utilisation des données historiques pour construire des modèles de comportements frauduleux et utiliser les techniques de data mining pour retrouver des instances similaires
- Exemples :
  - Assurances : détecter les groupes de personnes qui déclarent des accidents/vols pour les indemnités
  - Blanchiment d'argent : détecter les transactions suspectes (US Treasury's Financial Crimes Enforcement Network)
  - Assurance maladie : détecter les patients professionnels et les docteurs associés

Quelques exemples : *Web Mining*

- Organisation de sites Web :
  - Algorithmes de data mining appliqués aux journaux d'accès aux pages commerciales afin d'identifier les préférences et les comportements des clients et d'analyser les performances du marketing Web et l'organisation du site.
  - Exemple : GoogleAnalytics
- Outil de référencement de sites WEB
  - PageRank de Google
  - `www.igt.uni-stuttgart.de/eiserm/enseignement/google.pdf`

Quelques exemples : une application souvent citée

- Fouille de données sur les dépouillements de millions de tickets de caisse
- Mise en évidence par les magasins Wall-Mart d'une corrélation très forte entre l'achat de couches pour bébés et de bière le samedi après-midi
- Réorganisation des rayons : rapprochement des couches et des packs de bière
- Conséquence : augmentation des ventes

Positionnement de cette UV

- Domaine à la mode avec de nombreuses méthodes modernes
  - Ici, on se restreint à l'étude de quelques méthodes fondamentales
- Nombreuses technologies : Python, R, Hadoop/Spark, NoSQL, Julia
  - On utilisera le langage R.
  - Logiciel orienté statistique très répandu
- Nombreux sites où il est possible d'avoir des informations supplémentaires

Plan

Première partie : apprentissage non supervisé

- Introduction, types de données
- Méthodes exploratoires
- Représentation euclidienne
- ACP (1)
- ACP (2)
- Classification automatique (1)
- Classification automatique (2)

Deuxième partie : apprentissage supervisé

- Introduction, rappels
- Théorie de la décision
- Analyse discriminante
- Régression logistique
- Arbres de décision
- Sélection de modèle
- Régression linéaire multiple

Population–individus–caractéristiques

- Observation
  - Population composée d'individus
  - Les individus ont des caractéristiques
- Formalisation
  - La population est un ensemble Ω d'individu
  - Une caractéristique est une fonction :

$$X : \Omega \longrightarrow V_X$$

Tableaux individus–variables

- Population de n individus Ω
- Un nombre p de caractéristiques (attributs, variables, *features*) sur chaque individu : $X_1, \ldots, X_p$

	variable 1	$\cdots$	variable j	$\cdots$	variable p
individu 1	x_{11}		x_{1j}		x_{1p}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu i	x_{i1}		x_{ij}		x_{ip}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu n	x_{n1}		x_{nj}		x_{np}

Matrice X, n lignes, p colonnes

$$X_j(\text{individu } i) = x_{ij}$$

- Colonne : Variable fixée, tous les individus
- Ligne : Individu fixé, toutes ses caractéristiques

Type de variables

- **Quantitative** : $V_X \subset \mathbb{R}$
  - Discrète (V_X fini ou dénombrable) : résultat d'un comptage
  - Continue (V_X intervalle de $\mathbb{R}$) : poids
- **Qualitative** : V_X ensemble quelconque fini (modalités)
  - Nominale : couleur (rouge, bleu, vert)
  - Ordinale : taille (petit, moyen, grand), résultat d'une UV
- **Binaire** : variable qualitative particulière
  - Symétrique (Ex : féminin, masculin)
  - Ordre (Ex : Présence, Absence)

Retour sur le jeu de données *midge*

- Tableau individus–variables
  - 15 individus
  - 3 descripteurs
- Descripteurs :
- **Species** : Variable qualitative binaire symétrique
- **Ant.Length** : Variable quantitative continue
- **Wing.Length** : Variable quantitative continue

Species	Ant.Length	Wing.Length
Af	1.38	1.64
Af	1.4	1.7
Af	1.24	1.72
Af	1.36	1.74
Af	1.38	1.82
Af	1.48	1.82
Af	1.54	1.82
Af	1.38	1.9
Af	1.56	2.08
Apf	1.14	1.78
Apf	1.2	1.86
Apf	1.18	1.96
Apf	1.3	1.96
Apf	1.26	2
Apf	1.28	2

Exemples des félins : les variables

1. aspect du pelage	sans tâche tacheté rayé marbré
2. fourrure	poils ras poils longs
3. griffes	rétractiles non rétractiles
4. comportement prédateur	diurne diurne et nocturne nocturne
5. forme des oreilles	rondes et arrondies en pointe
6. os hyaoïde	présence absence
7. taille du garrot	$<$ 50cm entre 50cm et 70cm $>$ 70cm
8. poids de l'animal	$<$ 10kg entre 10kg et 80kg $>$ 80kg

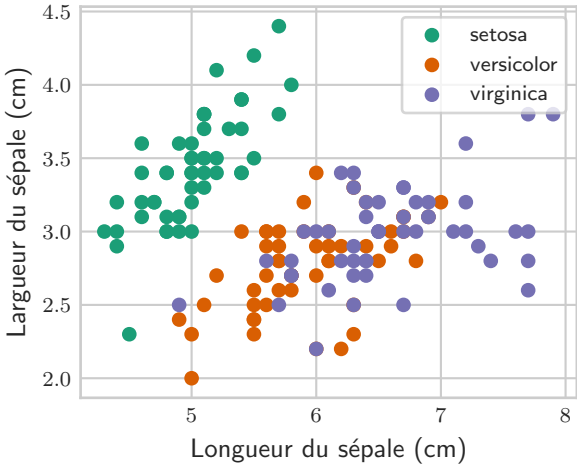
9. longueur du corps	$<$ 80cm entre 80cm et 150cm $>$ 150cm
10. longueur de la queue	petite moyenne longue
11. canines	très développées peu développées
12. type de proie	grosse grosse ou petite petite
13. monte aux arbres	monte aux arbres ne monte pas
14. chasse	à courre à l'affut
15. genre	panthera neofelis felis acinonyx

Iris de Fisher

- Exemple classique en statistique multidimensionnelle
- Proposé par Fisher pour illustrer les méthodes de discrimination
- 150 iris provenant de 3 familles différentes : Virginia, Versicolor et Setosa
- Longueur et la largeur du sépale et du pétale

	Sepal.Length	*Sepal.Width*	*Petal.Length*	*Petal.Width*	*Species*
1	5.1	3.5	1.4	0.2	*setosa*
2	4.9	3.0	1.4	0.2	*setosa*
3	4.7	3.2	1.3	0.2	*setosa*
4	4.6	3.1	1.5	0.2	*setosa*
...					
51	7.0	3.2	4.7	1.4	*versicolor*
52	6.4	3.2	4.5	1.5	*versicolor*
53	6.9	3.1	4.9	1.5	*versicolor*
54	5.5	2.3	4.0	1.3	*versicolor*
...					
101	6.3	3.3	6.0	2.5	*virginica*
102	5.8	2.7	5.1	1.9	*virginica*
103	7.1	3.0	5.9	2.1	*virginica*
104	6.3	2.9	5.6	1.8	*virginica*
...					
150	5.9	3.0	5.1	1.8	*virginica*

Diagramme de dispersion



Exemple de la reconnaissance des codes postaux

- Exemple classique *US zip codes*
- Ensemble de n images codées par une matrice de 28×28 pixels
- Tableau : $X \in \mathbb{R}^{n \times 784+1}$, n lignes, 785 colonnes

Bibliographie

- Bernard Flury. *A First Course in Multivariate Statistics*. Springer Science & Business Media, 1997.
- Gilbert Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.
- Jerome Friedman, Trevor Hastie et Robert Tibshirani. *The elements of statistical learning*. T. 1.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2006. isbn : 0387310738.
- Richard O Duda, Peter E Hart et David G Stork. *Pattern classification*. John Wiley & Sons, 2012.