Data Mining
(Concis et Pratique)

Julien Ah-Pine (julien.ah-pine@univ-lyon2.fr)

Université Lyon 2 - IUT Lumière

L3 CESTAT 2017/2018

Sommaire

1. Introduction et définitions
2. Exploration des données
3. Data Mining descriptif
4. Data Mining prédictif

Rappel du Sommaire

1. Introduction et définitions
   - Le Data Mining c'est quoi ?
   - Les différentes étapes en DM
   - Exemples d'applications réelles
   - Langage et librairies R pour le DM
   - Objectifs du cours

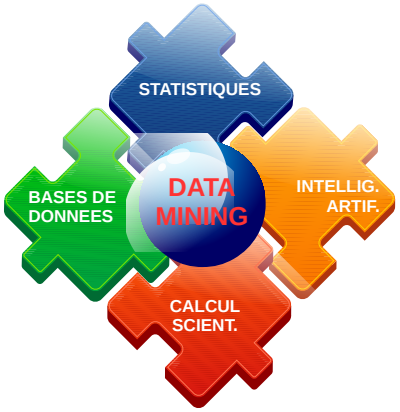
Une définition

- Data Mining/Knowledge Discovery = Fouille de données/Découverte de connaissances.
- Définition de Wikipédia (2017) :
  - L'**exploration de données**[1], connue aussi sous l'expression de **fouille de données**, forage de données, prospection de données, **data mining**, ou encore **extraction de connaissances à partir de données**, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.
  - Elle se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, l'intelligence artificielle ou l'informatique, pour construire des modèles à partir des données, c'est-à-dire trouver des structures intéressantes ou des motifs selon des critères fixés au préalable, et d'en extraire un maximum de connaissances.

1. Terme recommandé par la délégation générale à la langue française et aux langues de France (DGLFLF)

A la croisée de plusieurs disciplines

- Le DM tire profit des atouts et complémentarités de plusieurs disciplines :
  - Statistiques (ST)
  - Intelligence Artif. (IA)
  - Calcul scientif. (CS)
  - Base de données (BD)



- Comment est apparu le DM, où en sommes-nous et où allons-nous ?

Une rétrospective historique (Probabilités)

- Les **probabilités** comme origine des statistiques :
  - **17ème** siècle : Développement des calculs de probabilités avec Fermat, Pascal, Jacques Bernouilli (loi des grands nombres)...
  - **18ème** siècle : ... poursuivi par Laplace (théorème central limite), Condorcet, Bernouilli et Bayes et prémices des statistiques inférentielles avec le développement du concept d'erreurs vis à vis de la moyenne du point de vue probabiliste par Laplace.
  - 1ère moitié du **19ème** siècle :

    (ST) Développement de la méthode des **moindres carrés ordinaires** par Legendre et Gauss ; naissance de la science statistique comme discipline indépendante des probabilités ; ouverture du Conseil Supérieur de Statistique par Quetelet (démographe) en Belgique...

    (CS) Babbage invente la machine à calculer programmable (carte perforée) et (Ada) Lovelace écrit un programme pour calculer les nombres de Bernouilli.

Une rétrospective historique (Statistiques)

- **Statistiques "modernes"** et **intelligence artificielle** :
  - Fin du **19ème** siècle et début du **20ème** siècle :

    (ST) Définitions de concepts importants[2] en statistiques fréquentistes, tels que les plans d'expériences, la **vraisemblance**, les **tests d'hypothèses**, les intervalles de confiance par Fisher, Pearson, Neyman... sur des petits jeux de données.

    (ST) Fisher introduit l'**analyse discriminante linéaire** en 1936.

    (CS) En 1936 également, **Turing** propose sa "**machine**" qui est un modèle abstrait du fonctionnement des appareils mécaniques de calcul (concept formel de calculabilité/décidabilité -toute forme de calcul peut être représentée numériquement-).

2. Développement également de la théorie moderne des probabilités -théorie de la mesure- avec Borel.

Une rétrospective historique (Ordinateurs, IA, CS, BD)

- Avènement des **ordinateurs**, leurs impacts sur les statistiques et le début des disciplines sous-jacentes au DM :
  - Années **1940** :

    (CS) Avènement de l'informatique et des ordinateurs à la suite des travaux de Babbage, Turing, Von Neumann... avec l'impulsion industrielle d'IBM.

    (ST) Mise en oeuvre de méthodes statistiques avec des ordinateurs sur des petits jeux de données (30 individus et 10 variables).
  - Années **1950** :

    (IA) Turing publie en 1950 son article intitulé "Computing Machinery and Intelligence", prémices de l'**intelligence artificielle (IA)** cybernétique et de l'**apprentissage machine** (**machine learning**).

    (IA) McCullogh et Pitts présentent les premiers travaux sur les **réseaux de neuronnes** et Rosenblatt présente le modèle de **perceptron**.

    (ST) Cox introduit la **régression logistique** binomiale en 1958.

Une rétrospective historique (Ordinateurs, IA, CS, BD)

- Années **1960** :
  - (ST) Critiques du point de vue fréquentiste en statistique et renouveau des **statistiques bayésiennes** en incorporant aux modèles des informations subjectives (a priori) sous l'impulsion de Savage.
  - (ST) Benzécri et Escoffier introduisent l'**AFC** en 1962 et initient ainsi l'école française d'analyse des données.
  - (IA) Critique des réseaux de neuronnes (non-solvabilité des cas non-linéairement séparables comme XOR) et développement de l'**IA symbolique** (science cognitive, logique, bases de connaissance...) sous l'impulsion de Minsky et McCarthy (inventeur de LISP) du MIT.
  - (IA) Age d'or de l'IA symbolique : des ordinateurs résolvent des problèmes algébriques de mots, demontrent des théorèmes géométriques, apprennnent à parler l'anglais... Beaucoup d'engouement et d'investissement aux USA notamment.
  - (BD) Avènement des disques de données (par opposition aux cartes) et du concept de **base de données**.

Une rétrospective historique (Ordinateurs, IA, CS, BD)

- Années **1970 (Ko)** :
  - (IA) Premier hiver de l'IA : bcp de désilusions, promesses non tenues, arrêt des financements.
  - (IA) Développement par Werbos de l'**algorithme de rétro-propagation** pour l'apprentissage de réseaux de neuronnes multicouches et solution apportée aux cas non-linéairement séparables.
  - (ST) Du point de vue statistique/informatique : traitement d'un plus grand nombre de données et notamment de variables, développement des statistiques multidimensionnelles et de l'**analyse de données**.
  - (ST) Nelder et Wedderburn formalisent le **modèle linéaire généralisé** dans leur livre "Generalized Linear Models" en 1972.
  - (BD) Début des **bases de données relationnelles** (BDR) et **ordinateurs de bureau** : développement par Codd (IBM) des BDR décrivant une approche pour stocker et requêter des données à partir d'une base. Le langage **SQL** apparaît fin des années 70.

Une rétrospective historique (Ordinateurs, IA, CS, BD)

- Années **1980 (Mo)** :
  - (IA) Succès commercial des **systèmes experts** qui donnent un renouveau à l'IA symbolique (bases de connaissance) sous l'impulsion des travaux de Feigenbaum. Exemple : Mycin permet de diagnostiquer les maladies infectieuses du sang.
  - (IA) Quinlan introduit les **arbres de décision** en 1986.
  - (IA) Renouveau également de l'IA cybernétique suite aux travaux de Werbos, Rumelhart (rétro-propagation) et Hopfield (reseaux de neuronnes récurrent).
  - (CS) Ce nouvel élan est lié à l'algorithme de rétro-propagation mais également aux débuts du **calcul parallèle et distribué**.
  - (ST) Du côté statistique, se développent les méthodes non-paramétriques (on tente de s'affranchir du biais inductif).
  - (BD) Les données sont stockées sur plusieurs ordinateurs de bureau. Les besoins en analyses persistent. Les **entrepôt de données** introduits par Inmon émergent à la fin des années 80.
  - (DM) Au même moment, le terme "**Knowledge Discovery in Databases**" est utilisé pour la 1ère fois par Piatetsky-Shapiro.

Une rétrospective historique (Data Mining)

- Avènement du **Data Mining** (DM) :
  - Années **1990 (Go)** :
    - (DM) Le terme **Data Mining** apparaît au sein de la communauté BD pour caractériser les besoins en aide à la décision à partir de données. Le domaine est stimulé par des problémes opérationnels au sein d'entreprises. Début du **marketing quantitatif** et de la gestion de la relation client (**CRM**).
    - (DM) Agrawal and Srikant introduisent l'**algorithme apriori** pour la recherche de règles d'association dans des BD en 1994.
    - (IA) Deuxième hiver de l'IA : le développement et la puissance des ordinateurs de bureau (Apple et IBM) surpasse celle des ordinateurs programmés en LISP qui deviennent chers à maintenir. Nouveau gel des financements de l'industrie IA.
    - (www) Le **web** prend naissance et se développe rapidement : on passe de 26 sites web en 1992 à prés de 10 millions de sites en 1999.
    - (ST) Approches innovantes en statistiques : **SVM** de Vapnik et Cortes, **Boosting** de Freund et Schapire, **Bagging** et Arcing de Breiman, **LASSO** de Tibshirani, **GAM** de Hastie, Tibshirani...

Une rétrospective historique (Data science et Big data)

- **Data science** et **big data**, suites logiques des statistiques et du DM :
  - Années **2000 (To)** :

(www) Développement du **web social** et des **smartphones** : capacités accrues de stockage et d'échange de fichiers multimédia, expansion des activités économiques du e-commerce (recommandation, analyse de traces...). Tout ceci provoque un changement de paradigme : les **données sont non-structurées, complexes** et peuvent être de très **grande dimension** (texte, image...), les outils informatiques et les modèles d'analyse (statistiques, IA) doivent s'adapter à ces caractéristiques.

(ST) Développement de la branche **apprentissage statistique**. Importance des concepts **biais et variance** : ce qui importe c'est la qualité des connaissances découvertes et/ou des prédictions obtenues et non pas l'ajustement d'un modèle à des données.

(ST) Le livre de référence "The Elements of Statistical Learning" d'Hastie, Tibshirani et Friedman sort en 2001.

Une rétrospective historique (Data science et Big data)

(ST) Avènement du **Data Science** en 2001 par Cleveland dans un article intitulé "Data Science : An Action Plan for Expanding the Technical Areas of the Field of Statistics" montrant la nécessité d'intégrer les outils de calcul scientifique dans le champs d'expertises des statistiques. Le data science peut être vu comme les statistiques avec des outils modernes de calcul scientifique.

(BD) Introduction des bases de données **NoSQL** comme Bigtable de Google en 2004, pour le stockage distribué des données volumineuses et non-structuées.

(CS) Débuts du calcul distribué comme **MapReduce** de Google pour effectuer des requêtes et calculs sur des données non-structurées et distribués.

- Années **2010 (Po)** :

(DM) Avènement du **Big Data** (depuis la fin des années 2000) pour le stockage, le requêtage et l'analyse de données massives mais en mettant l'accent sur les 4V : Volume, Variété, Vélocité et Véracité. Le big data peut être vu comme du DM avec des technologies modernes de stockage et de calculs distribués.

Une rétrospective historique (Data science et Big data)

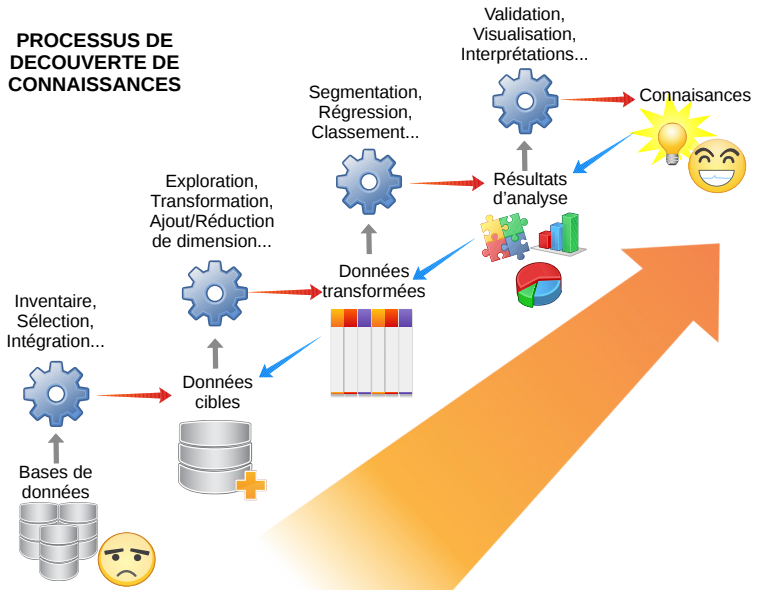
(BD)(CS) Développement à partir de 2011, de l'écosystème libre **Hadoop** et du calcul distribué MapReduce pour répondre aux besoins croissants de l'analyse de données massives.

(BD)(CS) Développement à partir de 2014 de l'outil libre **Spark**, qui va plus loin que le MapReduce classique en permettant des calculs distribués et itératifs nécessaires aux algorithmes de DM.

(IA) Retour en grâce des réseaux de neuronnes : la puissance des serveurs de calcul permettent de mettre en oeuvre ces méthodes sur des données très massives. Le **deep learning** est la méthode de DM la plus en vogue de ces dernières années : très performant pour de nombreux problèmes en image, texte, son...

(IA) L'algorithme AlphaGO de Google basé sur du deep learning bat un champion humain du jeu de Go en 2015 et en 2016.

Schéma



Inventaire, Sélection et Intégration des données

- En amont, il faut clairement définir et/ou identifier :
  - le phénomène : "qu'est-ce que j'étudie ?"
  - la tâche : "qu'est ce que je veux faire (découvrir ? prédire ?) ?"
  - la population : "quelles observations [3] je vais utiliser ?"
  - ▷ Cela demande une bonne connaissance "métier" càd le contexte et les données sur lesquels porte l'étude.
- Ensuite, il faut **rassembler des informations sur le phénomène** :
  - Faire l'inventaire des variables [4] existantes.
  - Sélectionner les variables en relation avec ma tâche.
  - Intégrer ces variables pouvant provenir de plusieurs sources/BD au sein d'un même jeu de données [5].
- Cette partie utilise des compétences en BD/DW/ETL mais nous n'aborderons pas ces aspects.
- Nous supposons que le jeu de données à été construit et nous nous consacrerons en particulier à l'analyse.

3. ou individus, ou objets, ou entités.
4. ou descripteurs ou attributs ou features.
5. Voire d'un entrepôt de données si besoin

Exploration, Transformation des données

- Il faut faire "**connaissance**" avec les données pour commencer à appréhender le phénomène à l'étude !
- Pour cela, les outils statistiques pertinents sont :
  - Les statistiques descriptives univariées (tendance centrale, de dispersion...) pour :
    - ⋆ appréhender les caractéristiques simples des variables,
  - Les graphiques (histogrammes, diagrammes en bâtons, camenberts, boîtes à moustache) pour :
    - ⋆ visualiser les caractéritiques simples, la distribution, des variables ;
  - Les statistiques descriptives bivariées (mesures de corrélation, d'association...) pour,
    - ⋆ identifier les variables qui sont corrélées ;
  - Les statistiques exploratoires multidimensionnelles (ACP, AFC, ACM...) pour :
    - ⋆ visualiser de façon synthétique les grandes tendances.

Exploration, Transformation des données (suite)

- Identifier, gérer les **données manquantes** :
  - Si le jeu de données et suffisamment grand, on pourra :
    - ⋆ enlever toute obs. comportant des données manquantes.
  - Si le jeu de données n'est pas suffisamment grand, on pourra :
    - ⋆ remplacer une donnée manquante par une moyenne,
    - ⋆ utiliser une méthode d'imputation plus élaborée.
- Identifier et traiter les **observations aberrantes** :
  - L'étude d'une boîte à moustache permet d'identifier les obs. ayant des valeurs extrêmes.
  - Il faut étudier ces obs. et les enlever si elles peuvent causer un biais dans l'estimation des méthodes utilisées :
    - ⋆ cas d'une obs. hors-norme,
    - ⋆ cas d'une obs. avec des erreurs de mesures.

Exploration, Transformation des données (suite)

- Transformer une variable quanti. en une variable quanti. :
  - Lorsque les variables son exprimées dans différentes échelles, celles-ci peuvent créer un biais dans les méthodes utilisées. Il est important dans ce cas de normaliser les variables en centrant et réduisant par exemple.
  - Lorsque les variables ne satisfont pas aux hypothèses d'un modèle utilisé. Dans ce cas, transformer la variable par une fonction permet de la ramener dans les hypothèses requises. Ex : dans le modèle linéaire gaussien, il est requis que les variables suivent des lois normales.
- Transformer une variable quanti. en une variable quali. :
  - Il est utilie de transformer une variable quanti. en quali., lorsque cela facilite l'interprétation ou lorsque cela est requis par la méthode utilisée. On parle de discrétisation. Elle peut être manuelle ou automatique.
- Transformer une variable quali. en une variable quali. :
  - Lorsque l'on a besoin de recoder les modalités (exprimées par du texte par exemple) d'une variable quali. en un autre format (code numérique).
  - Lorsque l'on souhaite regrouper plusieurs modalités en une seule.

Exploration, Transformation des données (suite)

- Transformer une variable quali. en une variable quanti. :
  - Cela est souvent pratiqué par l'école anglo-saxonne où l'on ramène une (ou plusieurs) variable quali. à un score numérique. Dans ce cas, les techniques factorielles telles que l'ACM peuvent être utilisées.
- Le problème plus général des **données mixtes** :
  - De nombreuses méthodes ne permettent pas de traiter simultanément des variables quanti. et quali. Dans ce cas, il est nécessaire de transformer les quanti. en quali. ou les quali. en quanti.

Segmentation, Régression, Classement

- Cette étape représente la **partie analyse** de la procédure de DM où l'objectif est l'**extraction des connaissances**.
- On peut faire la distinction entre deux types d'analyse :
  - L'**analyse descriptive** ou **apprentissage non-supervisé**. Le but est de mettre en évidence des régularités, tendances, corrélations. . . au sein des données afin d'obtenir des connaissances "cachées" sur le phénoméne à l'étude.
    On distingue (au moins) deux types de tâche :
    - ⋆ La **classification automatique** qui vise à partitionner la population en plusieurs classes. Chaque classe est un groupe homogène d'obs. qui sont plus similaires entre elles qu'elles ne le sont avec les obs. des autres groupes. Le but est aussi de savoir quelles sont les variables discriminantes de chaque classe.
    - ⋆ La **recherche de règles d'association** qui tente de déterminer quelles valeurs de quelles variables vont très souvent ensembles avec *quelles valeurs de quelles autres variables*. Les résultats obtenus sont des règles de type "si conditions alors *résultats*".

Segmentation, Régression, Classement (suite)

- L'**analyse prédictive** ou **apprentissage supervisé**. Dans ce cas et contrairement à l'approche descriptive, il existe une variable cible (d'où le terme supervisé) et l'objectif est d'estimer une fonction permettant de prédire pour une observation donnée la bonne valeur de la variable cible.
  Il existe deux types de problèmes :
  - ⋆ Les problèmes de **régression** : la variable cible est alors quantitative. Ex : prédiction des recettes d'un film étant donné les acteurs, les producteurs, le budget. . .
  - ⋆ Les problèmes de **classement** [6] : la variable cible est dans ce cas discrète. Ex : prédiction de l'avis général sur un film entre "nul, pas terrible, moyen, bon, super" étant donné les acteurs, les producteurs, le budget. . .

6. ou catégorisation ou classification supervisé.

Validation, Visualisation et Interprétation des résultats

- L'**étape de validation** est importante à plusieurs égards :
  - Il existe plusieurs méthodes pour les différentes tâches citées précédemment. Cette étape sert alors à déterminer laquelle des méthodes testées donne les meilleures performances. Il existe des **protocoles et critères pour comparer les méthodes** entre elles (cf ci-dessous).
  - Cette étape permet aussi d'avoir un retour expert sur la méthode retenue. Est-ce que les résultats permettent véritablement d'extraire des connaissances nouvelles ? C'est la **dimension "humaine" de la validation** qui est tout aussi primordiale.
- Pour la **classification automatique** on distingue :
  - **Validation externe** : on dispose d'une partition de référence et on compare le résultat de la méthode avec cette partition selon plusieurs critères (indice de Rand corrigé par ex.).
  - **Validation interne** : on mesure l'homogénéité des classes et de la partition obtenue à partir de plusieurs critères (intertie intra-classe, inter-classe par ex.).

Validation, Visualisation et Interprétation des résultats (suite)

- Pour les **règles d'association**, il existe plusieurs critères pour mesurer la pertinence d'une règle extraite (support, confiance, lift par ex.)
- En **apprentissage supervisé** (régression et classement), ce qui est important c'est de pouvoir prédire correctement sur des données non encore observées. Pour cela, on a recours classiquement à de la **validation croisée** afin d'avoir une estimation de l'**erreur en généralisation**.
  Il existe plusieurs mesures d'erreurs selon le problème considéré :
  - Pour un problème de **régression** : erreur quadratique moyenne (MSE) ou erreur moyenne en valeur absolue (MAE)...
  - Pour un problème de **classement** : taux d'erreur, précision, rappel, courbe ROC...

Validation, Visualisation et Interprétation des résultats (suite)

- Pour chaque tâche, chaque type de critère d'évalutaion, il peut exister plusieurs types de **graphique** permettant de **visualiser** les performances des méthodes et de les comparer entre elles.
- Il est important de comprendre les protocoles expérimentaux et les critères d'évaluation, afin d'avoir une **bonne interprétation des résultats d'expériences** dans le but de choisir la bonne méthode.

Quelques applications

- Vente, marketing :
  - Gestion de la relation client (ex : score d'appétence -achat-)
  - Segmentation de la clientèle...
- Banque, finance, assurance :
  - Détection de fraudes (ex : comportements atypiques),
  - Score de risque (ex : attribution ou pas de crédit)...
- Médecine, industrie pharmaceutique :
  - Réponse d'un patient vis à vis d'un traitement,
  - Identification de facteurs de risques...
- Génome humain, bio-informatique :
  - Relations entre l'ADN et des maladies,
  - Détection de rôles joués par des gènes...

▷ Le DM peut s'appliquer à tout phénomène dont on peut mesurer des observations (stockables dans une BD) et qu'on souhaite appréhender les caractéristiques et/ou qu'on souhaite prévoir le comportement.

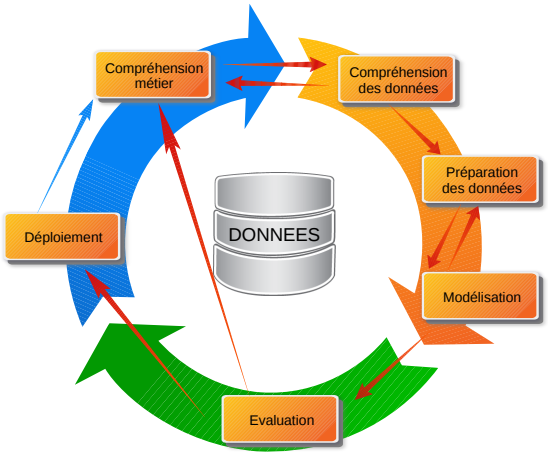
Industries qui utilisent le DM

- Source : kdnuggets.com [7] (site de G. Piatetsky-Shapiro)
- Industries - Fields where you applied Analytics, Data Mining, Data Science in 2016 ?
- The most popular areas were
  - CRM/Consumer analytics, still n. 1 at 16.3% but down from 18.6% share in 2015.
  - Finance, 15.0% (almost the same as in 2015)
  - Banking, 13.4% (slightly down)
  - Advertising, 12.0% (up 35% from 2015)
  - Science, 12.0% (almost the same)
  - Health care, 12.0% (11% down)
  - ...

7. http://www.kdnuggets.com/2016/12/poll-analytics-data-mining-data-science-applied-2016.html

CRISP-DM

- Cross Industry Standard Process for Data Mining [8] : procédure communément utilisée par les data miner (DM) en entreprise.



8. Fondé en 1996 par les entreprises NCR, SPSS, Daimer-Benz.

Remarques supplémentaires sur le déploiement

- Dans un cadre opérationnel, le **déploiement** fait suite à la validation et à cette étape le DM prend concrètement une **dimension business**.
- Prenons l'exemple d'une banque. L'équipe DM a mis en place une méthode de scoring pour une nouvelle offre de crédit. Le déploiement va consister à **diffuser auprès des opérationels** (les banquiers) soit la méthode (via un logiciel), soit les résultats de la méthode (via un rapport, une feuile de tabeur, une table dans une base de données...).
- L'équipe DM doit **présenter la méthode aux opérationels de façon accessible**, en évitant les détails techniques et en exposant : le but recherché, le principe de l'outil, son fonctionnement mais aussi ses limites. C'est la dimension aide à la décision du DM.
- Il est ensuite important de **suivre l'utilisation et les performances de la méthode**. Est-ce que la méthode de scoring est performante ? Est-ce que les clients à qui on a octroyé un crédit le remboursent véritablement ? Il s'agit ici de reporting qui permet d'enrichir la compréhension métier et on obtient ainsi un **cercle vertueux**.

Quelques outils pour le DM

- Ceux qui sont propriétaires et payants :
  - SAS
  - SPSS
  - SPAD
  - Knime
  - ...
- Ceux qui sont **open source et/ou gratuits** :
  - R
  - Python
  - Weka
  - Daiku
  - ...

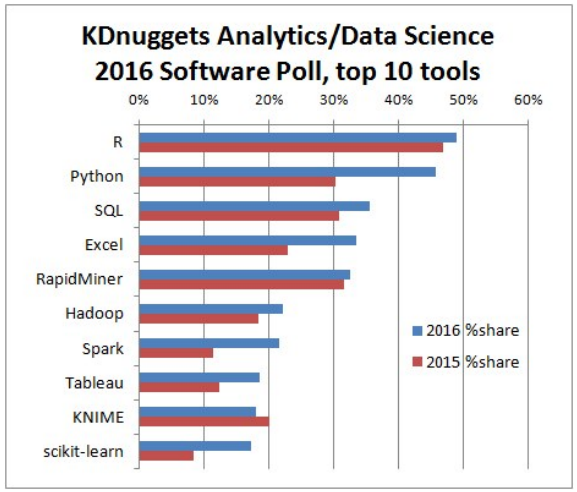
Pourquoi le langage R ?

- Le langage R [9] est, avec Python, l'un des deux principaux langages pour le DM/Data Science.
- Communauté active avec une conférence annuelle : useR !
- Beaucoup de librairies : Comprehensive R Archive Network [10].
- Une revue scientifique : The R Journal [11].
- Un IDE de référence qui est libre également : Rstudio [12].
- Plusieurs ressources "cheatsheets" disponibles [13].

9. https://www.r-project.org/
10. https://cran.r-project.org/
11. https://journal.r-project.org/
12. https://www.rstudio.com/
13. https://www.rstudio.com/resources/cheatsheets/, http://www.rdatamining.com/docs/RDataMining-reference-card.pdf?attredirects=0&d=1, http://www.thinkr.fr/le-blog/

R est en vogue !

- Source kdnuggets.com [14].



14. http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Objectifs

- Aborder chaque étape du DM à partir d'exemples réels.
- Rappeler/introduire des méthodes classiques de façon **concise**.
- Savoir mettre en oeuvre ces méthodes avec le langage R.
- Savoir mettre en place le protocole expérimental adéquat.
- Savoir interpréter les résultats.
- ▷ Etre opérationnel !

Organisation des séances

- Introduction brève des méthodes ∼30 minutes.
- Déroulement des TP ∼60-70 minutes.
- Correction et interprétation des résultats ∼15 minutes.

- La dernière séance programmée sera l'examen :
  - Cas à analyser.
  - Code R et interprétations à restituer.

- Les supports de cours et sujets de TP sont disponibles au fil de l'eau sur mon site eric.univ-lyon2.fr/~jahpine/.

Rappel du Sommaire

2 Exploration des données
- Commandes utiles et imports de données
- Statistiques descriptives univariées et bivariées
- Quelques tests statistiques
- Manipulation et transformation de données

Commandes utiles

- Gestion des librairies :
  - Liste des librairies chargées : `search`
  - Installation : `install.packages`
  - Chargement : `library`
  - L'utilisation de l'onglet `Packages` de Rstudio est très pratique !
- Gestion des variables :
  - Sauvegarde de variables : `save` (fichier `Rdata`)
  - Chargement d'un ensemble de variables : `load`
  - Supprimer une variable de l'envir. de travail : `rm`
- Export des données :
  - Format texte CSV : `write.csv`
  - Format Excel : `write.xlsx` [`xlsx`]
- Commentaires : #
- Remarque sur la notation : `commande` [`librairie`] (le cas échéant)
- Ressources en-ligne :
  - http://www.rdatamining.com/docs/introduction-to-data-mining-with-r-and-data-import-export-in-r

Import de données

- Format natif de R (`.Rdata`) :
  - `load`
- Format texte CSV (`.csv`) :
  - `read.table`
  - `read.csv`
- Format Excel (`.xlsx`) :
  - `read.xlsx` [`xlsx`]
- Format SAS (`.xpt`), SPSS (`.sav`) :
  - `sasxport.get` [`Hmisc`]
  - `spss.get` [`Hmisc`]
- Ressources en-ligne :
  - http://www.statmethods.net/input/importingdata.html
  - http://www.rdatamining.com/docs/introduction-to-data-mining-with-r-and-data-import-export-in-r

Tendances centrales

- Principe : résumer la distribution d'une variable en un nombre.
- Variable quantitative :
  - Moyenne (`mean`)
  - Médiane (`median`)
- Variable qualitative :
  - Mode (`summary`)
- Commandes donnant plusieurs indicateurs :
  - `summary`
  - `describe` [`Hmisc`]
  - `describe` [`psych`]
- Ressources en-ligne :
  - http://www.statmethods.net/stats/descriptives.html
  - http://www.rdatamining.com/docs/data-exploration-and-visualization-with-r

Indicateurs de dispersion/répartition

- Principe : mesurer la dispersion/concentration d'une variable.
- Variable quantitative :
  - Variance (`var`)
  - Ecart-type (`sd`)
  - Etendue (`range`)
  - Quartiles (`quantile`)
- Variable quanlitative :
  - Fréquences (`table`)
- Ressources en-ligne :
  - http://www.statmethods.net/stats/descriptives.html
  - http://www.rdatamining.com/docs/data-exploration-and-visualization-with-r

Corrélations et associations

- Principe : mesurer une relation de dépendance entre deux variables.
- Variables quanti./quanti. :
  - Covariance (`cov`)
  - Corrélation de Bravais-Pearson (`cor`)
- Variables quali./quali. :
  - Table de contingence (`table`)
  - Coefficient Chi2 (`chisq.test(table)`)
  - Coefficient Phi [15] (`Phi` [DescTools])
  - Coefficient de Tchuprow [16] (`TschuprowT` [DescTools])
- Variables quanti./quali. :
  - Statistiques univariées d'une variable quanti. par groupe de modalités d'une variable quali. (`aggregate(quanti~quali,summary)`)
  - Rapport de corrélation [17] (`eta2` [`BioStatR`])

15. https://en.wikipedia.org/wiki/Phi_coefficient
16. https://en.wikipedia.org/wiki/Tschuprow's_T
17. https://en.wikipedia.org/wiki/Correlation_ratio

Graphiques de statistiques univariées

- Variable quantitative :
  - Boîte à moustache (`boxplot`)
  - Histogramme (`hist`)
  - Estimation à noyau de la densité (`plot(density)`)
- Variable qualitative :
  - Camembert (`pie`)
  - Diagramme à bâtons (`barplot(table)`)
- Variables quanti./quali. :
  - Boîte à moustache par modalité (`boxplot(quanti~quali)`)

Graphiques de statistiques bivariées

- Variables quanti./quanti. :
  - Nuage de points [18] entre deux variables (`plot`)
  - Nuage de points entre plusieurs couples de variables (`pairs`)
- Variables quali./quali. :
  - Table de contingence "graphique" (`balloonplot(table)` [`gplots`])
  - Table de contingence des résidus du test de Chi2 "graphique" (`assoc(table)` [`vcd`])
- Variables quantis./quali. :
  - Coordonnées parallèles de plusieurs variables quanti. par modalité d'une variable quali. (`parcoord` [`MASS`])
- Références en-ligne :
  - http://www.statmethods.net/graphs/scatterplot.html
  - http://www.statmethods.net/advgraphs/mosaic.html
  - http://www.rdatamining.com/docs/data-exploration-and-visualization-with-r

18. https://en.wikipedia.org/wiki/Scatter_plot

Tests statistiques

- Tests d'adéquation à une loi donnée pour variable quantitative :
  - Loi normale : test de Shapiro-Wilk [19] (`shapiro.test`)
  - Loi quelconque : test de Kolmogorov-Smirnov [20] (`ks.test`)
- Tests de corrélation entre deux variables quantitatives :
  - Tests de Pearson ou Kendall ou Spearman (`cor.test`)
- Test d'indépendance entre entre deux variables qualitatives :
  - Tests du Chi2 (`chisq.test`)
- Test de comparaison de populations :
  - ANOVA à 1 facteur [21] (`aov`)
  - Test de Kruskal-Wallis [22] (`kruskal.test`)

19. https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
20. https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test
21. https://en.wikipedia.org/wiki/One-way_analysis_of_variance
22. https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance

Détection des points aberrants

- Par analyse graphique :
  - Un point extrême d'une boîte à moustache peut être considéré comme aberrant si :
    - si sa valeur absole dépasse la moyenne + 3 fois l'écart type,
    - si sa valeur est au-dessus de Q3+1.5(Q3-Q1) ou au-dessous de Q1-1.5(Q3-Q1) où Q1 et Q3 sont les 1er et 3ème quartiles (règle de Tukey).
- Par test statistique :
  - Test de Dixon [23] (`dixon.test` [`outliers`])
  - Test de Grubbs [24] (`grubbs.test` [`outliers`])

23. https://en.wikipedia.org/wiki/Dixon%27s_Q_test
24. https://en.wikipedia.org/wiki/Grubbs%27_test_for_outliers

Manipulation de données avec `dplyr`

- Introduction à la librairie [`dpylr`].
- Principe : Sélectionner, croiser des variables et automatiser certains calculs/prétraitements dans une grammaire lisible.
- Sélection de variables (colonnes d'un `data.frame`) :
  - Extraction de variables (`select`).
  - Matching sur nom des variabes (`starts_with`, `contains`, `ends_with`, ...)
- Sélection d'individus (lignes d'un `data.frame`) :
  - Extraction d'individus (`filter` + conditions logiques)
  - Détection des doublons (`distinct`)
  - Echantillonage (`sample_frac`,`sample_n`)
- Ressources en ligne :
  - https://www.rstudio.com/wp-content/uploads/2016/01/data-wrangling-french.pdf

Résumés de données (calculs d'indicateurs) avec `dplyr`

- Principe : à partir d'un sous-ensemble de lignes, calculer un indicateur.
- Quelques actions possibles (liste non exhaustive) :
  - Action sur une variable (`summarize`).
  - Action sur toutes les variables (`summarize_all`)
  - Action de dénombrement sur une variable (qualitative) (`count`)
- Quelques indicateurs classiques : `min, max, mean, sd` ...
- On peut définir son propre indicateur par une fonction !

Construction de nouvelles variables avec `dplyr`

- Principe : transformer une ou plusieurs variables afin d'obtenir une nouvelle variable.
- Quelques actions possibles (liste non exhaustive) :
  - Action sur une variable (`mutate`)
  - Action sur toutes les variables (`mutate_all`)
- Quelques transformations classiques : `pmin, pmax` ...
- On peut définir sa propre transformation par une fonction !
- Utilisation pour le recodage des variables :
  - Variable Quali. $\rightarrow$ Quali. (`recode_factor`)
  - Variables Quanti.+Quali. $\rightarrow$ Quali. (`mutate` + `case_when` + conditions logiques ...)

Regroupement de données et %>% avec dplyr

- Principe du regroupement de données :
  analyser une variables qualitative ou le croisement de plusieurs variables qualitatives.
  - Regroupement des lignes selon les modalités de variable(s) qualitative(s) (`group_by`)
- Principe du "pipe" %>% :
  enchaîner plusieurs opérations, l'input d'une opération étant l'output de l'opération précédente.
- Exemple :
  `iris %>% group_by(Species) %>% summarize(m=mean(Sepal.Length))`
- Le %>% permet d'avoir une lecture simple du "workflow".
- Il existe d'autres opérateurs de ce type ([magrittr]).

Discrétisation de variables quantitatives

- Principe : transformer une variable quantitive en une variable qualitative en définissant manuellement ou automatiquement des intervalles.
- Nous appliquerons la discrétisation manuelle :
  - Découpage selon des points définissant des intervalles (`cut`)
- Il existe des méthodes plus avancées. Vous pouvez utiliser par exemple la librairie [smbinning] avec la ressource en-ligne suivante
  http://blog.revolutionanalytics.com/2015/03/r-package-smbinning-optimal-binning-for-scoring-modeling.html.

Gestion des données manquantes

- Les données manquantes en R sont symboliquement marquées par `NA` ("Non Attributed").
- Quelques commandes de base pour gérer les données manquantes :
  - Test si présence d'une donnée manquante (`is.na`).
  - Test de lignes complètes (`complete.cases`).
- Il existe plusieurs méthodes d'imputation mais nous utiliserons des approches classiques.
- En particulier, nous mettons en oeuvre les outils proposés par [dplyr] cités précédemment.
- Pour aller plus loin, il existe la librairie [mince] et la ressource en ligne suivante par exemple
  https://datascienceplus.com/imputing-missing-data-with-r-mice-package/.

Rappel du Sommaire

3 Data Mining descriptif
- Analyse de données
- Classification automatique
- Règles d'associations

Analyse descriptive

- Nous disposions d'une table de données avec n individus et p variables que l'on notera $\mathbf{x}^1, \ldots, \mathbf{x}^p$.
- L'objectif est d'explorer ces données par des méthodes statistiques afin d'en extraire/découvrir des informations pertinentes. On parle d'apprentissage non-supervisé car il ne s'agit pas de modéliser une variable en particulier.
- Les méthodes de réduction de dimension (ACP, AFC, ACM) en analyse de données permettent de représenter les données dans des espaces réduits et ce faisant, elles mettent en valeur les tendances principales en déterminant les associations/oppositions entre individus et variables de façon simultanée.
- Les méthodes de classification automatique agissent praticulièrement au niveau des individus. Elles viennent souvent compléter les méthodes de réduction de dimension en déterminant de façon claire les contours de groupes homogènes conduisant à une typologie de la population.
- Dans le cas de données qualitatives, l'extraction de règles d'association permet de déterminer des groupes de modalités qui

Méthodes de réduction de dimension

- Principe : représenter de façon "efficace" et "intelligente" l'information contenue dans une table au travers de graphiques présentant les données dans une espace géométrique de dimension faible.
- Concepts sous-jacents :
  - Le terme information est ici de nature géométrique et repose pincipalement sur la notion de variance d'un nuage de points : de combien en moyenne les points sont distants du barycentre.
  - Les notions de distances (métriques) sont donc fondamentales. On détermine un sous-espace vectoriel de faible dimension au sein duquel le nuage projeté est le moins déformé possible.
  - On montre que déterminer ce sous-espace vectoriel revient à déterminer la décomposition spectrale (recherche de valeurs et vecteurs propres) d'une matrice carrée symétrique définie positive.

Méthodes de base

- Les tables de données peuvent être de différentes natures et selon le type de variables on a une méthode particulière.
- Si les variables sont quantitatives on parle d'Analyse en Composantes Principales (ACP).
- Si on étudie le croisement de deux variables qualitatives on parle d'Analyse Factorielle des Correspondances Simple (AFC).
- Si les variables (plus de deux) sont qualitatives, on parle d'Analyse (Factorielle) des Correspondances Multiples (ACM).
- Si les variables sont un mélange de quanti. et quali. on parle d'Analyse Factorielle de Données Mixtes (AFDM).

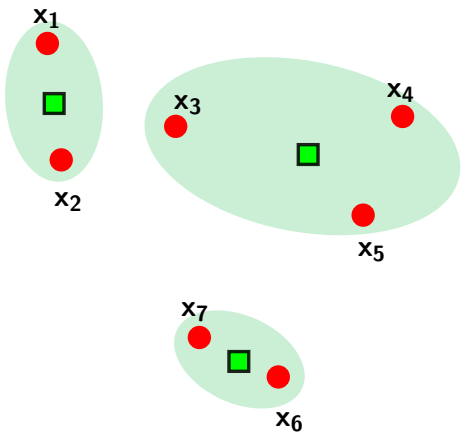
Outils en R et ressources en ligne

- Il existe plusieurs outils en R pour faire de l'ADD :
  - Commandes de base (`svd`, `eigen`, `prcomp`, `princomp`)
  - Plusieurs librairies (`[MASS]`, `[sca]`, ...)
  - Ressource en ligne : https://cran.r-project.org/web/views/Multivariate.html (sections "Projection methods" ou "Correspondance analysis")
- Nous utiliserons principalement la librairie (française) `[FactoMineR]`.
- Ressources en ligne :
  - Site de la librairie : http://factominer.free.fr/.
  - Article : http://factominer.free.fr/docs/article_FactoMineR.pdf.
- Livre associé : F. Husson, S. Lê, J. Pages, "Analyse de données avec R", Presses Universitaires de Rennes.

L'algorithme des k-moyennes (k-means)

- Principe : Affecter un individu à la classe dont le barycentre est le plus proche. Mettre à jour l'affection de cet individu et le barycentre de son ancienne et nouvelle classe. Itérer ces opérations pour tous les individus et jusqu'à convergence.
- Remarques importantes :
  - On raisonne dans un espace euclidien et les variables sont donc continues.
  - La mesure de proximité utilisée est la distance euclidienne (avec poids uniforme sur les variables).
  - Le barycentre est le vecteur moyen et par défaut les individus ont tous un poids uniforme.
  - D'un point de vue optimisation, la procédure diminue la variance intra-classe et augmente la variance inter-classes à chaque itération.
  - Complexité en $O(n)$ (si p et k sont petits).
  - Cet algorithme détecte des classes qui sont de forme sphérique dans leur représentation géométrique.
- Commande `kmeans`.

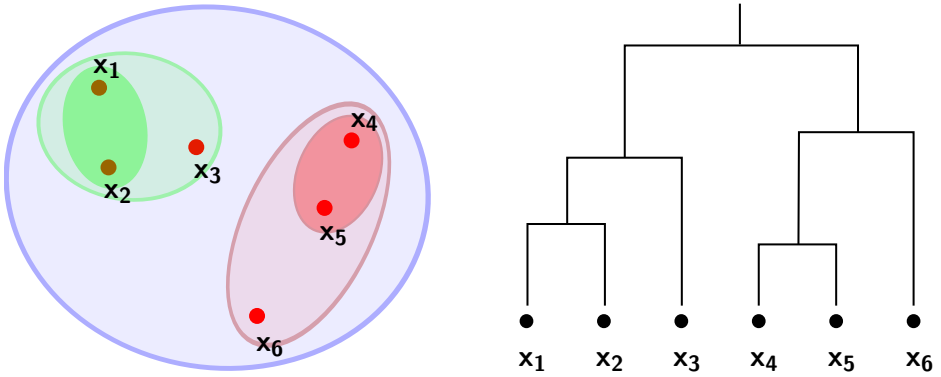
L'algorithme des k-moyennes (k-means)



La classification ascendante hiérarchique (CAH)

- Principe : Regrouper itérativement les deux classes les plus proches jusqu'à ce que tous les individus soient regroupés en une seule classe. On obtient une succession de classes emboîtées.
- Remarque importante :
  - L'input classique de cette procédure est une matrice de dissimilarités.
  - On ne fixe pas *a priori* le nombre de classes et on peut couper *a posteriori* l'arbre afin d'avoir une partition en k classes.
  - Peut traiter tout type de données (quanti., quali., mixtes, relationnelles) à condition d'avoir une matrice de dissimilarités.
  - Plusieurs méthodes existent pour calculer la dissimilarité entre une nouvelle classe et les autres classes mais la formule paramétrique de Lance-Williams permet d'unifier sept techniques particulières.
  - L'ensemble de ces techniques permet de tenir compte de nombreuses situations. Contrairement aux k-means, on peut détecter des classes de forme non sphérique (méthode single par exemple).
  - Complexité en $O(n^3)$ (donc plus coûteux que les k-means).
- Commandes `hclust, cutree`.

La classification ascendante hiérarchique (CAH)



L'algorithme des k-modes

- Principe : Extension de l'algorithme des k-moyennes aux données qualitatives. La procédure est la même. C'est le concpet de barycentre qui change : le vecteur représentant d'une classe est le vecteur dont chaque variable est donnée par le mode (modalité la plus fréquente) parmi les membres de la classe.
- Remarques importantes :
  - La mesure de proximité par défaut est basée sur le "simple matching distance" : pour deux individus, on compte le nombre de variables dont les modalités ne sont pas les mêmes.
  - On montre que le "vecteur des modes" est celui qui minimise les distances de "simple matching" avec les individus de la classe (de la même façon que le barycentre est le vecteur qui mimise les distances euclidiennes avec les individus de la classe).
  - Complexité en $O(n)$ (si q (nombre total de modalités) et k sont petits).
- Commande `kmodes [klaR]`.

CAH de Ward et réduction de dimension

- Principe : les méthodes de réduction de dimension permettent de mettre en lumière différents groupes d'individus sans pour autant en dessiner des contours exacts. On peut alors utiliser une méthode de classification automatique pour détecter des classes.
- Remarques importantes :
  - La représentation dans l'espace réduit est utilisée comme représentation euclidienne des données. Dans le cas des données qualitatives, l'ACM permet d'avoir une représetation continue des données.
  - Les méthodes factorielles reposent sur des critères intertiels. C'est aussi le cas des k-moyennes ou de la CAH de Ward. Ces méthodes sont donc particulièrement en adéquation avec la représentation factorielle.
  - Pour éviter de fixer k, c'est la CAH de Ward qui est associée classiquement aux méthodes de réduction de dimension. Mais, en pratique, les k-moyennes sont aussi utilisés soit pour faire face au pb de complexité de le CAH, soit pour améliorer la partition à k classes obtenue en coupant l'arbre de la CAH.
- Commande `HCPC [FactoMineR]`.

Evaluation et comparaison de partitions

- Principe : évaluer la qualité du résulat d'une méthode de classification automatique. On distingue la **validation interne** de la **validation externe**. Dans ce dernier cas, on dispose de la vérité terrain.
- Remarques importantes :
  - La validation interne est basée sur des mesures caractérisant l'homogénéité des classes obtenues, en analysant les distances entre les membres d'une même classe ou entre les membres de classes distinctes. On utilisera typiquement la **variance intra-classe et la variance inter-classe** mais d'autres indices existent comme la valeur moyenne de la silhouette.
  - La validation externe confronte la vraie partition à celle obtenue par une méthode. Dans ce cas, les critères de validation sont des mesures de similarité ou d'association entre deux partitions. L'**indice corrigé de Rand** est un critère typique mais d'autres mesures existent.
  - Les mesures de validation externe peuvent être aussi utilisées pour comparer les partitions obtenues par deux méthodes différentes.
- Commande `cluster.stats [fpc]`.

Ressources en-ligne

- La classification automatique est une sous-discipline majeure de la fouille de données.
- Concernant R, voici quelques pointeurs :
  - https://cran.r-project.org/web/views/Cluster.html
  - http://www.statmethods.net/advstats/cluster.html
  - http://www.rdatamining.com/docs/data-clustering-with-r
  - https://www.stat.berkeley.edu/~s133/Cluster2a.html
  - https://rstudio-pubs-static.s3.amazonaws.com/33876_1d7794d9a86647ca90c4f182df93f0e8.html
  - https://www.r-bloggers.com/search/clustering/

Règles d'associations

- Principe : on cherche des règles du type si conditions alors *résultats* qui soient vraies pour au moins Y% des cas et qui se rencontrent globalement pour au moins X% des individus de la base. Y% est appelé indice de confiance et X% indice de support.
- Remarques importantes :
  - L'**indice de support** est formellement :
$$P(\text{conditions} \wedge \textit{résultats})$$
  - On cherche des associations entre l'observation d'une conjonction de modalités d'un ensemble de variables et l'observation d'une modalité d'une autre variable.
  - L'**indice de confiance** est formellement :
$$\frac{P(\text{conditions} \wedge \textit{résultats})}{P(\text{conditions})} = P(\textit{résultats}|\text{conditions})$$
  - Cette tâche s'applique sur des données qualitatives. Il faut donc discrétiser les variables quantitatives si l'on souhaite les utiliser.

Quelques définitions

- Une règle est de type conditions $\rightarrow$ *résultats*. On écrit également antécédent $\rightarrow$ *conséquent*.
- Exemple des tickets de caisse : Si "couche" $\wedge$ "samedi" $\rightarrow$ *"bière"*.
- Les conditions sont une conjonction de type :
$$\text{conditions}=(V^1 = a) \wedge (V^2 = b) \wedge (V^3 = c)$$
Chaque élément est appelé **item**.
- Le *résultat* est un unique item (et non pas une conjonction d'items) qui ne fait pas partie des conditions.
- Les indices de support et de confiance permettent de sélectionner des règles pouvant être intéressantes. En pratique, il faut compléter ces critères par des mesures dites d'intérêt. Une mesure classique est le **lift** défini formellement comme suit :
$$\frac{P(\text{conditions} \wedge \textit{résultats})}{P(\text{conditions})P(\textit{résultats})}$$

L'algorithme Apriori

- Principe : algorithme basé sur le support et la confiance. Recherche dans un premier temps de sous-ensembles d'items ayant un support suppérieur à un seuil X. Puis, il décompose chacun des sous-ensemble d'items en paires (conditions, *Résultat*) de sorte à ce que l'indice de confiance soit supérieur à un seuil Y.
- Remarques importantes :
  - Première étape : on exclut tous les sous-ensembles d'items peu fréquents. Si un sous-ensemble de taille p est fréquent alors un sous-ensemble de taille $p-1$ de ce sous-ensemble est également fréquent. On n'a besoin que d'une seule passe sur les individus.
  - Deuxième étape : soit un sous-ensemble fréquent de taille p, il existe $2^{p-1}-1$ règles possibles conditions $\rightarrow$ *Résultat*. Apriori permet d'identifier rapidement les règles dépassant un seuil de confiance.
  - Malgré cela, il exite en pratique bcp de règles peu intéressantes et il faut donc d'une part imposer un seuil de confiance très fort (>75%) et plus fort que le seuil de support (règles rares) ; d'autre part continuer à filtrer les règles par d'autres mesures d'intérêt.

Rappel du Sommaire

4 Data Mining prédictif
- Régression
- Classement
- Arbres de decision et fôrets aléatoires

Analyse prédictive

- Nous disposions d'une table de données avec n individus et p variables que l'on notera $\mathbf{x}^1, \ldots, \mathbf{x}^p$. Nous disposons de plus d'une **variable cible** notée $\mathbf{y}$.
- L'objectif est de modéliser $\mathbf{y}$ en fonction de $\mathbf{x}^1, \ldots, \mathbf{x}^p$ dans le but ultime de faire des prédictions. Contrairement à l'analyse descriptive où il n'y a pas de variable d'intérêt, on parle ici d'**apprentissage supervisé** car c'est la variable cible $\mathbf{y}$ qui nous intéresse en particulier.
- La variable $\mathbf{y}$ peut être typiquement soit quantitative continue, soit qualitative nominale (discrète). Dans le 1er cas, on a un problème de **régression** tandis que le 2ème cas est un problème de **classement**.
- Pour la régression, le modèle linéaire est l'approche la plus fréquente : on suppose la relation $\mathbf{y} = a_0 + a_1\mathbf{x}^1 + \ldots + a_p\mathbf{x}^p + \boldsymbol{\epsilon}$. Il existe en revanche plusieurs façons d'inférer les paramètres $\{a_0, a_1, \ldots, a_p\}$.
- Pour le classement, il existe plusieurs modèles qui sont inspirés de la statistique mais également de l'intelligence artificielle.

Régression - Nature des variables explicatives

- La variable à expliquer $\mathbf{y}$ est **quantitative**.
- Les variables explicatives $\mathbf{x}^1, \ldots, \mathbf{x}^p$ peuvent être de plusieurs natures également :
  - $\mathbf{x}^1, \ldots, \mathbf{x}^p$ sont toutes quantitatives : régression linéaire multiple,
  - $\mathbf{x}^1, \ldots, \mathbf{x}^p$ sont toutes qualitatives : analyse de la variance (à plusieurs facteurs),
  - $\mathbf{x}^1, \ldots, \mathbf{x}^p$ forment un mélange de var. quanti. et quali. : analyse de la covariance.
- On se restreint aux problèmes de régression linéaire multiple et on (re)voit les moindres carrés ordinaires, la régression sur composantes principales et les moindres carrés partiels.
- D'autres techniques existent comme la régression pénalisée (ridge, lasso, elasticnet), les machines à vecteurs supports (svm)...

Régression par moindres carrés ordinaires MCO

- On se restreint au cas où $\mathbf{x}^1, \ldots, \mathbf{x}^p$ sont toutes quantitatives.
- Principe : On suppose la relation linéaire suivante $\mathbf{y} = a_0 + a_1\mathbf{x}^1 + \ldots + a_p\mathbf{x}^p + \boldsymbol{\epsilon}$ et pour inférer les paramètres on cherche à minimiser $\sum_{i=1}^{n} \epsilon_i^2$ (somme des carrés des résidus).
- Remarques importantes :
  - Si on fait de plus l'hyopothèse que ϵ_i sont i.i.d. selon $\mathcal{N}(0, \sigma^2)$ alors on parle de **modèle linéaire gaussien**. Dans ce cas, l'estimateur du maximum de vraisemblance (MV) est identique à l'estimateur des moindres carrés ordinaires. On peut alors compléter l'estimation ponctuelle par des intervalles de confiance et des tests de significativité (tests de Student, Fisher, ...).
  - Attention ! Pour que les méthodes de statistiques inférentielles soient valides, il faut vérifier que les hypothèses de gaussianité, d'indépendance et d'homoscédasticité soient vérifiées.
  - Le théorème de Gauss-Markov nous dit que l'estimateur du MV est celui de variance minimale parmi les estimateurs linéaires sans biais.
- Commande `lm`.

Régression sur composantes principales (PCR)

- Principe : lorsque les variables $\mathbf{x}^1, \ldots, \mathbf{x}^p$ ne sont pas linéairement indépendantes alors la méthode des MCO n'est pas identifiable. Une approche consiste à réduire l'espace de description et d'appliquer les MCO dans cet espace. La méthode classique consiste à faire une ACP et de faire la régression sur les premières composantes principales.
- Remarques importantes :
  - En effet, on sait que les axes principaux sont mutuellement orthogonaux et il n'y a donc plus de problèmes de colinéarité.
  - Comme en ACP, on n'utilise pas ici toutes les composantes principales. On profite donc ici du principe de sélection d'information propre aux méthodes de réduction de dimension. En théorie, la régression PCR est donc moins sensible aux données aberrantes.
  - Même si on régresse sur des composantes principales sachant qu'elles sont des combinaisons linéaires des variables initiales, on peut toujours se ramener à une expression du modèle en fonction de ces dernières.
- Commande `pcr [pls]`.

Régression par moindres carrés partiels (PLS)

- Principe : la méthode PLS est aussi une régression sur des composantes (càd des variables synthétiques). Mais contrairement à la régression PCR, la méthode PLS détermine des composantes qui privilégient les variables explicatives fortement corrélées à la variable à expliquer.
- Remarques importantes :
  - Les composantes PLS sont déterminées itérativement. A chaque étape k, on cherche $\mathbf{t}^k$, la combinaison linéaire de la partie résiduelle des variables explicatives la plus corrélée à $\mathbf{y}^k$, la partie résiduelle de la variable à expliquer.
  - On fait ensuite une régression linéaire simple par MCO de $\mathbf{y}^k$ sur $\mathbf{t}^k$.
  - La partie résiduelle est la part des données non encore expliquées. C'est la projection des variables sur l'espace engendré par les résidus $\boldsymbol{\epsilon}^k$.
  - On montre que les composantes PLS, $\mathbf{t}^1, \ldots, \mathbf{t}^m$ sont des combinaisons linéaires des variables initiales et qu'elles sont orthogonales entre elles (similairement à la méthode PCR).
- Commande `pls [pls]`.

Sélection de modèles

- Il existe plusieurs méthodes/modèles pour résoudre une tâche de DM.
- Etant donné une étude de cas, comment choisir un modèle ?
- Approche "axiomatique" :
  - Chaque modèle repose sur des hypothèses.
  - Une bonne maîtrise de ces hypothèses et des données permet de choisir un sous-ensemble approprié de méthodes.
- Approche empirique :
  - On teste plusieurs méthodes (sans faire trop attention à leurs fondements) sur les données de l'étude.
  - On choisit le modèle donnant les meilleurs résultats.
- On attend d'un modèle qu'il fasse de **bonnes prédictions sur des données non observées** !
- D'autres critères de sélection existent comme la possibilité d'interpréter un modèle par opposition aux méthodes "boîtes noires".
- Dans ce cours, nous faisons de la pratique et utiliserons donc l'approche empirique.

Protocol expérimental

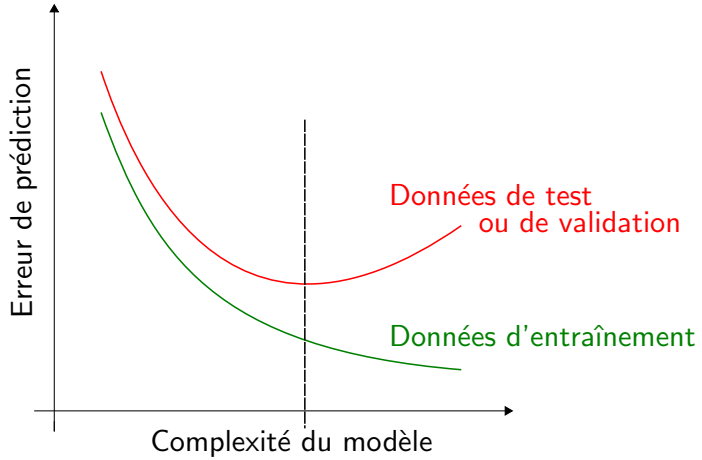
- On dispose d'un jeu de données annotées $\mathbb{O}$ avec n individus.
- On découpe $\mathbb{O}$ en deux sous-ensembles disjoints $\mathbb{O} = \mathbb{E} \cup \mathbb{T}$ où :
  - $\mathbb{E}$: ensemble d'**entraînement** ou d'apprentissage,
  - $\mathbb{T}$: ensemble de **test**.
- On infère les paramètres du modèle à partir des données $\mathbb{E}$.
- On teste le modèle estimé sur les données non observées $\mathbb{T}$.
- On distingue deux types d'erreur :
  - l'erreur du modèle estimé sur $\mathbb{E}$ est l'**erreur d'entraînement**,
  - l'erreur du modèle estimé sur $\mathbb{T}$ est l'**erreur en généralisation**.
- Attention ! En DM, c'est l'erreur en généralisation qu'il est important de minimiser !

Arbitrage biais-variance

- **Sous-apprentissage** : le modèle repose sur des hypothèses trop restreintes (modèle trop simpliste) et on est sûr d'avoir une erreur d'entraînement et en généralisation forte.
- **Sur-apprentissage** : le modèle repose sur des hypothèses trop larges (modèle trop complexe), on obtient une erreur d'entraînement très faible mais une erreur en généralisation très forte.
- Modèle trop simpliste : si les données d'entraînement changent, les estimations du modèle changent peu (faible variance) mais l'erreur d'entraînement est élevée (fort biais). L'erreur en généralisation est forte malgré une faible variance en raison du fort biais.
- Modèle trop complexe : si les données d'entraînement changent, les estimations du modèle changent beaucoup (forte variance) mais l'erreur d'entraînement est très faible (faible biais). Cependant, l'erreur en généralisation peut être forte en raison de la forte variance.
- Le choix d'un bon modèle réside en un bon équilibre entre le **biais** et la **variance** !

Arbitrage biais-variance

- Illustration de l'arbitrage biais-variance :

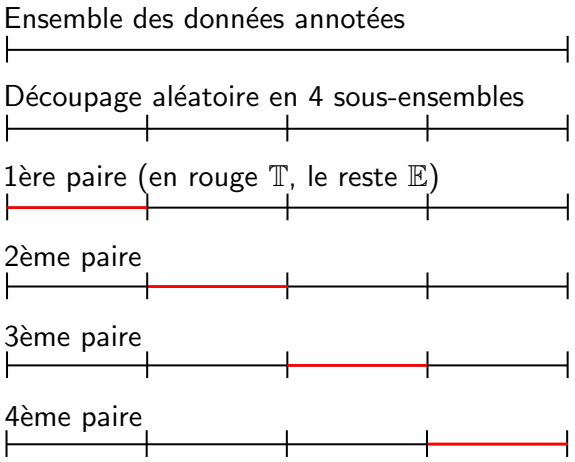


Procédure de validation croisée

- Rappel : si on change les données d'entraînement, on change les estimations du modèle et donc les performances de celui-ci.
- Principe : découper le jeu de données en k sous-ensembles de taille identique. On apprend sur l'union de $k-1$ sous-ensemble et on teste sur le sous-ensemble restant. On procède ainsi k fois (l'ensemble de test change à chaque itération). On obtient ainsi k estimations de la performance du modèle. On moyenne pour avoir une estimation plus robuste de la performance.
- Plus formellement :
  - Une k validation croisée signifie que l'on a k paires $(\mathbb{E}_j, \mathbb{T}_j)_{j=1,\ldots,k}$.
  - Pour chaque $j = 1, \ldots, k$, on apprend sur $\mathbb{E}_j$ et on teste sur $\mathbb{T}_j$.
  - Chaque paire j nous donne une estimation d'une mesure d'erreur ou de performance. On aboutit donc à k estimations distinctes. La moyenne est une estimation plus robuste que celle obtenue par chaque paire j.
  - Si $k = n$ on parle de "leave one out cross validation" (LOOCV).

Procédure de validation croisée (suite)

- Illustration d'une validation croisée à 4 échantillons.



Mesures de performances

- Dans le cas de la régression, on utilise classiquement les mesures d'erreur suivantes pour évaluer un modèle de prédiction f :
  - La Moyenne des carrés des résidus ("Mean Squared Error") :

$$mse(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2$$

  - La racine carrée de la moyenne des carrés des résidus ("Root Mean Squared Error") :

$$rmse(f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2}$$

  - La moyenne des résidus en valeurs absolues ("Mean Absolute Error") :

$$mae(f) = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(\mathbf{x}_i)|$$

Classement - Nature des variables explicatives

- La variable à expliquer **y** est cette fois-ci **qualitative**. On supposera dans la suite que les mesures dans **y** peuvent prendre q différentes valeurs : $\{C_1, \ldots, C_k, \ldots, C_q\}$.
- Les variables explicatives $\mathbf{x}^1, \ldots, \mathbf{x}^p$ peuvent être de plusieurs natures également :
  - $\mathbf{x}^1, \ldots, \mathbf{x}^p$ sont toutes quantitatives : analyse discriminante (linéaire et quadratique).
  - $\mathbf{x}^1, \ldots, \mathbf{x}^p$ forment un mélange de var. quanti. et quali. : régression logistique (binomiale ou multinomiale).
- Nous verrons essentiellement les méthodes mentionnées ci-dessus.
- Mais beaucoup d'autres méthodes existent ! comme les machines à vecteurs supports (svm), les réseaux de neuronnes, les réseaux bayésiens. . .

Analyse linéaire discriminante (LDA)

- Principe : les données sont issues d'un mélange de lois normales et chaque classe possède un vecteur moyen distinct. Formellement, étant donné une classe C_k on suppose que $\mathbf{x}|C_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ où $\boldsymbol{\mu}$ est le vecteur moyen et $\boldsymbol{\Sigma}$ la matrice de variance-covariance. On cherche alors à estimer $\boldsymbol{\Sigma}$ et pour chaque classe C_k, le vecteur $\boldsymbol{\mu}_k$. Puis, on donne un score et une prédiction à un point $\mathbf{x}$ quelconque à l'aide de $P(C_k|\mathbf{x})$ que l'on calcule grâce à la règle de Bayes.
- Remarques importantes :
  - Le cas linéaire considère que toute classe C_k possède la même matrice de variance-covariance $\boldsymbol{\Sigma}$: hypothèse d'homoscédasticité.
  - On aboutit alors à une fonction de score qui est linéaire en $\mathbf{x}$ appelée également **score de Fisher**.
  - La méthode peut être vue telle une technique de réduction de dimension où l'on cherche simultanément à maximiser l'inertie inter-classe et minimiser l'inertie intra-classe.
- Commande `lda [MASS]`.

Analyse quadratique discriminante (QDA)

- Principe : c'est le même cadre formel que précédemment càd pour chaque classe C_k on suppose que $\mathbf{x}|C_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Mais ici toute classe a un vecteur moyen et une matrice de variance-covariance distincte $\boldsymbol{\Sigma}_k$. L'abandon de l'hypothèse d'homoscédasticité conduit à une fonction de score qui est quadratique en $\mathbf{x}$.
- Remarques :
  - Pour LDA, la fonction de score est linéaire en $\mathbf{x}$ ce qui veut dire que dans l'espace de description, on peut tracer des frontières linéaires (hyperplans) séparant les classes C_k entre elles.
  - Pour QDA, la fonction de score est quadratique en $\mathbf{x}$ ce qui veut dire que les frontières séparant les classes dans l'espace de description sont des courbes.
  - QDA est plus flexible que LDA mais demande plus de calculs car il y a q matrices $\boldsymbol{\Sigma}_k$ à estimer.
- Commande `qda [MASS]`.

Régression logistique

- Principe : pour chaque classe C_k on modélise directement la probabilité $P(C_k|\mathbf{x})$ par une forme paramétrique. On suppose ensuite une loi de distribution pour l'observation du nb d'occurrence des classes. Les paramètres sont obtenus par maximum de vraissemblance.
- Précisément, dans la régression logistique on a :
  - La 1ère hypothèse concerne la forme paramétrique de $P(C_k|\mathbf{x})$:

$$P(C_k|\mathbf{x}) = \frac{\exp(a_{k0} + \mathbf{a}_k^\top \mathbf{x})}{\sum_{l=1}^{q} \exp(a_{l0} + \mathbf{a}_l^\top \mathbf{x})}$$

  - Ceci est équivalent à supposer $q-1$ fonctions logits des odds-ratio :

$$\log \frac{P(C_k|\mathbf{x})}{P(C_q|\mathbf{x})} = a_{k0} + \mathbf{a}_k^\top \mathbf{x}$$

  - La 2ème hypothèse concerne la distribution de probabilité du nb d'occurrences des classes. La régression logistique suppose une loi multinomiale qui généralise au cas $q > 2$ la loi binomiale.
- Commande `multinom [nnet]` [25]

25. La regression logistique multinomiale est un cas simple de réseau de neuronnes : c'est q perceptrons en parallèle.

Rappels sur la sélection de modèles

- Ce qui a été évoqué précédemment aux slides concernant les problèmes de régression reste valable pour les problèmes de classement :
  - l'arbitrage biais-variance et les problèmes de sous et sur-apprentissage,
  - l'importance de l'erreur en généralisation,
  - le protocol expérimental, l'estimation plus robuste par validation croisée de l'erreur en généralisation,
  - les approches "axiomatique" et empirique pour la sélection de modèle et l'accent mis dans ce cours sur l'approche empirique.
- Ce qui est spécifique aux problèmes de classement :
  - les mesures d'erreur/de performance pour comparer les modèles,
  - dans le cas de deux classes ($q = 2$), la décision de prédiction est prise en comparant un score et un seuil ("si score de $\mathbf{x}$ est supérieur à un seuil θ alors je mets dans la classe C_1"). Dans ce cas, on peut utiliser un outil supplémentaire pour l'évaluation des modèles appelé courbe ROC.

Mesures de performance

- Principe : la variable cible $\mathbf{y}$ étant discrète, les mesures d'erreur reposent principalement sur la matrice confusion qui est une table de contingence croisant la vérité terrain et les prédictions du modèle.
- Matrice de confusion dans le cas $q = 2$:

		$\hat{f}(\mathbf{x})$		
		C_1	C_2	Total
y	C_1	a	b	$a+b$
	C_2	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d=n$

  - a =Nb d'objets C_1 correctement catégorisés
  - b =Nb d'objets C_1 catégorisés en C_2
  - c =Nb d'objets C_2 catégorisés en C_1
  - d =Nb d'objets C_2 correctement catégorisés

Mesures de performance (suite)

- A partir de la matrice de confusion on définit :
  - Taux d'erreur ("Error rate" ou "Misclassification Rate") :

$$err(\hat{f}) = \frac{b+c}{n}$$

  - Taux de réussite ou de reconnaissance ("Accuracy Rate") :

$$acc(\hat{f}) = \frac{a+d}{n} = 1 - err(\hat{f})$$

  - Taux de vrais positifs[26] ("True positive rate" ou "Sensitivity") :

$$tp(\hat{f}) = \frac{a}{a+b}$$

  - Taux de faux positifs ("False positive rate" ou "False alarm rate") :

$$fp(\hat{f}) = \frac{c}{c+d}$$

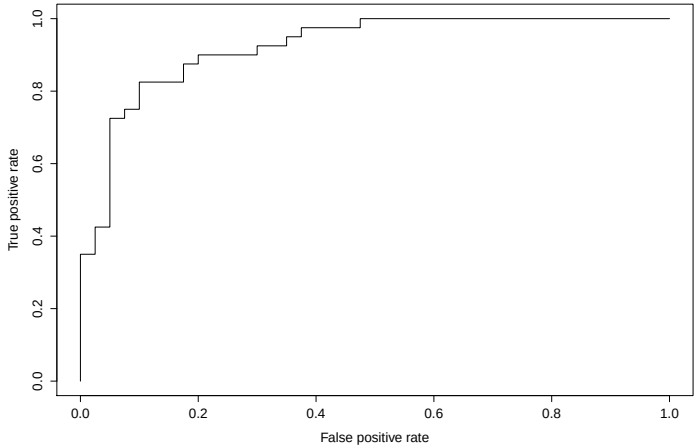
26. On suppose ici que C_1 est la classe positive.

Courbe ROC

- Principe : dans les problèmes de classement binaire ($q = 2$), la plupart des méthodes prennent la décision d'affecter dans C_1 au travers d'une fonction de score g. On a $f(\mathbf{x}) = C_1$ ssi $g(\mathbf{x}) > \theta$ où θ est un seuil. Ce seuil est donné par défaut mais si on le change alors les performances du modèle varient également. La courbe ROC ("Receiver Operating Characteristics") permet d'étudier la sensibilité d'un modèle vis à vis de ce seuil.
- Quelques précisions :
  - La courbe est tracée dans un plan ou chaque axe correspond à deux mesures de performance. Typiquement, *fpr* en abscisse et *tpr* en ordonnée. Le point $(0, 1)$ correspond au modèle/seuil optimal.
  - La 1ère bissectrice du plan correspond à un modèle de prédiction aléatoire. Il faut donc avoir une courbe au-dessus de la 1ère bissectrice.
  - Si on a deux modèles, on peut tracer deux courbes ROC et celle qui est au-dessus de l'autres correspond au meilleur modèle.
  - L'aire en dessous de la courbe ROC (auc) est une valeur synthétisant la performance d'un modèle. 1 correspond au modèle optimal.

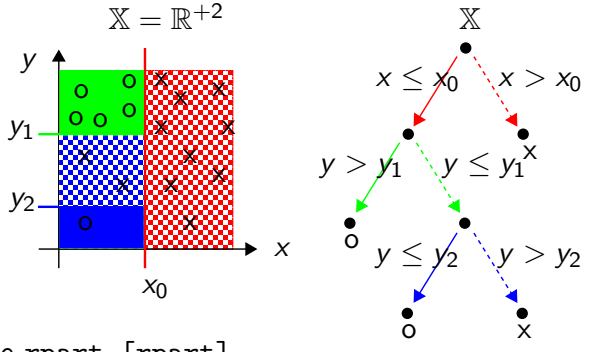
Courbe ROC (suite)

- Illustration de la courbe ROC :



Arbres de décision

- Principe : la méthode consiste à découper l'espace de description (càd celui engendré par les variables explicatives) en hyper-rectangles. Chaque hyper-rectangle est défini comme la conjonction de plusieurs règles simples chacune basée sur une variable explicative. A chaque hyper-rectangle on associe une valeur de la variable cible.



- Commande `rpart [rpart]`.

Arbres de décision (suite)

- Remarques importantes :
  - A chaque étape une variable est choisie afin de séparer en deux un hyper-rectangle existant. Cette séparation est simplement modélisée par une règle. L'ensemble de ces règles est représenté par un arbre binaire.
  - La méthode permet de traiter les problèmes de régression et de classement (binaire ou multi-classe). De plus, elle permet de tenir compte des données mixtes.
  - En régression, les hyper-rectangles sont définis de sorte à minimiser la somme des carrés des résidus. La valeur prédite associée à un hyper-rectangle est la moyenne des y des individus de l'hyper-rectangle.
  - En classement, les hyper-rectangles sont définis afin de minimiser l'impureté qui est en général mesurer par l'entropie. La valeur prédite associée à un hyper-rectangle est la classe majoritaire des y des individus de l'hyper-rectangle.
  - La méthode est sujette au sur-apprentissage et pour y remédier, on emploie des méthodes d'élagage (on enlève des branches de l'arbre).
  - Un avantage certain des arbres de décision est qu'il fournit un modèle interprétable sous forme de règles "si ... alors ...".

Bootstrap et Bagging

- Principe :
  - Bootstrap : c'est une méthode de ré-échantillonnage avec remise permettant de disposer de plusieurs échantillons afin d'avoir plusieurs estimations de modèles et d'erreurs en généralisation.
  - Bagging ("bootstrap + averaging") : c'est un paradigme de méthode d'ensemble en apprentissage supervisé qui repose sur le bootstrap. L'idée est d'estimer une même méthode sur plusieurs échantillons bootstrap et de faire une prédiction basée sur un consensus de ces différents modèles estimés qui représentent autant d'opinions distinctes.
- Remarques importantes :
  - Le consensus pour les pbs de régression est en général une moyenne tandis que pour les pbs de classement c'est le vote majoritaire.
  - Le bagging est souvent appliqué avec les arbres de décision car il permet de réduire la variance de ces derniers (pb de sur-apprentissage des ad).

Forêts aléatoires

- Principe : il s'agit du bagging appliqué avec les arbres décisionnels et auquel on ajoute un échantillonnage sur les variables explicatives. En effet, lorsqu'un arbre est appris à partir d'un échantillon bootstrap, à chaque itération, on choisit la variable de séparation dans un sous-ensemble des variables explicatives qui est pris aléatoirement.
- Remarques importantes :
  - Choisir aléatoirement un sous-espace de représentation à chaque étape permet de rendre davantage indépendants les échantillons bootstrap (qui ne le sont pas à la base en raison de le remise). En théorie, cela permet de réduire la variance globale du modèle. En pratique, les forêts aléatoires donnent souvent d'excellents résultats.
  - La méthode ne nécessite pas de validation croisée ! Pour chaque échantillon bootstrap, on peut tester le modèle sur les individus n'appartenant pas à l'échantillonet mesurer l'erreur. On moyenne ensuite toutes ces mesures, on parle alors de "out of bag error".
  - Les forêts aléatoires sont des "boîtes noires" mais, il est possible d'évaluer l'importance de chaque variable dans le modèle estimé.

Forêts aléatoires (suite)

- Commande `rf [randomForest]`.