Arbres de classification

Chapitre 3

Marie Chavent

à partir des cours d'Adrien Todeschini et Robin Genuer

Master MIMSE - Université de Bordeaux

2015-2016

Références

- Classification and regression trees. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Chapman & Hall, 1984.
- Pattern classification. R. Duda, P. Hart and D. Stork. Wiley, New York, 2000.
- Bagging Predictors. L. Breiman. Machine Learning 26 :123-140.
- Random Forests. L. Breiman. Machine Learning 45 :5-32.

Introduction

- On parle d'arbres de décision pour désigner à la fois des arbres de classification et de régression
- Ici : arbres de classification

- **Principe général** : un arbre de classification est obtenu par un partitionnement récursif de l'espace d'entrée

- CART peut s'appliquer à des données comportant des variables d'entrée quantitatives et qualitatives et à une variable de sortie multiclasse.

- Fournit une représentation graphique simple du prédicteur obtenu, qui facilite l'interprétation des résultats.

- Deux étapes principales de CART :
  1. construction de l'arbre maximal
  2. élagage

Introduction

- On parle d'arbres de décision pour désigner à la fois des arbres de classification et de régression
- Ici : arbres de classification

- **Principe général** : un arbre de classification est obtenu par un partitionnement récursif de l'espace d'entrée

- CART peut s'appliquer à des données comportant des variables d'entrée quantitatives et qualitatives et à une variable de sortie multiclasse.

- Fournit une représentation graphique simple du prédicteur obtenu, qui facilite l'interprétation des résultats.

- Deux étapes principales de CART :
  1. construction de l'arbre maximal
  2. élagage

Introduction

- On parle d'arbres de décision pour désigner à la fois des arbres de classification et de régression
- Ici : arbres de classification
- **Principe général** : un arbre de classification est obtenu par un partitionnement récursif de l'espace d'entrée
- CART peut s'appliquer à des données comportant des variables d'entrée quantitatives et qualitatives et à une variable de sortie multiclasse.
- Fournit une représentation graphique simple du prédicteur obtenu, qui facilite l'interprétation des résultats.
- Deux étapes principales de CART :
  1. construction de l'arbre maximal
  2. élagage

Introduction

- On parle d'arbres de décision pour désigner à la fois des arbres de classification et de régression
- Ici : arbres de classification

- **Principe général** : un arbre de classification est obtenu par un partitionnement récursif de l'espace d'entrée

- CART peut s'appliquer à des données comportant des variables d'entrée quantitatives et qualitatives et à une variable de sortie multiclasse.

- Fournit une représentation graphique simple du prédicteur obtenu, qui facilite l'interprétation des résultats.

- Deux étapes principales de CART :
  1. construction de l'arbre maximal
  2. élagage

Introduction

- On parle d'arbres de décision pour désigner à la fois des arbres de classification et de régression
- Ici : arbres de classification

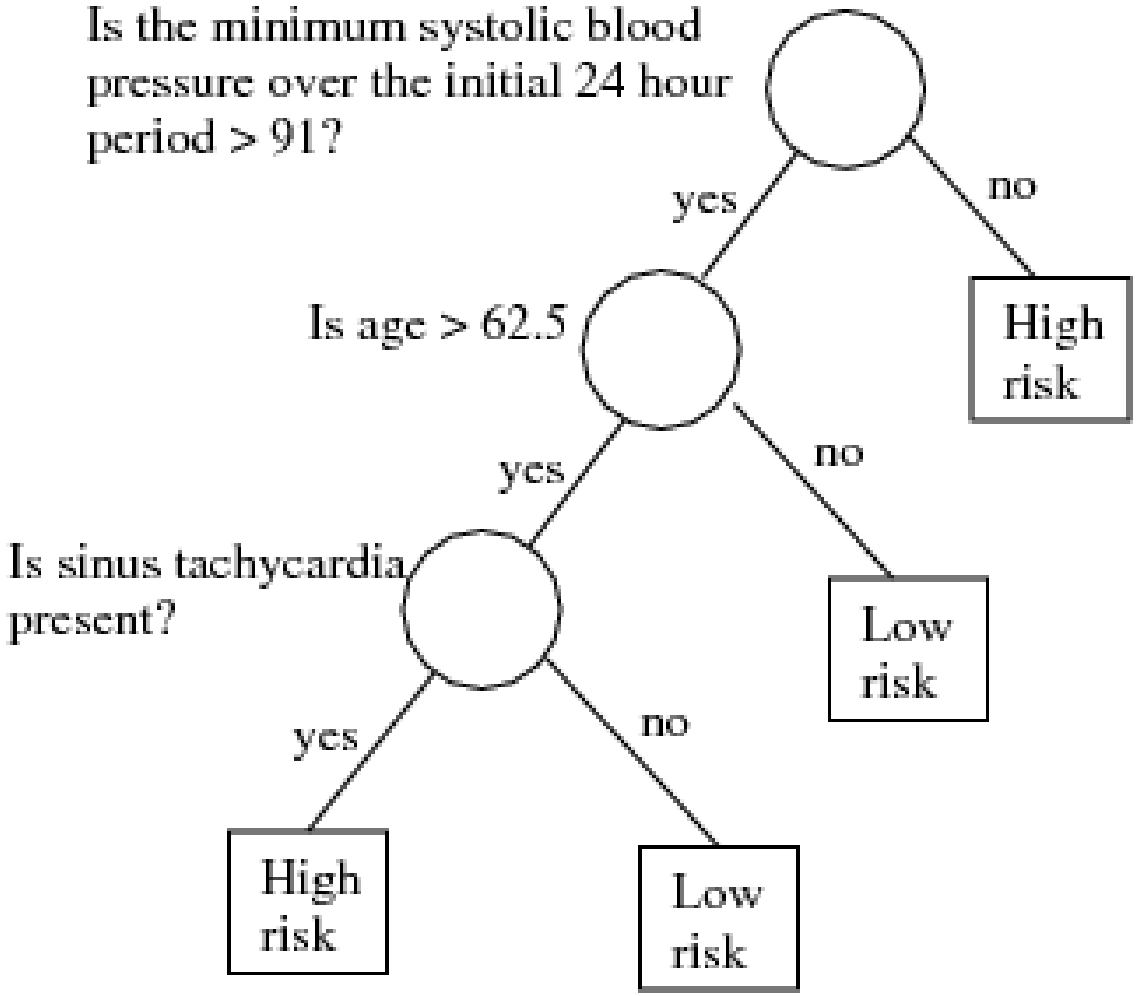
- **Principe général** : un arbre de classification est obtenu par un partitionnement récursif de l'espace d'entrée

- CART peut s'appliquer à des données comportant des variables d'entrée quantitatives et qualitatives et à une variable de sortie multiclasse.

- Fournit une représentation graphique simple du prédicteur obtenu, qui facilite l'interprétation des résultats.

- Deux étapes principales de CART :
  1. construction de l'arbre maximal
  2. élagage

Introduction

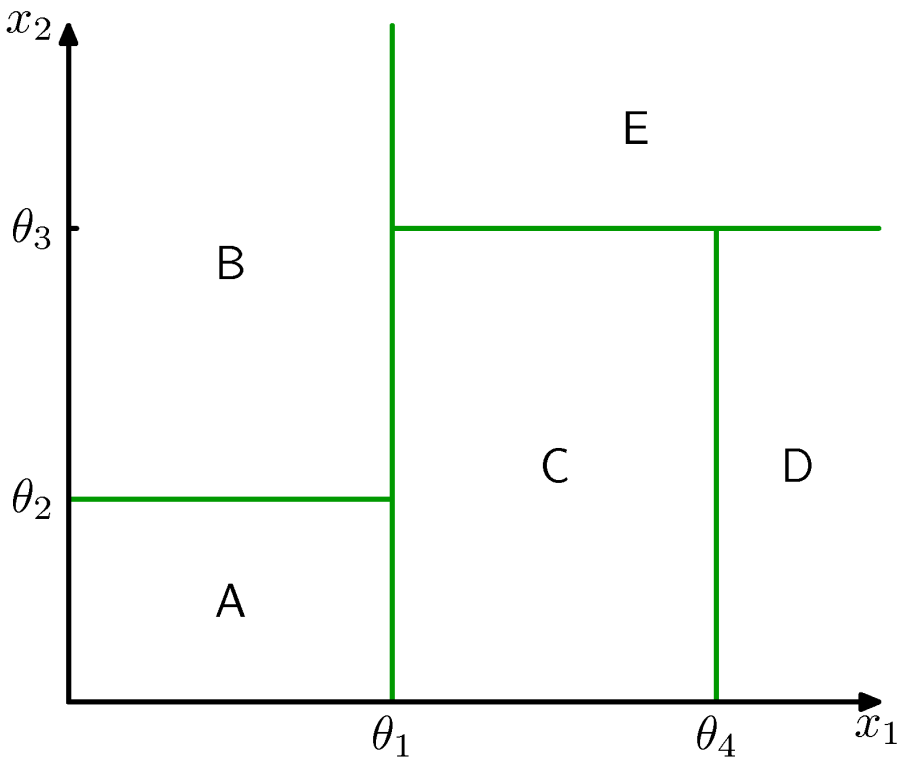


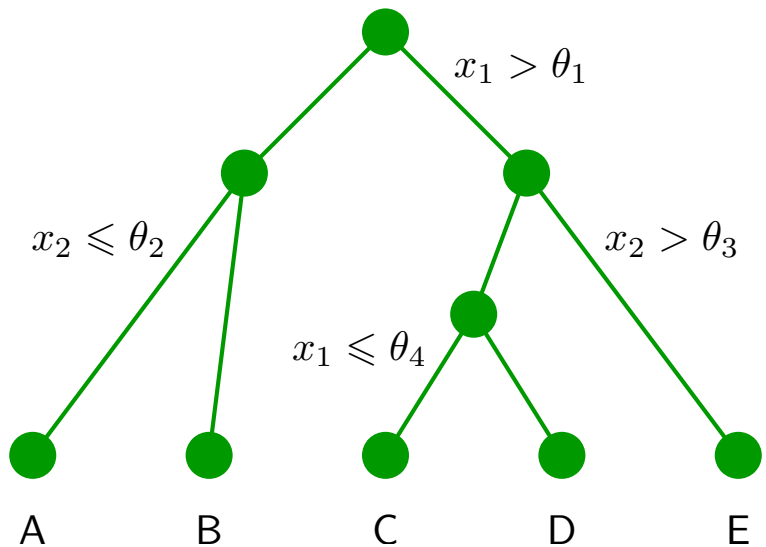
1. **Construction de l'arbre maximal**
2. Elagage

Notations

- Variables d'entrées : $\boldsymbol{X = (X^1, ..., X^p) \in \mathcal{X}}$, quantitatives ou qualitatives
- Variable de sortie : $\boldsymbol{Y \in \mathcal{Y} = \{1, ..., K\}}$, qualitative à $\boldsymbol{K}$ classes.
- On cherche à construire un classifieur associé à un arbre : $\boldsymbol{g : \mathcal{X} \to \mathcal{Y}}$ sur la base des observations d'apprentissage $\boldsymbol{(X_i, Y_i)}$, $\boldsymbol{i = 1, ..., n}$

Exemple d'arbre





Vocabulaire

- A chaque noeud t de l'arbre on associe une coupure (*split*)
- A chaque coupure est associée une variable de coupure X^{j_t} selon laquelle on découpe le noeud

- Variable **quantitative** $X_i^{j_t} \in \mathbb{R}$: on lui associe un seuil de coupure θ_t
  - si $X_i^{j_t} \leq \theta_t$ alors l'observation i va dans le noeud fils **gauche**
  - si $X_i^{j_t} > \theta_t$ alors l'observation i va dans le noeud fils **droit**

- Variable **qualitative** $X_i^{j_t} \in A$: on lui associe une partition des modalités en 2 groupes $\{A_t, {}^cA_t\}$:
  - si $X_i^{j_t} \in A_t$ alors l'observation i va dans le noeud fils **gauche**
  - si $X_i^{j_t} \notin A_t$ alors l'observation i va dans le noeud fils **droit**

Vocabulaire

- A chaque noeud t de l'arbre on associe une coupure (*split*)
- A chaque coupure est associée une variable de coupure X^{j_t} selon laquelle on découpe le noeud

- Variable **quantitative** $X_i^{j_t} \in \mathbb{R}$: on lui associe un seuil de coupure θ_t
  - si $X_i^{j_t} \leq \theta_t$ alors l'observation i va dans le noeud fils **gauche**
  - si $X_i^{j_t} > \theta_t$ alors l'observation i va dans le noeud fils **droit**

- Variable **qualitative** $X_i^{j_t} \in A$: on lui associe une partition des modalités en 2 groupes $\{A_t, {}^cA_t\}$:
  - si $X_i^{j_t} \in A_t$ alors l'observation i va dans le noeud fils **gauche**
  - si $X_i^{j_t} \notin A_t$ alors l'observation i va dans le noeud fils **droit**

Vocabulaire

- A chaque noeud t de l'arbre on associe une coupure (*split*)
- A chaque coupure est associée une variable de coupure X^{j_t} selon laquelle on découpe le noeud

- Variable **quantitative** $X_i^{j_t} \in \mathbb{R}$: on lui associe un seuil de coupure θ_t
  - si $X_i^{j_t} \leq \theta_t$ alors l'observation i va dans le noeud fils **gauche**
  - si $X_i^{j_t} > \theta_t$ alors l'observation i va dans le noeud fils **droit**

- Variable **qualitative** $X_i^{j_t} \in A$: on lui associe une partition des modalités en 2 groupes $\{A_t, {}^cA_t\}$:
  - si $X_i^{j_t} \in A_t$ alors l'observation i va dans le noeud fils **gauche**
  - si $X_i^{j_t} \notin A_t$ alors l'observation i va dans le noeud fils **droit**

Mesure de l'impureté (ou hétérogénéité)

- On va chercher, parmi toutes les coupures possibles celle qui sépare au mieux les classes :
  - i.e. donne deux noeuds fils les plus homogènes possibles
  - i.e. minimise une certaine fonction d'impureté

- On note R_t l'ensemble des données d'apprentissage appartenant au noeud t
  - La racine de l'arbre $R_1 = \{1, ..., n\}$ contient l'ensemble des données

- On définit la fréquence de la classe k dans le noeud t

$$\hat{p}_{t,k} = \frac{1}{Card(R_t)} \sum_{i \in R_t} \mathbb{1}_{Y_i = k}$$

- On note $\hat{p}_t = (\hat{p}_{t,1}, ..., \hat{p}_{t,K})$

Mesure de l'impureté (ou hétérogénéité)

- On va chercher, parmi toutes les coupures possibles celle qui sépare au mieux les classes :
  - i.e. donne deux noeuds fils les plus homogènes possibles
  - i.e. minimise une certaine fonction d'impureté

- On note $\boldsymbol{R_t}$ l'ensemble des données d'apprentissage appartenant au noeud $\boldsymbol{t}$
  - La racine de l'arbre $\boldsymbol{R_1 = \{1, ..., n\}}$ contient l'ensemble des données

- On définit la fréquence de la classe k dans le noeud t

$$\hat{p}_{t,k} = \frac{1}{Card(R_t)} \sum_{i \in R_t} \mathbb{1}_{Y_i = k}$$

- On note $\hat{p}_t = (\hat{p}_{t,1}, \ldots, \hat{p}_{t,K})$

Mesure de l'impureté (ou hétérogénéité)

- On va chercher, parmi toutes les coupures possibles celle qui sépare au mieux les classes :
  - i.e. donne deux noeuds fils les plus homogènes possibles
  - i.e. minimise une certaine fonction d'impureté

- On note $\boldsymbol{R_t}$ l'ensemble des données d'apprentissage appartenant au noeud $\boldsymbol{t}$
  - La racine de l'arbre $\boldsymbol{R_1 = \{1, ..., n\}}$ contient l'ensemble des données

- On définit la fréquence de la classe $\boldsymbol{k}$ dans le noeud $\boldsymbol{t}$

$$\hat{\boldsymbol{p}}_{\boldsymbol{t,k}} = \frac{\mathbf{1}}{\boldsymbol{Card(R_t)}} \sum_{\boldsymbol{i \in R_t}} \mathbb{1}_{\boldsymbol{Y_i = k}}$$

- On note $\hat{\boldsymbol{p}}_{\boldsymbol{t}} = (\hat{\boldsymbol{p}}_{\boldsymbol{t,1}}, ..., \hat{\boldsymbol{p}}_{\boldsymbol{t,K}})$

Mesure de l'impureté (ou hétérogénéité)

Fonctions d'impureté standard :

- Indice de Gini :

$$\psi(\hat{p}_t) = \sum_{k=1}^{K} \hat{p}_{t,k}(1 - \hat{p}_{t,k})$$

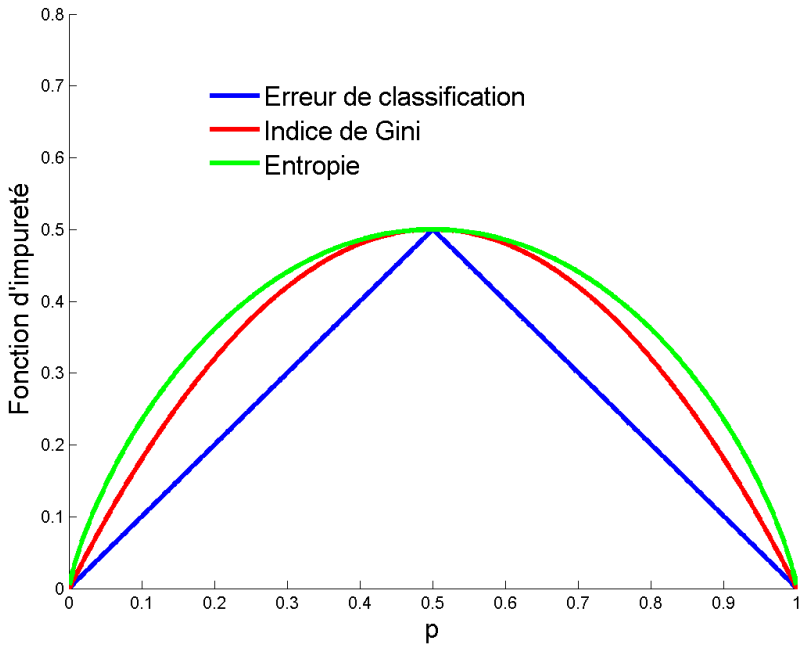
- Entropie :

$$\psi(\hat{p}_t) = -\sum_{k=1}^{K} \hat{p}_{t,k} \log(\hat{p}_{t,k})$$

Mesure de l'impureté (ou hétérogénéité)

Cas binaire : $K = 2$

- Gini : $\psi(\hat{p}_t) = \hat{p}_{t,1}(1 - \hat{p}_{t,1}) + \hat{p}_{t,2}(1 - \hat{p}_{t,2}) = 2\hat{p}_{t,1}(1 - \hat{p}_{t,1})$
- Entropie : $\psi(\hat{p}_t) = -\hat{p}_{t,1}\log(\hat{p}_{t,1}) - (1 - \hat{p}_{t,1})\log(1 - \hat{p}_{t,1})$

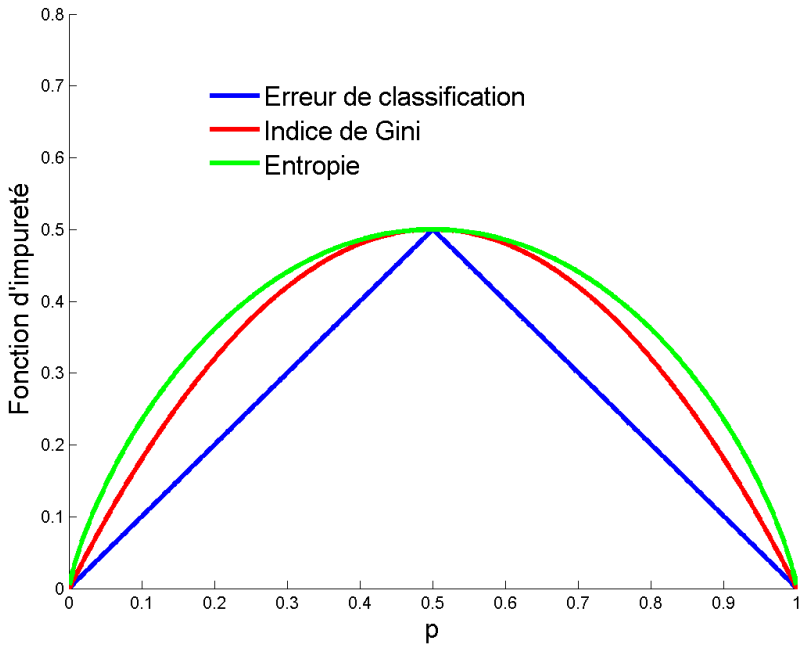


- positive
- nulle si le noeud est constitué d'observations de la même classe (noeud est pur)
- maximale si la moitié des obs. sont de classe 1 et l'autre moitié de classe 2.

Mesure de l'impureté (ou hétérogénéité)

Cas binaire : $\boldsymbol{K = 2}$

- Gini : $\psi(\hat{p}_t) = \hat{p}_{t,1}(1 - \hat{p}_{t,1}) + \hat{p}_{t,2}(1 - \hat{p}_{t,2}) = 2\hat{p}_{t,1}(1 - \hat{p}_{t,1})$
- Entropie : $\psi(\hat{p}_t) = -\hat{p}_{t,1}\log(\hat{p}_{t,1}) - (1 - \hat{p}_{t,1})\log(1 - \hat{p}_{t,1})$



- positive
- nulle si le noeud est constitué d'observations de la même classe (noeud est pur)
- maximale si la moitié des obs. sont de classe 1 et l'autre moitié de classe 2.

Mesure de l'impureté (ou hétérogénéité)

- On note t_L et t_R les noeuds fils gauche et droit du noeud t.
- On mesure la qualité d'une coupure c_t par

$$\Delta(c_t) = \psi(\hat{p}_t) - \pi_L \psi(\hat{p}_{t_L}) - \pi_R \psi(\hat{p}_{t_R})$$

où π_L et π_R sont les proportions de données de R_t appartenant à R_{t_L} et R_{t_R}.

- On choisit la coupure c_t associée au noeud t qui vérifie

$$c_t = \underset{c_t = \{X^{j_t}, \theta_t\} \text{ ou } c_t = \{X^{j_t}, A_t\}}{\arg\max} \Delta(c_t)$$

Mesure de l'impureté (ou hétérogénéité)

- On note t_L et t_R les noeuds fils gauche et droit du noeud t.
- On mesure la qualité d'une coupure c_t par

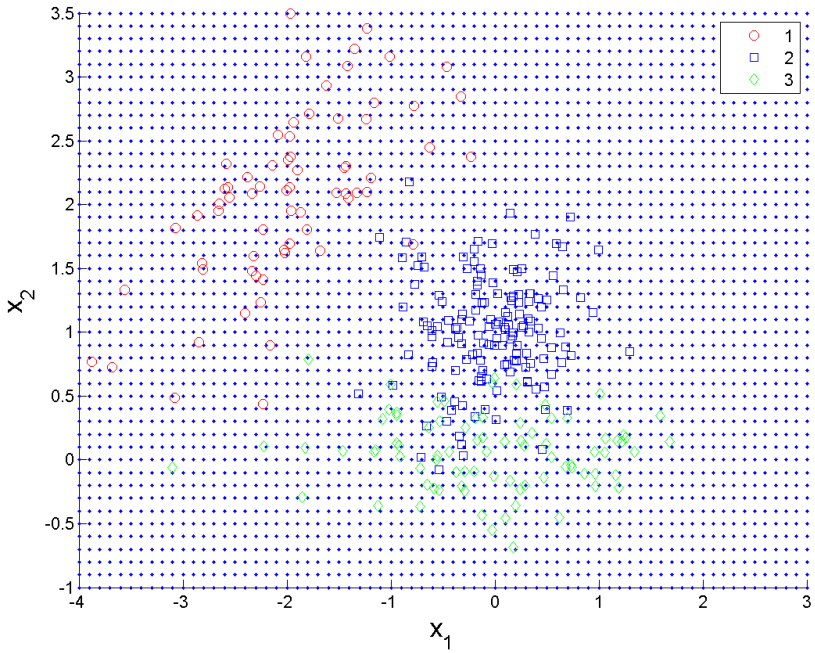
$$\Delta(c_t) = \psi(\hat{p}_t) - \pi_L \psi(\hat{p}_{t_L}) - \pi_R \psi(\hat{p}_{t_R})$$

où π_L et π_R sont les proportions de données de R_t appartenant à R_{t_L} et R_{t_R}.

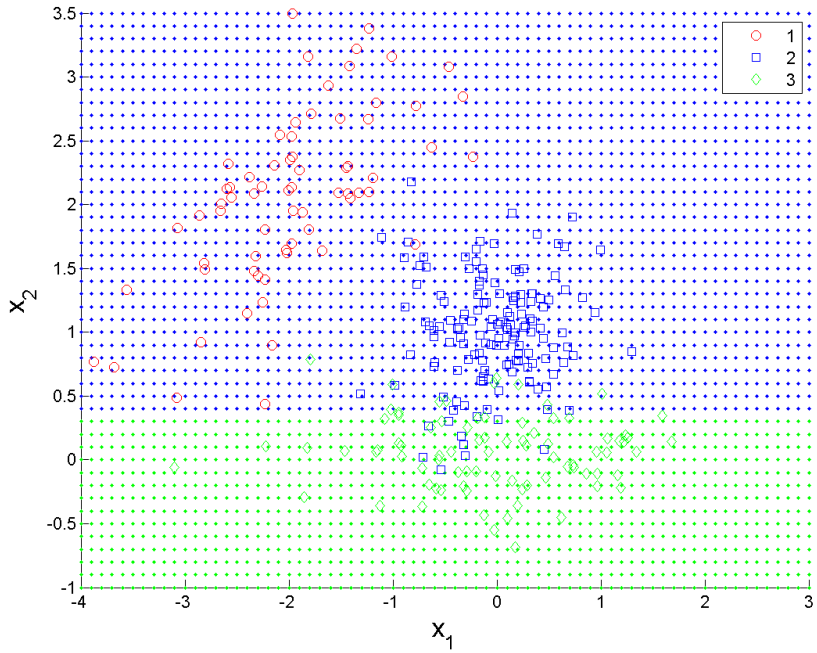
- On choisit la coupure c_t associée au noeud t qui vérifie

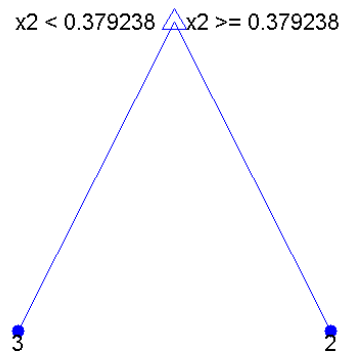
$$c_t = \underset{c_t = \{X^{j_t}, \theta_t\} \text{ ou } c_t = \{X^{j_t}, A_t\}}{\arg\max} \Delta(c_t)$$

Exemple : données synthétiques

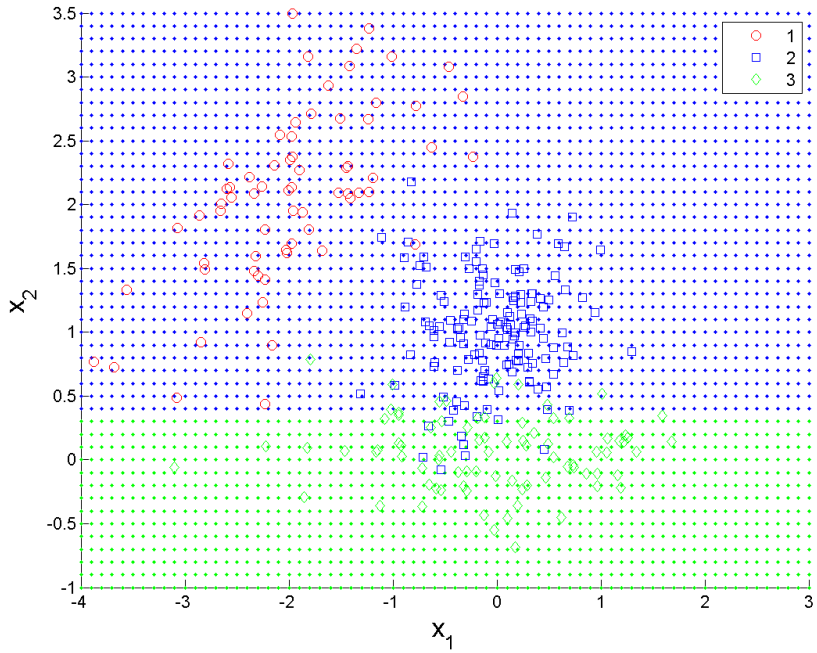


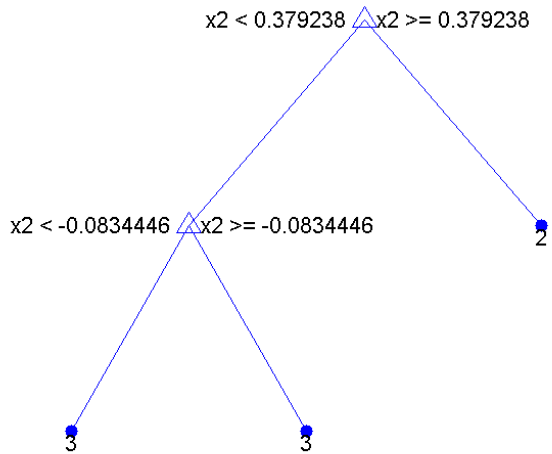
Exemple : données synthétiques



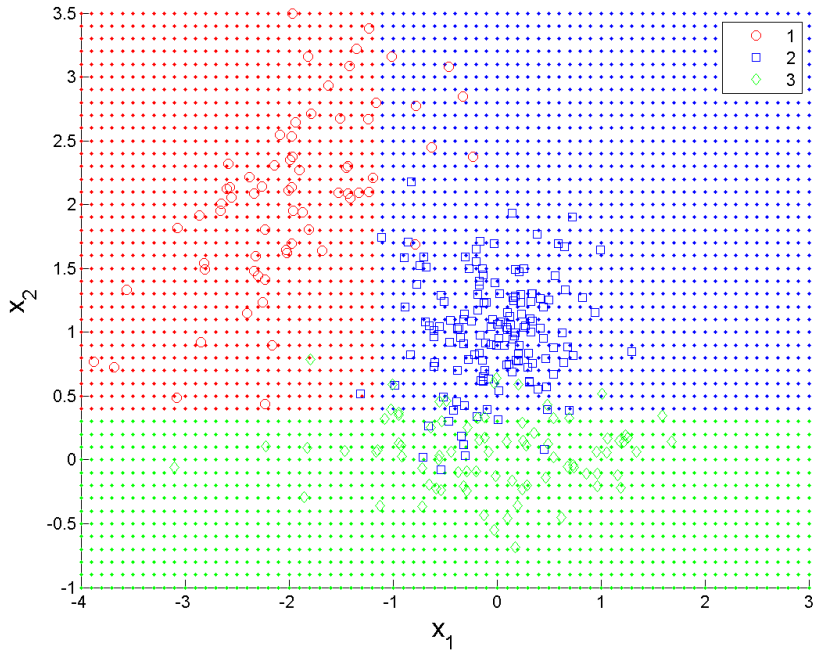


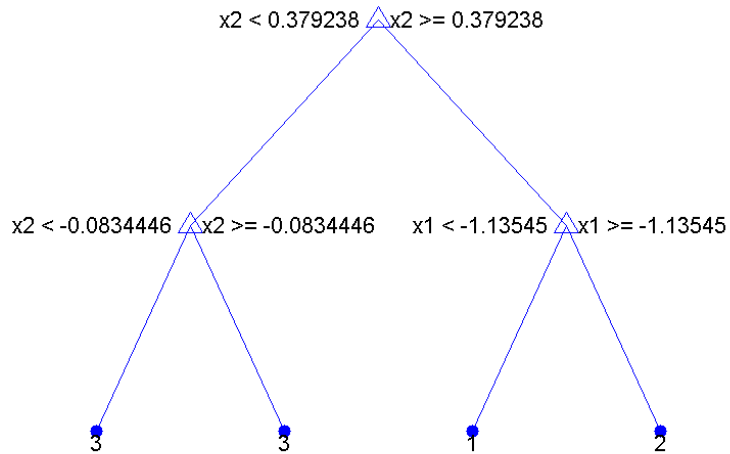
Exemple : données synthétiques



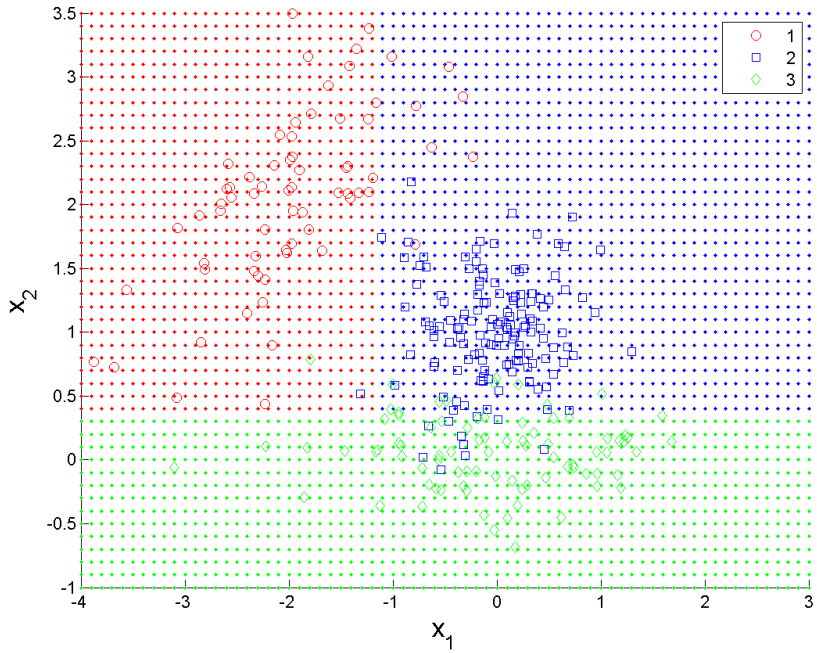


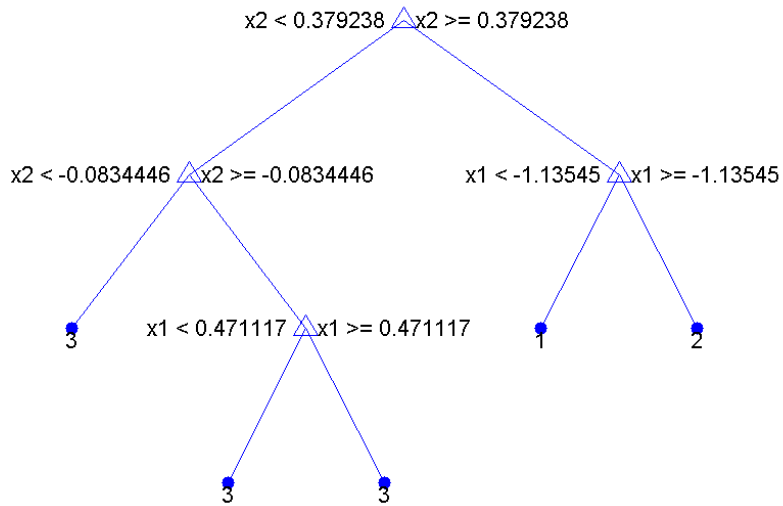
Exemple : données synthétiques



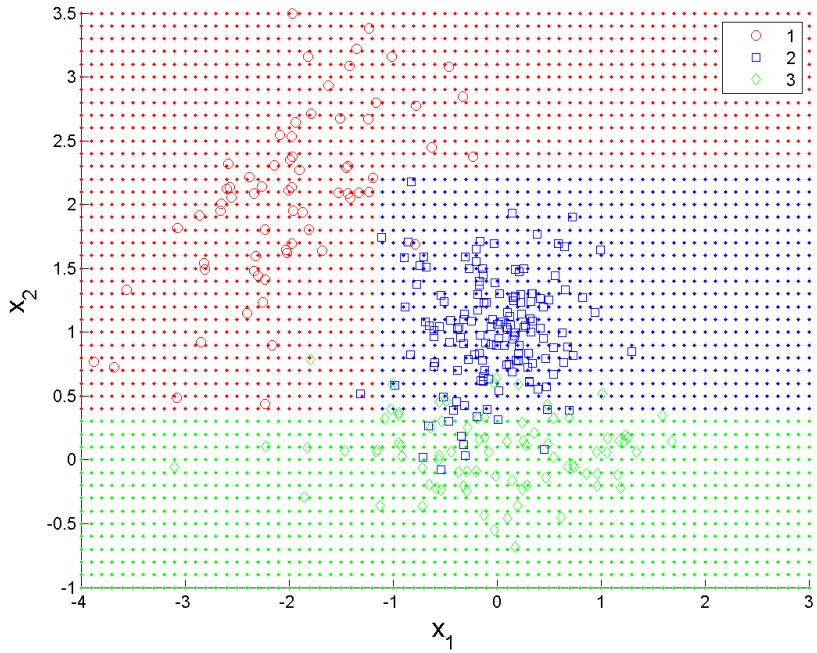


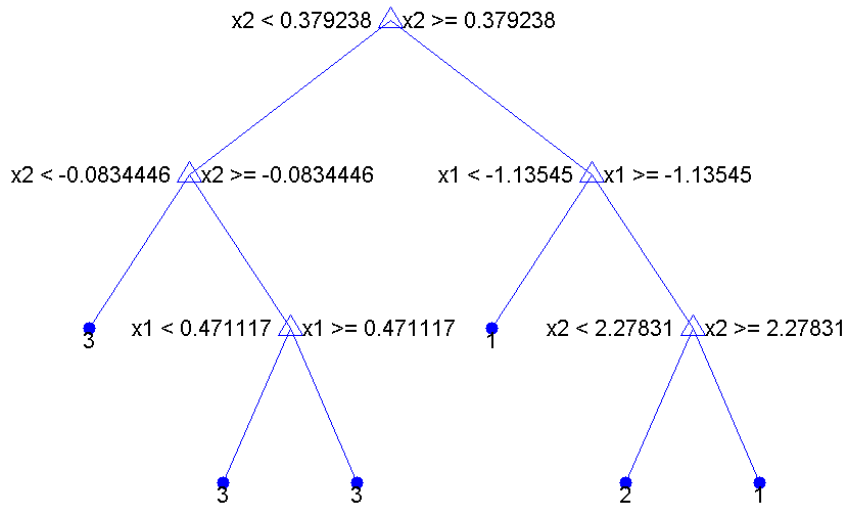
Exemple : données synthétiques



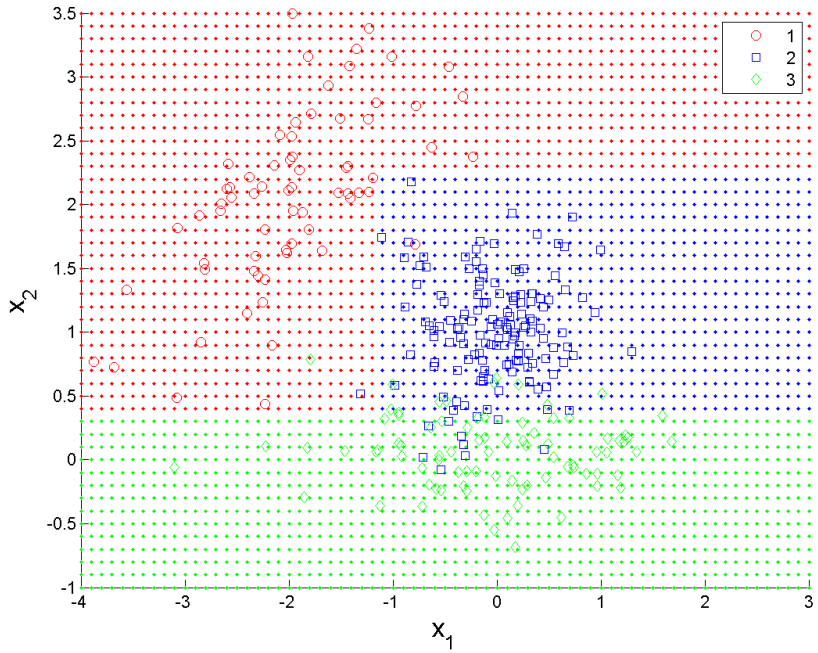


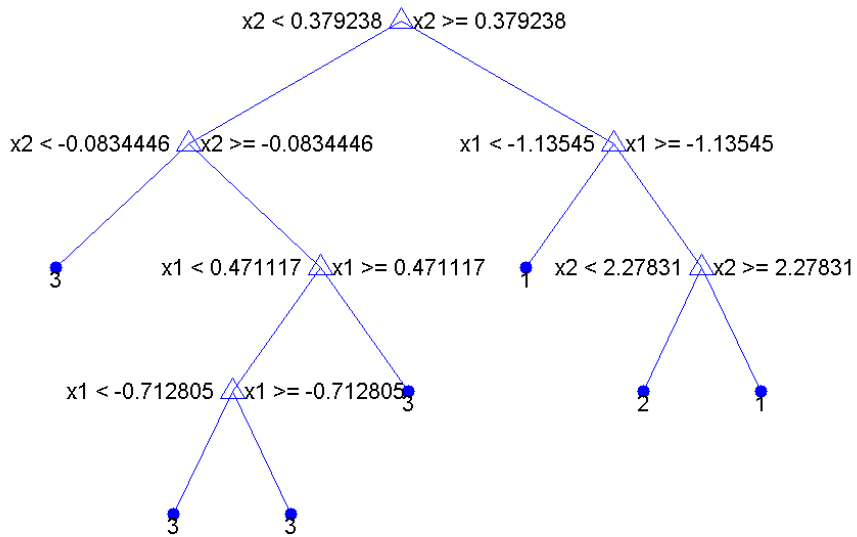
Exemple : données synthétiques



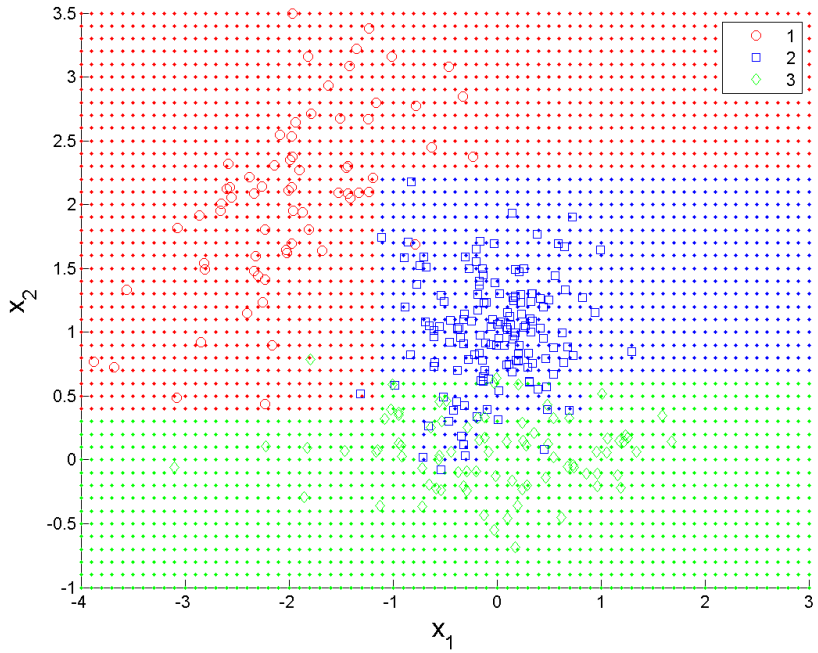


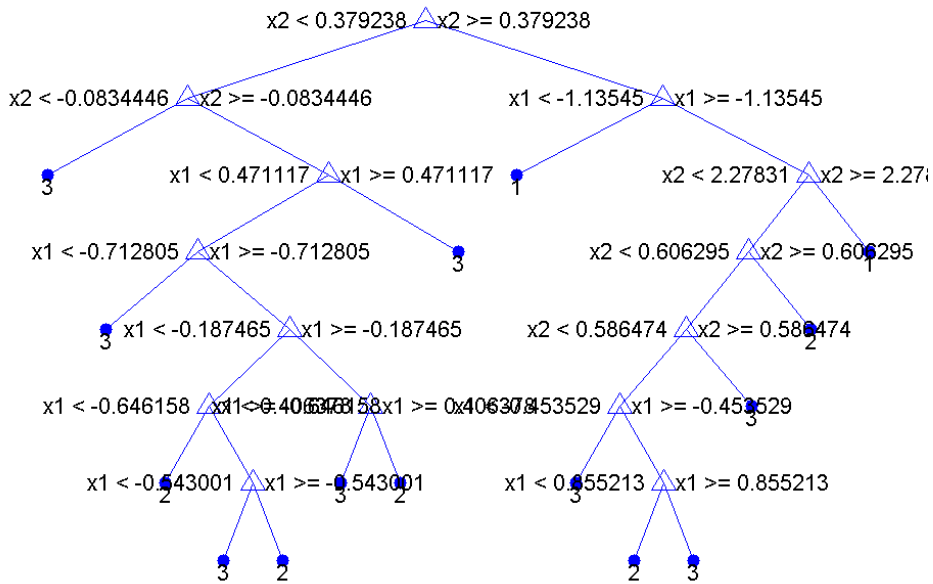
Exemple : données synthétiques



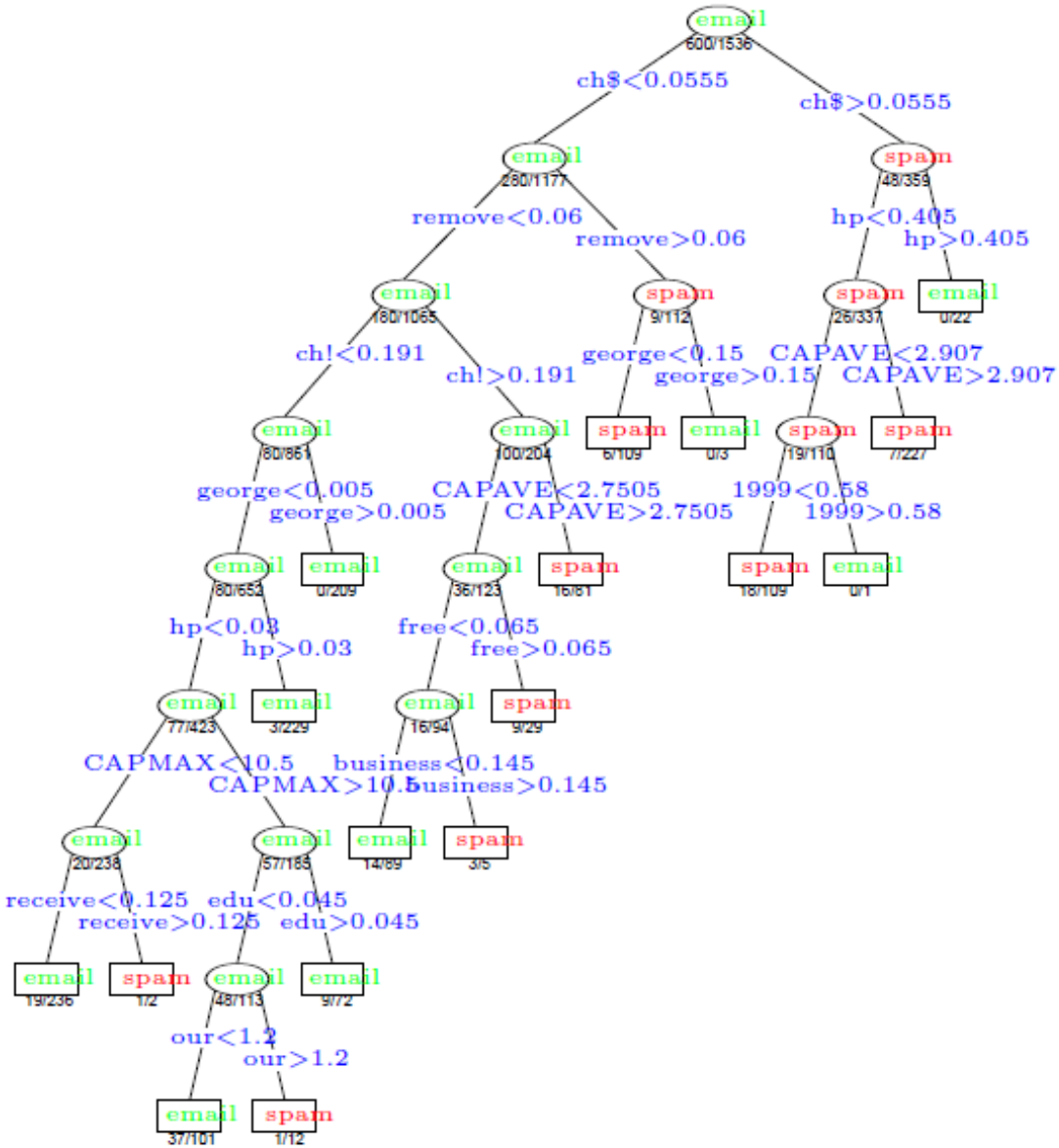


Exemple : données synthétiques





Exemple : spam



Critère d'arrêt

De façon générale, on ne découpe pas un noeud pur.
Ensuite, plusieurs choix possibles

- On ne découpe pas un noeud qui contient moins de n_{min} données.
  - Si $n_{min} = 1$, on découpe les données au maximum (arbre maximal).
  - Si $n_{min} > 1$, on obtient un arbre moins profond.
  - choix de n_{min} difficile
- On ne découpe pas un noeud si $\max_{c_t} \Delta(c_t) < \epsilon$, donc si le gain en pureté n'est pas suffisant
  - choix de ϵ difficile

Critère d'arrêt

De façon générale, on ne découpe pas un noeud pur.
Ensuite, plusieurs choix possibles

- On ne découpe pas un noeud qui contient moins de n_{min} données.
  - Si $n_{min} = 1$, on découpe les données au maximum (arbre maximal).
  - Si $n_{min} > 1$, on obtient un arbre moins profond.
  - choix de n_{min} difficile

- On ne découpe pas un noeud si $\max_{c_t} \Delta(c_t) < \epsilon$, donc si le gain en pureté n'est pas suffisant
  - choix de ϵ difficile

Règle de classification

- A chaque noeud terminal (ou feuille) on associe la classe majoritaire des observations appartenant à ce noeud :

$$k_t = \arg\max_k \hat{p}_{t,k}$$

- La règle de classification est :

$$g(x) = k_{F(x)}$$

où $F(x)$ est la feuille dans laquelle tombe l'observation x.

1. Construction de l'arbre maximal
2. **Elagage**

Elagage

Les critères d'arrêt étant difficiles à régler en pratique, Breiman et al. ont proposé une meilleure stratégie :

- Partant de l'arbre maximal $\boldsymbol{T_{max}}$, on construit une suite de sous-arbres élagués de $\boldsymbol{T_{max}}$
- On choisit l'arbre final parmi cette collection d'arbres

Construction de la suite d'arbres

Critère de coût-complexité à minimiser :

$$C_\alpha(T) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g_T(X_i)\neq Y_i} + \alpha\frac{|T|}{n}$$

où $|T|$ est le nombre de feuilles de l'arbre T

- Si $\alpha = 0$, $T_{max} = \arg\min_T C_\alpha(T)$ avec $C_0(T_{max}) = 0$ (taux d'erreur nul)
- On augmente légèrement α, l'arbre qui minimise $C_\alpha(T)$ est un arbre élagué de T_{max}, noté $T_{|T|-1}$, pour lequel on a regroupé deux feuilles dans leur noeud père, qui devient alors un noeud terminal.
- On repart de $T_{|T|-1}$ et on augmente encore α
- et ainsi de suite ...

Construction de la suite d'arbres

Critère de coût-complexité à minimiser :

$$C_\alpha(T) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g_T(X_i)\neq Y_i} + \alpha\frac{|T|}{n}$$

où $|T|$ est le nombre de feuilles de l'arbre T

- Si $\alpha = 0$, $T_{max} = \arg\min_T C_\alpha(T)$ avec $C_0(T_{max}) = 0$ (taux d'erreur nul)
- On augmente légèrement α, l'arbre qui minimise $C_\alpha(T)$ est un arbre élagué de T_{max}, noté $T_{|T|-1}$, pour lequel on a regroupé deux feuilles dans leur noeud père, qui devient alors un noeud terminal.
- On repart de $T_{|T|-1}$ et on augmente encore α
- et ainsi de suite ...

Construction de la suite d'arbres

- On obtient une suite de sous-arbres élagués :

$$T_{max} = T_{|T|} \supset T_{|T|-1} \supset ... \supset T_1 = \text{racine}$$

- Par construction, l'arbre T_l est élagué de l'arbre T_{l-1}
- On peut montrer que

$$T_l = \arg\min_{T} C_{\alpha_l}(T)$$

où α_l est la valeur de α pour laquelle on a déterminé T_l

- On choisit $l_{opt} \in \{1, ..., |T|\}$ et l'arbre T_{opt} associé par validation croisée

Limites des arbres de classification

- **Instabilité**
  - Un faible changement dans les données d'apprentissage peut donner un arbre très différent
  - Du à la construction hiérarchique de l'arbre
  - Solution : méthodes d'ensemble bagging et random forests (Breiman 1996, 2001)

- **Partitions binaires ?**
  - On pourrait considérer un partitionnement en plus de deux noeuds.
  - Cette stratégie est en général moins bonne, les données d'apprentissage étant trop vite fragmentées

- **Combinaisons linéaires**
  - On peut considérer des règles de partition de la forme $\sum a_j X^j \leq \theta_t$
  - Poids a_j peuvent être obtenus par optimisation
  - Mais perte de l'interprétation et augmentation significative du temps de calcul

Limites des arbres de classification

- **Instabilité**
  - Un faible changement dans les données d'apprentissage peut donner un arbre très différent
  - Du à la construction hiérarchique de l'arbre
  - Solution : méthodes d'ensemble bagging et random forests (Breiman 1996, 2001)

- **Partitions binaires ?**
  - On pourrait considérer un partitionnement en plus de deux noeuds.
  - Cette stratégie est en général moins bonne, les données d'apprentissage étant trop vite fragmentées

- **Combinaisons linéaires**
  - On peut considérer des règles de partition de la forme $\sum a_j X^j \leq \theta_t$
  - Poids a_j peuvent être obtenus par optimisation
  - Mais perte de l'interprétation et augmentation significative du temps de calcul

Limites des arbres de classification

- **Instabilité**
  - Un faible changement dans les données d'apprentissage peut donner un arbre très différent
  - Du à la construction hiérarchique de l'arbre
  - Solution : méthodes d'ensemble bagging et random forests (Breiman 1996, 2001)

- **Partitions binaires ?**
  - On pourrait considérer un partitionnement en plus de deux noeuds.
  - Cette stratégie est en général moins bonne, les données d'apprentissage étant trop vite fragmentées

- **Combinaisons linéaires**
  - On peut considérer des règles de partition de la forme $\sum a_j X^j \leq \theta_t$
  - Poids a_j peuvent être obtenus par optimisation
  - Mais perte de l'interprétation et augmentation significative du temps de calcul