UP2 : Apprentissage Statistique et Analyse de données

Arbre de décision et Forêt aléatoire

ANIS S. HOAYEK

Octobre 2025

Plan du cours

1. Apprentissage statistique pour l'aide à la décision.
2. Méthodes de Classification : Principes généraux
3. Arbres de décision.
4. Validité : Sensibilité, Spécificité, ROC, AUC, LIFT, etc.
5. Forêts aléatoires.
6. Lien avec Bagging et Boosting.
7. Forêts aléatoires non supervisées (Isolation Forest).
8. Applications classiques TD (à la main) + TP (sous R et/ou Python).
9. Évaluation : Examen écrit (13 Novembre 2025) + TP noté (compte rendu par groupes selon la liste prédéfinie).

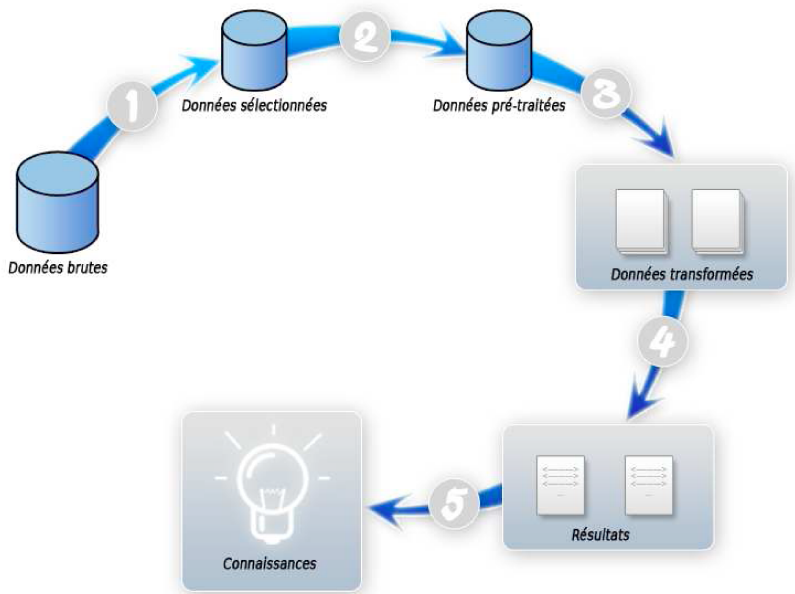
Apprentissage statistique pour l'aide à la décision

Introduction

1. La statistique est incontournable dans les sciences expérimentales : données dans le but spécifique de confirmer ou d'infirmer des hypothèses scientifiques.
2. Données existantes avant même qu'on se pose des questions spécifiques.
3. Apprendre de ces données les réponses à des questions $\Longrightarrow$ La statistique permet de faire cet apprentissage $\Longrightarrow$ prendre des décisions.
4. Méthodes de la statistique dans un contexte expérimental :
    1. Planification du recueil des données (Échantillonnage, Plan d'expérience, Data management)
    2. Description et présentation des données (Statistique descriptive)
    3. Induction de règles générales à partir de ce qui a été observé (Modélisation et Inférence statistique)

Définition.

Le data-mining, est un processus d'extraction des connaissances qui comprend plusieurs étapes :
Pré-traitement des données $\Longrightarrow$ Fouille des données pré-traitées $\Longrightarrow$ Apprentissage à partir de ce qui a été observé



Domaines d'application

Entreprises stockent dans des entrepôts de données (data warehouse) des des picas octets de données relatives à leurs activités à des fins de :

1. Gestion des stocks, des services, des ressources humaines, des clients, etc.
2. Contrôle de qualité : suivi en ligne des paramètres de production, traçabilité, etc.
3. Secteur bancaire : Identification du risque de crédit d'un client en fonction de sa probabilité de défaut de paiement.
4. Médecine : Identification des patients à risque et les tendances de la maladie.
5. Marketing : Identification du taux de désabonnement des clients.

Remarque.

Les données n'ont pas été recueillies à ces fins !

Étape 1 : Nettoyage et mise en forme des données

1. S'assurer de la fiabilité des données,
2. Éliminer les sujets atypiques ou non pertinents pour l'analyse
3. Obtenir un "feeling" du jeu de données :
    1. Analyses descriptives unidimensionnelles adaptées à la nature des variables.
    2. Analyses statistiques bivariées adaptées à la nature des couples de variables pour faire apparaître les liens entre elles.
    3. Analyses statistiques multivariées.

BUT : Repérer les variables ayant des distributions statistiques "bizarres". Le cas échéant, décider s'il convient d'éliminer ou de transformer les données.

Étape 2 : Analyse exploratoire des données

1. Réduire la dimension du vecteur de variables
2. Éliminer les variables redondantes

Via des techniques statistiques qui ne sont pas l'objet de ce cours : ACP, AFC, AFD, etc.

Étape 3 : Classification/Régression

1. Une variable d'intérêt (ou cible) Y qualitative/discrète ou quantitative, définissant des groupes de la population
2. Une ou plusieurs variables explicatives ou prédictives $X_1, \ldots, X_p$.
3. Sur un individu, on va observer uniquement la valeur $(x_1, \ldots, x_p)$ de $(X_1, \ldots, X_p)$.
4. On voudra lui assigner sa valeur y de Y (son groupe ou sa valeur) en minimisant les risques d'erreur.
5. Pour établir le classifieur qui fera ce travail, on dispose de données du type $(x_1, \ldots, x_p, y) \Longrightarrow$ On veut à partir de ces données établir la meilleure règle de classification/régression possible pour classer par la suite les éventuels prospects.

Étape 3 bis : Clustering

1. On dispose des valeurs observées $(x_1, \ldots, x_p)$ sur n individus des variables $(X_1, \ldots, X_p)$.
2. Regrouper, en K groupes les plus dissemblables entre eux possibles, les individus de façon à ce qu'à l'intérieur de chaque groupe, les données soient le plus homogène possible.
    1. K connu : Clustering supervisée
    2. K inconnu : Clustering non supervisée
3. Analyser les particularités de chacun des K groupes

Remarque.

Le clustering revient à "créer" une variable qualitative/discrète indiquant les groupes auxquels peuvent appartenir les individus.

Étape 3 bis : Modélisation

1. La modélisation a pour objectif d'étudier les liens plus fins entre :
   1. Une (ou plusieurs) variables d'intérêt (ou cible) Y
   2. Une ou plusieurs variables explicatives (ou prédictives) $X_1, \ldots, X_p$
2. Elle se fait notamment via les méthodes de :
   1. Régression linéaire simple, multiple, non-linéaire : Y et $X_1, \ldots, X_p$ quantitatives/continues
   2. ANOVA, ANCOVA : Y quantitative/continue, au moins une des $X_1, \ldots, X_p$ qualitative/discrète
   3. Régression logistique : Y qualitative/discrète et $X_1, \ldots, X_p$ quantitatives/continues
   4. Régression log-linéaire : Y et $X_1, \ldots, X_p$ qualitatives/discrètes
   5. Autres méthodes : séries chronologiques etc...

Remarque.

Dans ce cours, on se concentre sur les méthodes d'apprentissage que sont la classification, et la régression.

Stratégies d'une application classique du data-mining

1. Extraction des données de l'entrepôt, via éventuellement un sondage,
2. Nettoyage : analyse graphique, validation des codages etc..
3. Exploration : réduction de la dimension, élimination de variables redondantes,
4. Analyse (classification, régression, clustering, modélisation)
5. Exploitation du modèle et diffusion des résultats.

Pièges du data-mining

1. Attaque du problème sans but précis.
2. Data-snooping : “If you torture the data long enough, it will confess to anything” (Ronald Coase)
3. Incompréhension des algorithmes utilisés par les logiciels.
4. Confiance aveugle dans les sorties des logiciels informatiques de plus en plus ”tout intégré” et de plus en plus ”boîte noire”.
5. Absence d’interrogation et d’esprit critique dans le choix des ”valeurs par défaut” fait par les logiciels.

Méthodes de Classification : Principes généraux

Principes généraux

Problème : Déterminer à quel groupe $\mathcal{G}_1, ..., \mathcal{G}_K$ appartient un individu :

- L'appartenance à un groupe est déterminée par la valeur y d'une v.a. $Y \in \{1, 2, ..., K\}$.
- Si $y = k$, alors l'individu $\in \mathcal{G}_k$.

Exemple : Un client demande un prêt à une banque. La banque veut savoir si le client est en mesure de rembourser son prêt ($y = 1$) ou non ($y = 2$). Évidemment, l'observation de y est impossible, sauf à attendre l'issue du prêt.

Idée : Mesurer sur l'individu un certain nombre de variables $X = (X_1, \ldots, X_p)$, (e.g. : ancienneté, compte courant, compte sur livret, assurance-vie, etc...) dites discriminantes, classifiantes ou de "procuration", et déterminer à partir de la valeur observée $x = (x_1, \ldots, x_p)$ de X si l'individu fait (fera) parti de $\mathcal{G}_1$ ($y = 1$), $\mathcal{G}_2$ ($y = 2$), ..., $\mathcal{G}_K$ ($y = K$).

Postulat de base : X est un vecteur aléatoire. Il a donc une loi de probabilité (notée $\mathcal{L}_k$) qui diffère selon la valeur de Y (c-à-d : si X provient de $\mathcal{G}_k$, $X \sim \mathcal{L}_k$)

Classifieur

Définition.

Un classifieur est une fonction $d(\cdot) : \mathbb{R}^p \longrightarrow \{1, 2, ..., K\}$ qui, à chaque $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, associe un et un seul $k \in \{1, 2, ..., K\}$.
La classification se fait comme suit : Si un individu a x pour valeur de X, on l'assigne à $\mathcal{G}_K$ si $d(x) = k$.

Comment déterminer $d(\cdot)$? On aura besoin :

1. D'une méthode pour construire $d(\cdot)$
2. D'information sur le contexte du problème ($\mathcal{L}_k$, probabilités a priori d'appartenance aux groupes)
   et/ou
3. D'un échantillon $\mathcal{E}$ composé de valeurs observées $(x_i, y_i), i = 1, \ldots, n$.

Difficulté : $\exists$ une infinité de façons de choisir $d(\cdot)$. Quels critères utiliser pour ce choix ?

Information disponible pour construire un classifieur

La construction d'un classifieur dépend de l'information dont on dispose.

1. Si on connaît parfaitement la loi $\mathcal{L}_k$ de X (de densité $f_k(\cdot)$) quand $y = k$, on peut utiliser cette loi pour construire $d(\cdot)$ (cas d'école car il n'utilise pas $\mathcal{E}$)
2. Si on connaît imparfaitement cette densité (i.e. à des paramètres près), on peut utiliser $\mathcal{E}$ pour estimer ces paramètres, puis fonctionner comme en 1) (classifieur du Maximum de Vraisemblance avec paramètres estimés)
3. Si on n'a aucune information sur les différentes lois $\mathcal{L}_k$, on doit utiliser $\mathcal{E}$ pour construire un classifieur (e.g. arbre de décision). C'est le cas le plus réaliste.

Information disponible pour construire un classifieur

À ces informations peut s'ajouter :

- Une information a priori sur les "chances" qu'un individu $\in \mathcal{G}_k$. Cette information prend la forme de probabilités a priori π_k, $k = 1, \ldots, K$ avec $0 < \pi_k < 1$ et $\pi_1 + \ldots + \pi_K = 1$
- Une information sur les coûts d'une mauvaise classification $C(k', k)$ (par. ex coût de consentir un prêt à un client mauvais payeur)

Remarque.

Nous allons nous concentrer sur le cas 3) qui est celui rencontré en pratique.

Évaluation d'un Classifieur : Probabilités d'assignation correcte et incorrecte

- L'individu sera assigné à $\mathcal{G}_k$ selon la valeur observée x de X.
- Cette assignation est correcte si l'individu provient bien de $\mathcal{G}_k$. Sinon, l'assignation est incorrecte.
- Comme cette assignation se fait à partir de la valeur observée x de X, qui est aléatoire, il existe donc des probabilités de bonne et mauvaise assignation.

Définition.

La probabilité d'assigner un individu à $\mathcal{G}_{k'}$ alors qu'il est issu du groupe $\mathcal{G}_k$ est définie par

$$\rho_{k'k} = \mathbb{P}\left[X \longmapsto \mathcal{G}_{k'} \mid X \in \mathcal{G}_k \text{ ou } X \in \mathcal{L}_k\right] = \int_{\mathcal{G}_{k'}} f_k(x)\,\mathrm{d}x,$$

où $f_k(\cdot)$ est la densité de X (donc de $\mathcal{L}_k$) quand l'individu est issu de $\mathcal{G}_k$ (note : premier indice = là où il est envoyé ; deuxième indice, d'où il provient)

Évaluation d'un Classifieur : Probabilités d'assignation correcte et incorrecte

- Si les densités $f_k(\cdot)$ des lois $\mathcal{L}_k$ sont connues, on peut en principe calculer les $\rho_{k'k}$ (mais ce n'est pas nécessairement facile).
- ρ_{kk} $(k = 1, \ldots, K)$ sont les probabilités d'assignation correcte d'un individu provenant de $\mathcal{G}_k$.
- $1 - \rho_{kk}$ $(k = 1, \ldots, K)$ est la probabilité d'assigner de façon incorrecte un individu provenant de $\mathcal{G}_k$.
- Ces quantités peuvent être regroupées dans une matrice dite matrice de confusion dont la somme de chaque colonne $= 1$.

Mat. Confusion		vérité			
		$\mathcal{G}_1$	$\mathcal{G}_2$	$\ldots$	$\mathcal{G}_K$
prédiction	$\mathcal{G}_1$	ρ_{11}	ρ_{12}	$\ldots$	ρ_{1K}
	$\mathcal{G}_2$	ρ_{21}	ρ_{22}	$\ldots$	ρ_{2K}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$\mathcal{G}_K$	ρ_{K1}	ρ_{K2}	$\ldots$	ρ_{KK}
	total	1	1	...	1

Comparaison de deux classifieurs : Admissibilité et Probabilité globale d'assignation incorrecte

Définition.

Soit d un classifieur avec matrice de confusion $\{\rho_{k'k}, k', k = 1, ..., K\}$ et $\widetilde{d}$ un second classifieur avec matrice de confusion $\{\widetilde{\rho}_{k'k}, k', k = 1, ..., K\}$. On dit que d est aussi bon que $\widetilde{d}$ si $\rho_{kk} \geq \widetilde{\rho}_{kk} \, \forall k$. On dit que d est meilleur que $\widetilde{d}$ si $\rho_{kk} > \widetilde{\rho}_{kk}$ pour au moins un k.
Si d est un classifieur pour lequel il n'existe pas de meilleur classifieur, alors d est dit admissible.

Comparaison de deux classifieurs : Admissibilité et Probabilité globale d'assignation incorrecte

- Dans la définition précédente, les classifieurs sont jugés en fonction des éléments diagonaux de la matrice de confusion. Ce n'est pas la seule possibilité.
- Supposons que l'on dispose de l'information supplémentaire suivante :
  - On sait qu'a priori l'individu a une probabilité π_k ($\in]0,1[$) d'appartenir à $\mathcal{G}_k$.
  - La probabilité globale (ou a posteriori) d'assignation correcte d'un classifieur d avec matrice de confusion $\{\rho_{k'k}\}$ est définie par :

$$\begin{aligned}\mathbb{P}\left[\text{Individu} \longmapsto \text{à son groupe}\right] &= \sum_{k=1}^{K} \mathbb{P}\left[\text{Individu} \longmapsto \mathcal{G}_k \mid \text{Individu} \in \mathcal{G}_k\right] \mathbb{P}\left[\text{Individu} \in \mathcal{G}_k\right], \\ &= \sum_{k=1}^{K} \rho_{kk}\pi_k .\end{aligned}$$

Comparaison de deux classifieurs : Admissibilité et Probabilité globale d'assignation incorrecte

La probabilité globale d'assignation incorrecte de d est ainsi définie par :

$$\rho(d) = 1 - \sum_{k=1}^{K} \rho_{kk}\pi_k = \sum_{k=1}^{K}\sum_{k' \neq k} \rho_{k'k}\pi_k$$

Définition.

Supposons données les probabilités a priori π_k ($\in]0,1[$) d'appartenir à $\mathcal{G}_k$. Soit d un classifieur avec matrice de confusion $\{\rho_{k'k}, k', k = 1, ..., K\}$ et probabilité globale d'assignation incorrecte $\rho(d)$. Soit $\widetilde{d}$ un second classifieur avec matrice de confusion $\{\widetilde{\rho}_{k'k}, k', k = 1, ..., K\}$ et probabilité globale d'assignation incorrecte $\rho\left(\widetilde{d}\right)$. On dit que d est aussi bon que $\widetilde{d}$ si $\rho(d) = \rho\left(\widetilde{d}\right)$. On dit que d est meilleur que $\widetilde{d}$ si $\rho(d) < \rho\left(\widetilde{d}\right)$.

Note : En pratique, quand $\rho(d) < 0.2$, on considère que le classifieur est "bon".

Estimation des probabilités de bonne et mauvaise assignation

De façon générale, avec K groupes on peut estimer les $\rho_{k'k}$ par

$$p_{k'k} = \frac{n_{k'k}}{n_k}$$

où $n_{k'k}$ = nombre d'observations issues de $\mathcal{G}_k$ et assignées à $\mathcal{G}_{k'}$ et n_k = nombre d'observations de $\mathcal{E}$ provenant de $\mathcal{G}_k$. L'estimateur ainsi obtenu est appelé l'estimateur par resubstitution. Le tableau des $n_{k'k}$ est appelé la matrice d'incidence par resubstitution.

Mat. Incidence		vérité $\mathcal{G}_1$	$\mathcal{G}_2$	...	$\mathcal{G}_K$	total
prédiction	$\mathcal{G}_1$	n_{11}	n_{12}	...	n_{1K}	n_{1+}
	$\mathcal{G}_2$	n_{21}	n_{22}	...	n_{2K}	n_{2+}
	⋮	⋮	⋮	⋱	⋮	⋮
	$\mathcal{G}_K$	n_{K1}	n_{K2}	...	n_{KK}	n_{K+}
	total	n_1	n_2	...	n_K	n

Estimation des probabilités de bonne et mauvaise assignation

Le tableau des $p_{k'k}$ est la matrice de confusion estimée par resubstitution :

Mat. Conf. Est.		vérité			
		$\mathcal{G}_1$	$\mathcal{G}_2$	...	$\mathcal{G}_K$
prédiction	$\mathcal{G}_1$	p_{11}	p_{12}	...	p_{1K}
	$\mathcal{G}_2$	p_{21}	p_{22}	...	p_{2K}
	⋮	⋮	⋮	⋱	⋮
	$\mathcal{G}_K$	p_{K1}	p_{K2}	...	p_{KK}
	total	1	1	...	1

Ces estimateurs sont en général trop optimistes car **les mêmes données servent à la fois à construire le classifieur et à estimer ces probabilités.**

Estimation des probabilités de bonne et mauvaise assignation

Il existe plusieurs façons de contourner ce problème.
Nous en évoquons une ici rapidement. Elle consiste à partitionner l'échantillon $\mathcal{E}$ en deux sous-échantillons : $\mathcal{E}_{\text{app}}$ et $\mathcal{E}_{\text{est}}$
$\mathcal{E}_{\text{app}}$ = échantillon d'apprentissage (avec n_{app} données) sur lequel on construit le classifieur d_{app}.
$\mathcal{E}_{\text{est}}$ = échantillon d'estimation (avec n_{est} données) qu'on "passe" dans le classifieur pour obtenir les estimateurs :

- $p_{k'k}^{\text{est}}$, l'estimateur de $\rho_{k'k}$ pour le classifieur d
- $R^{\text{est}}(d)$, l'estimateur de $\rho(d) = 1 - \sum_{k=1}^{K} \rho_{kk}\pi_k$ (si les π_k sont disponibles).

Estimation des probabilités de bonne et mauvaise assignation

- Ces dernières quantités estiment en fait les $\rho_{k'k}$ de d_{app} et $\rho(d_{\text{app}})$ et non pas $\rho_{k'k}$ de d et $\rho(d)$
- Si n est grand, n_{app} et n_{est} le seront aussi. Ainsi $d \simeq d_{\text{app}} \Longrightarrow p^{\text{est}}_{k'k} \approx \rho_{k'k}$ de d et $R^{\text{est}}(d) \approx \rho(d)$.
- Ainsi, cette méthode est bien adaptée au cas où n est grand. On appelle cette méthode la validation externe ou encore "**out-of sample validation**". Elle dépend cependant du choix (aléatoire) de $\mathcal{E}_{\text{est}}$. Si n est grand, les $p^{\text{est}}_{k'k}$ est et $R^{\text{est}}(d)$ ne devraient pas trop varier d'un $\mathcal{E}_{\text{est}}$ à l'autre.

Estimation des probabilités de bonne et mauvaise assignation

- Si n est petit, une variante est la validation croisée ou interne (leave one out cross validation)
- L'idée consiste à prendre $\mathcal{E}_{\text{est}} = \{x_1\}$ et $\mathcal{E}_{\text{app}} = \mathcal{E} \setminus \{x_1\}$. On calcule le classifieur sur $\mathcal{E}_{\text{app}}$; celui-ci ne va pas différer de beaucoup de celui obtenu de $\mathcal{E}$.
- On regarde ensuite si l'individu de $\mathcal{E}_{\text{est}}$ est bien classé.
- On répète successivement cette procédure avec chacune des données. Au final, on estime les $\rho_{k'k}$ par le nombre de fois qu'une observation de $\mathcal{G}_k$ est classée en $\mathcal{G}_{k'}$.
- Des variantes utilisent des $\mathcal{E}_{\text{est}}$ de taille > 1 (**V-fold cross validation**).
- Le prix est un temps de calcul plus long.

Arbres de Décision

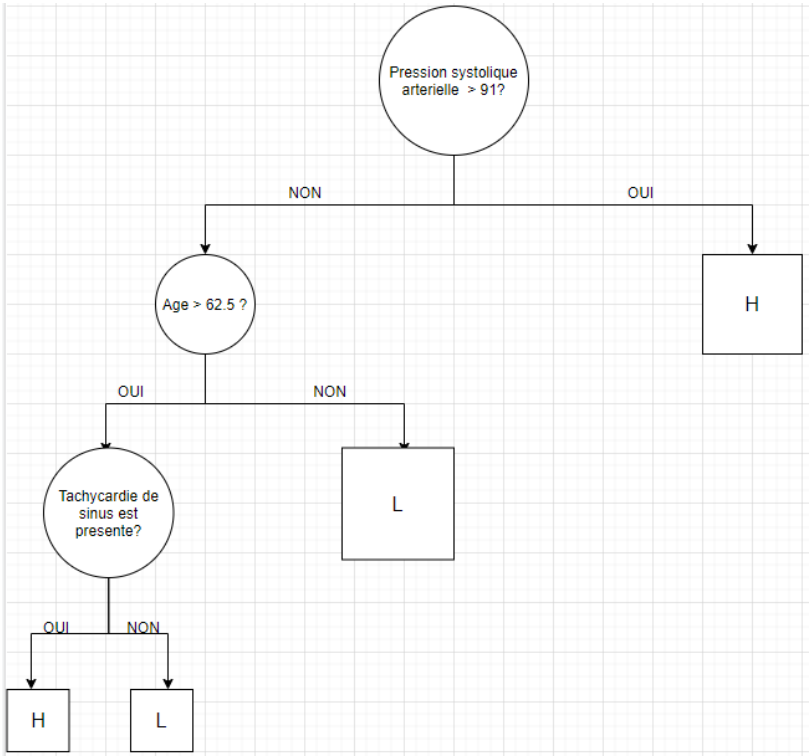
Fait.

Référence "classique" : Breiman, L. Friedman, J. H. Olshen, R.A., Stone, C. (1984) : Classification and regression trees. Wadsworth International Group. Belmont, California.

1. Algorithme de classification/régression supervisé.
2. Méthode statistique non-paramétrique.
3. Permet de classer un ensemble d'individus décrits par des variables qualitatives et quantitatives.
4. Produit des classes les plus homogènes possibles.
5. Classifications compréhensibles pour l'utilisateur (dans les méthodes classiques (hiérarchique, k-means,...) l'information est perdue dans les classes).

Introduction

1. Dans un hôpital, pour chaque nouveau patient avec une crise cardiaque, on mesure 15 variables pendant les premières 18 heures. Parmi les variables : la pression artérielle, l'age et 13 autres caractéristiques résumant les différents symptômes.
2. L'objectif de l'étude est d'identifier les patients à haut risque (ceux qui ne survivrons pas au moins 30 jours).



Introduction

- Représentation :
  - La racine : χ.
  - Un nœud : sous ensemble de χ (représenté par un cercle).
  - Nœuds terminaux : sous-ensembles qui ne sont plus divisés (représentés par des boites).
  - Chaque nœud terminal est marqué par une classe cible qui est une des valeurs y d'un attribut cible Y.
- Construction d'un arbre de décision :
  - Un arbre de classification se construit par segmentations récursives de l'échantillon $\mathcal{E} = \{(x_i, y_i), i = 1, ..., n\}$.
  - On sélectionne la variable qui sépare "le mieux" les données.
  - Le processus se répète pour chaque sous-groupe.
  - On s'arrête quand les sous-groupes atteignent la taille minimale, ou quand il n'y plus d'amélioration.

Ainsi, la construction d'un arbre de décision nécessite :

- Sélectionner les coupes.
- Décider de déclarer un nœud terminal (convertir le nœud en feuille) ou continuer de le scinder à nouveau. (L'idée principale est de choisir chaque split de façon à ce que les deux nœuds descendants soient chacun plus **"purs"** que le nœud parent).
- Affecter une classe à chaque nœud terminal.

Notations

- n : taille de l'échantillon (nombre des observations).
- K : nombre de classes de la variable cible Y.
- $N(t)$: nombre d'observations dans le nœud t.
- $N_k(t)$: nombre d'observations de la classe $k \in \{1, 2, \ldots, K\}$ dans le nœud t.
- $p(k|t)$: proportion d'observations dans le nœud t appartenant à la classe $k \in \{1, 2, \ldots, K\}$

$$p(k|t) = \frac{N_k(t)}{N(t)}.$$

- $p(t)$: vecteur de proportions correspondant au nœud t

$$p(t) = [p(1|t), p(2|t), \ldots, p(K|t)].$$

- $Y(t)$: classe attribuée au nœud t

$$Y(t) = \arg \max_{k=1,\ldots,K} p(k|t).$$

Définition.

Une mesure d'impureté d'un nœud t dans un arbre de décision ayant une variable cible Y de K classes est une fonction ayant la forme :

$$Imp(t) = \phi(p(t)),$$

où ϕ est une fonction non-négative de $p(t)$ qui satisfait les conditions suivantes :

1. ϕ atteint son maximum unique en $p(t) = \left[\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right]$.
2. ϕ atteint le minimum en $[1, 0, \ldots, 0], [0, 1, \ldots, 0], \ldots, [0, 0, \ldots, 1]$.
3. ϕ est une fonction symétrique de $p(1|t), p(2|t), \ldots, p(K|t)$.

Remarque.

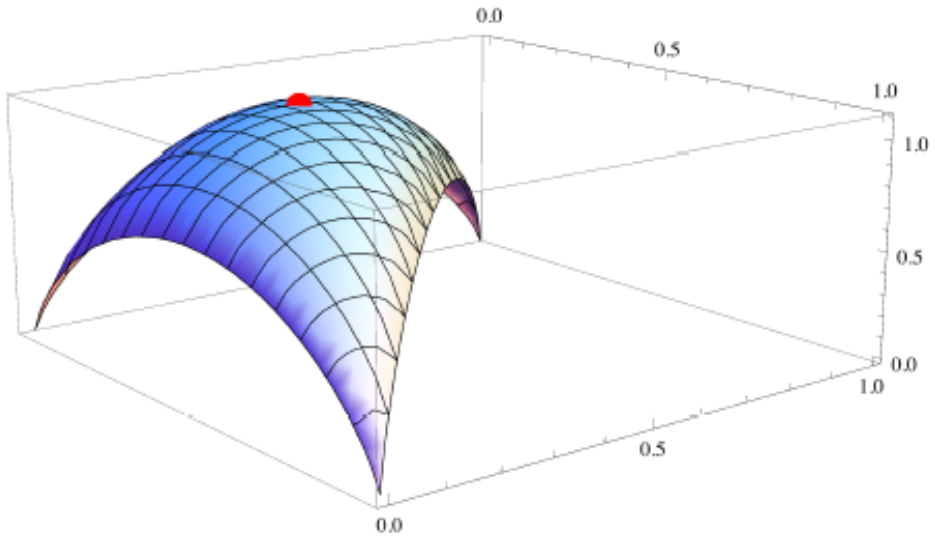
$Imp(t)$ est maximale quand toutes les classes sont mélangées avec des "parts égales" (Distribution uniforme/équiprobable) et est minimale quand le nœud ne contient qu'une seule classe (certitude totale).

Remarque.

Comme $p(1|t) + p(2|t) + \ldots + p(K|t) = 1$, on a

$$\begin{aligned} Imp(t) &= \phi(p(t)), \\ &= \phi(p(1|t), p(2|t), \ldots, p(K-1|t), \\ &\quad 1 - p(1|t) - p(2|t) - \ldots - p(K-1|t)) \end{aligned}$$

Ce qui montre que $Imp(t)$ est en fait une fonction de $K-1$ variables. Quand $K = 3$, on peut représenter une telle fonction et un exemple est : (le • indique le maximum de la fonction, atteint en $(1/3, 1/3)$) :



Exemples des mesures d'impureté :

1. Entropie :

$$Imp(t) = \mathcal{H}(t) = -\sum_{k=1}^{K} p(k|t) \log_2 p(k|t),$$

avec : $0 \log_2 0 = 0$.

2. Indice de Gini :

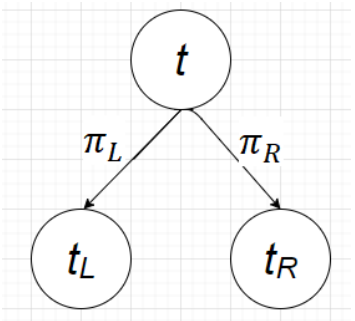
$$Imp(t) = \mathcal{G}(t) = 1 - \sum_{k=1}^{K} p^2(k|t).$$

Remarque.

Un nœud est pure s'il contient des données d'une seule classe. Dans ce cas $\mathcal{H}(t) = \mathcal{G}(t) = 0$.

Couper ou ne pas couper ?

On considère un nœud t avec deux nœuds fils t_L et t_R. On note cette opération de coupe par $\mathcal{S}$.



avec :

- π_L = proportion d'observations de t qui vont vers t_L.
- π_R = proportion d'observations de t qui vont vers t_R.

Couper ou ne pas couper ?

Qualité de la coupe $\mathcal{S}$ est définie par la variation de la mesure d'impureté :

$$\Phi(\mathcal{S}, t) = \Delta Imp(t) = Imp(t) - \pi_L Imp(t_L) - \pi_R Imp(t_R).$$

L'idée est de choisir une coupe $\mathcal{S}$ qui maximise $\Phi(\mathcal{S}, t)$. On a :
$\Phi(\mathcal{S}, t) \in [0, Imp(t)]$,

- Si $\Phi(\mathcal{S}, t) = 0$, le split n'a pas diminué l'impureté.
- Si $\Phi(\mathcal{S}, t) = Imp(t)$, alors $Imp(t_L) = Imp(t_R) = 0$ et les 2 nœuds descendants sont purs : le split a parfaitement séparé les groupes de t.

Définition 1.

L'impureté globale d'un arbre de décision T est définie par :

$$Imp(T) = \sum_{t \in \widetilde{T}} \pi(t)\, Imp(t),$$

où, $\widetilde{T}$ = l'ensemble des nœuds terminaux
et $\pi(t)$ = la proportion de la population globale en nœud t.