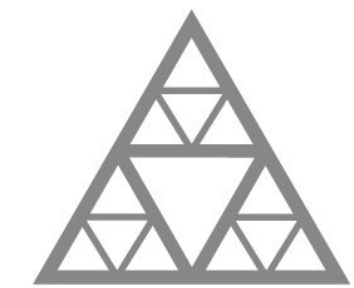
Pratique de l'apprentissage statistique

5. Arbres de régression et de décision

V. Lefieux



École des Ponts
ParisTech

Plan

Plan

Une référence

(Breiman et collab., 1984)

Objectifs

Prévoir :

- une variable continue $Y \in \mathbb{R}$ (régression),
- une variable qualitative Y à K classes $\{1, \ldots, K\}$ (classification supervisée),

à partir de p covariables $(X_1, \ldots, X_p)$.

Principes I

Classification And Regression Tree, CART (Breiman et al, 1984), est une méthode récursive :

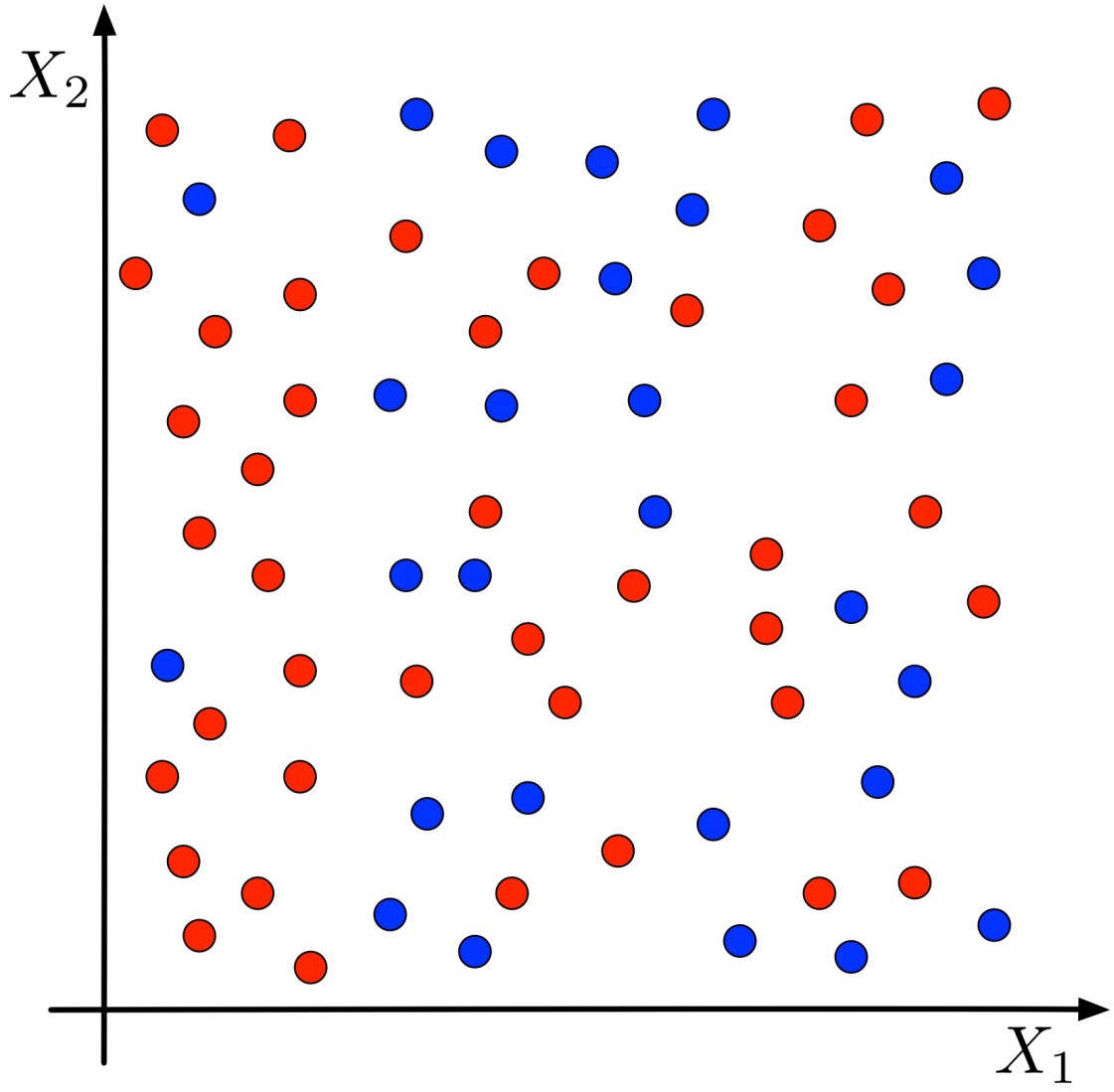
- A la racine on trouve tout l'échantillon.
- Chaque noeud de l'arbre divise l'échantillon en 2 branches, selon une variable discrète, continue ou ordinale (seuil) ou une variable nominale (ensemble de catégories).
- Un noeud terminal est appelé feuille.

Principes II

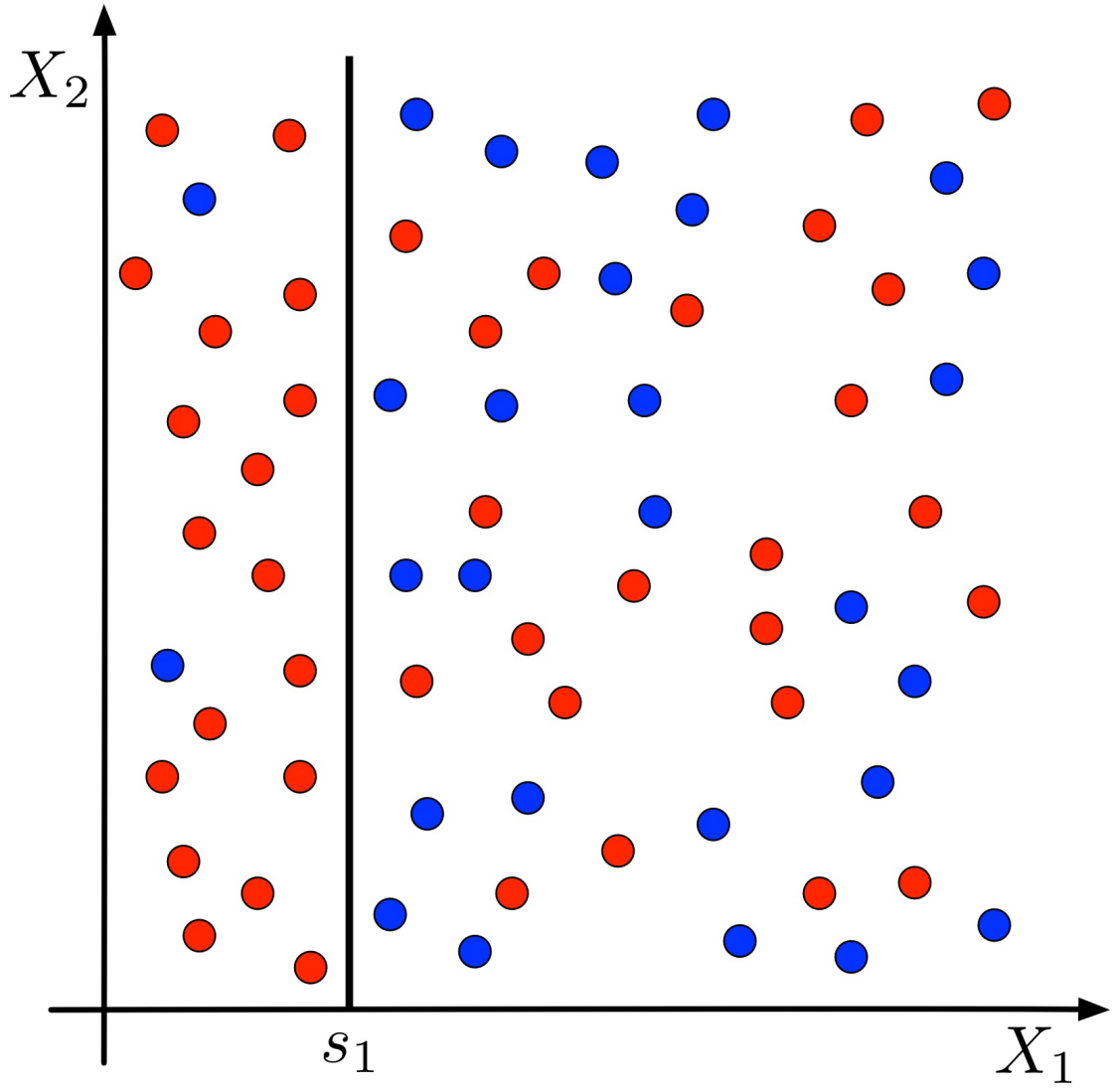
- On obtient une partition de l'espace en pavés (pavage binaire récursif).
- On ajuste sur chaque pavé un modèle simple :
  - Cas de la classification supervisée : vote majoritaire.
  - Cas de la régression : moyenne.

Plan

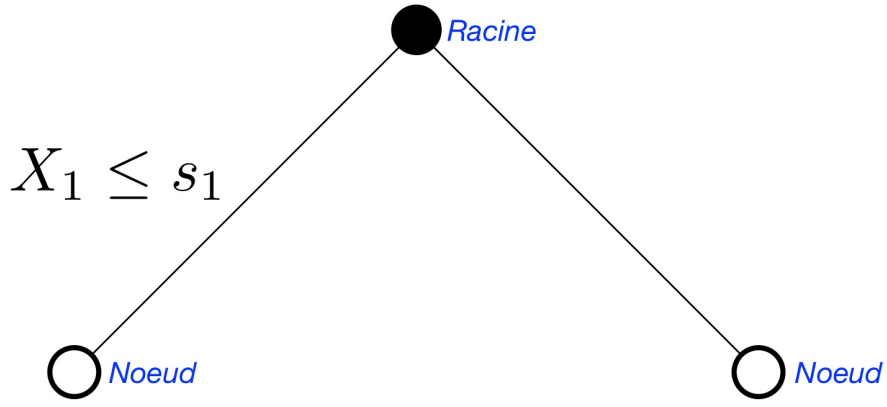
Exemple (Classification) I



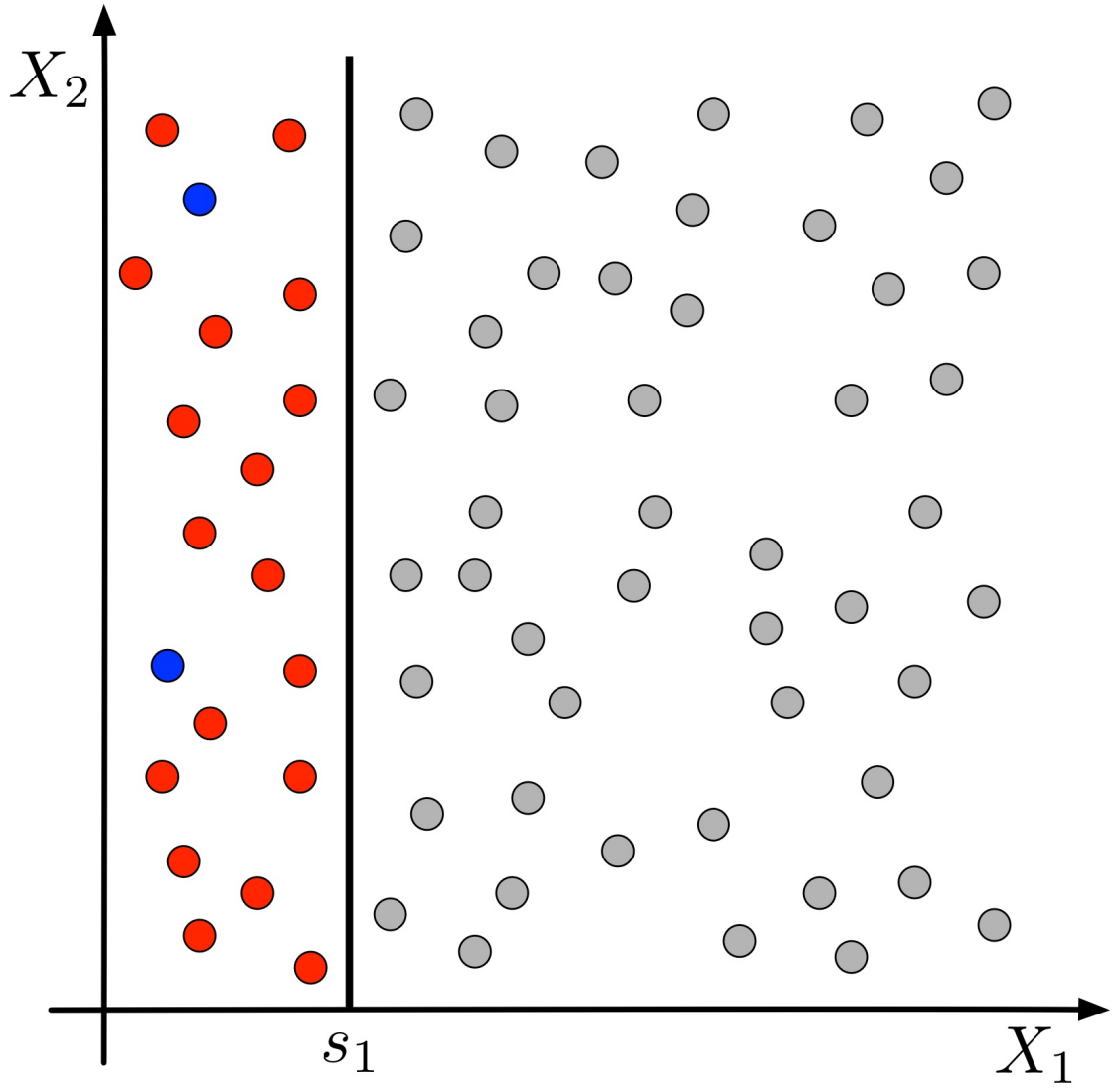
Exemple (Classification) II



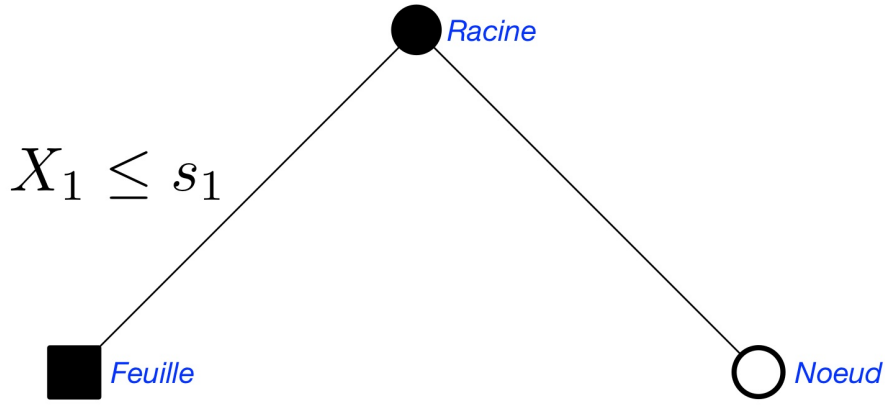
Exemple (Classification) III



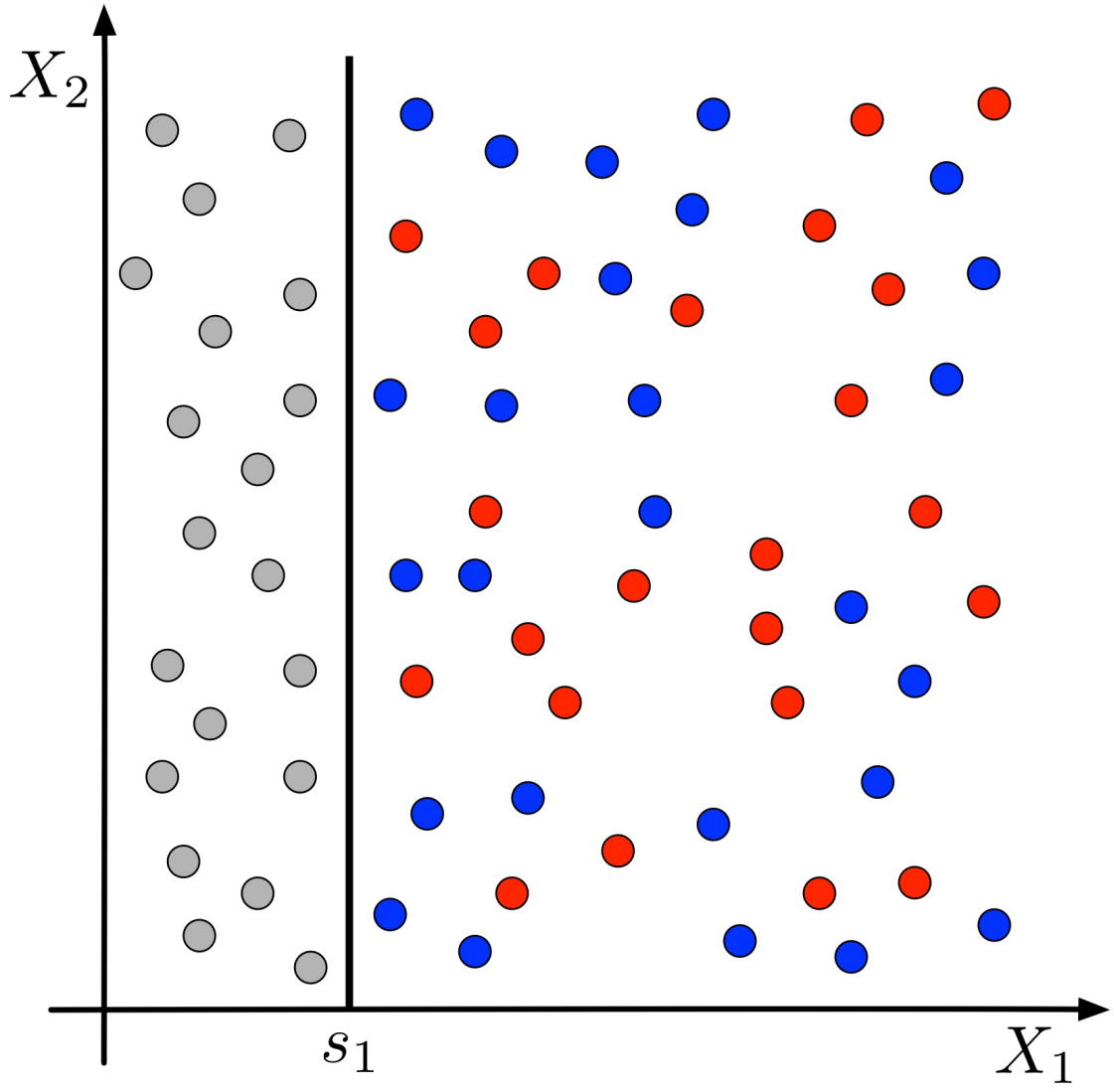
Exemple (Classification) IV



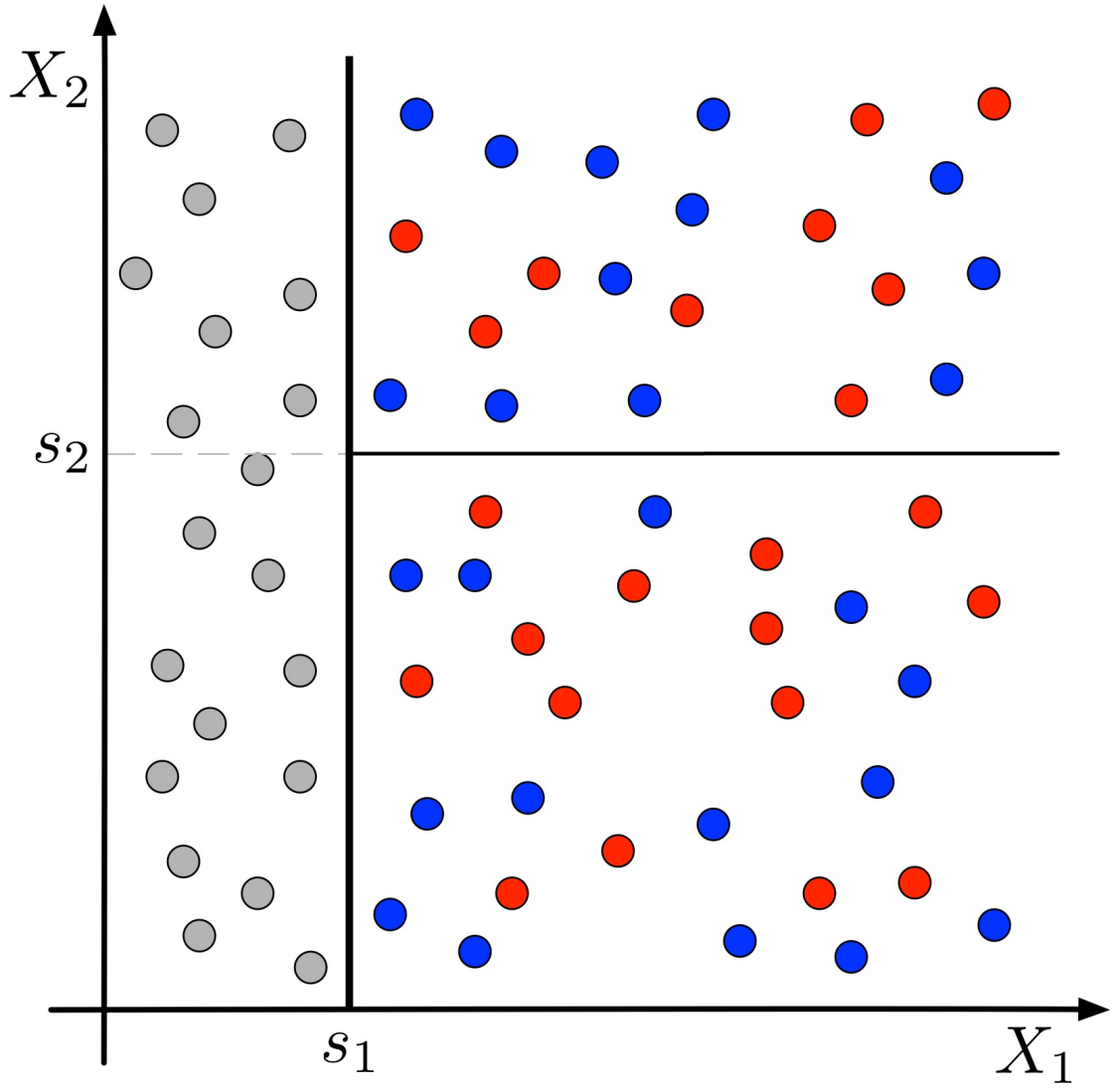
Exemple (Classification) V



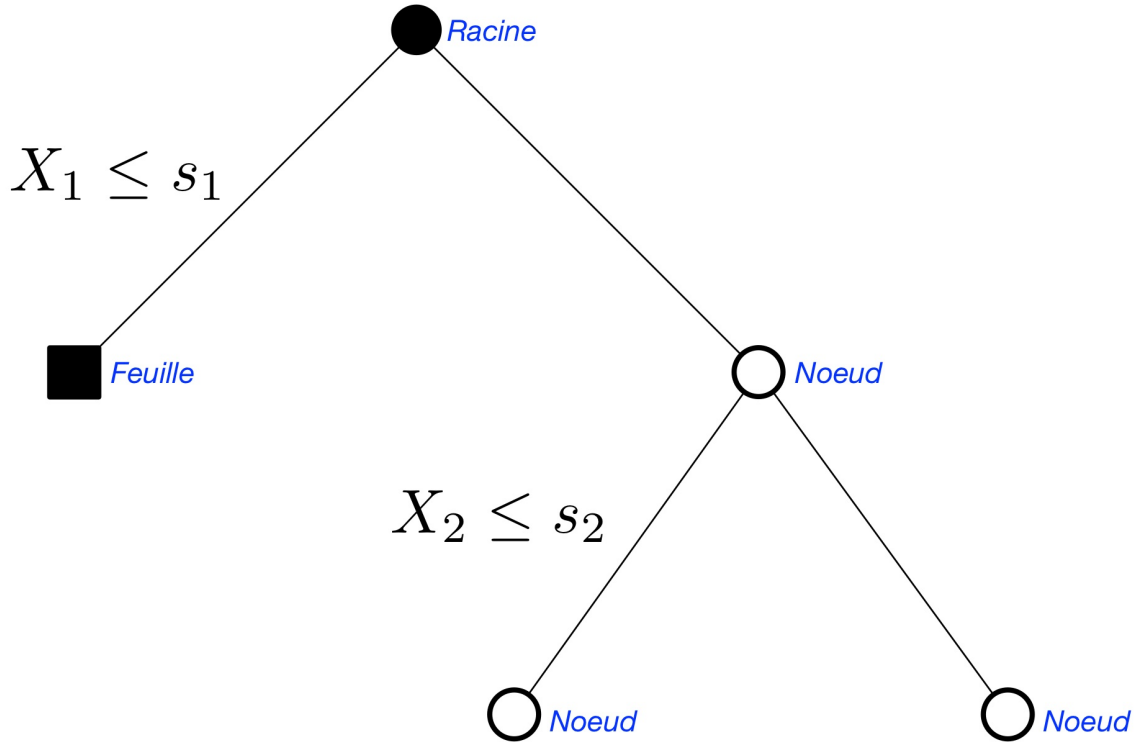
Exemple (Classification) VI



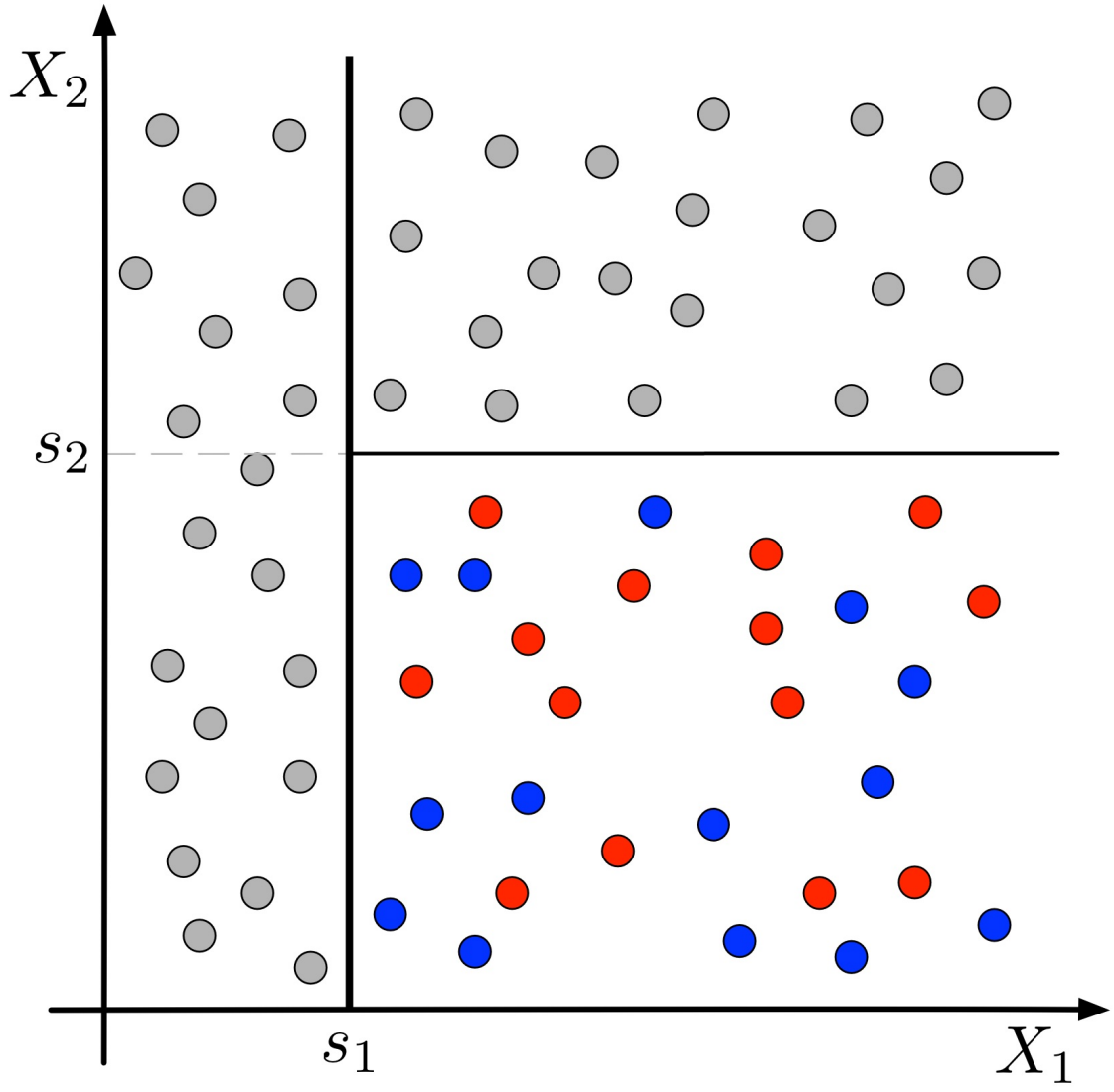
Exemple (Classification) VII



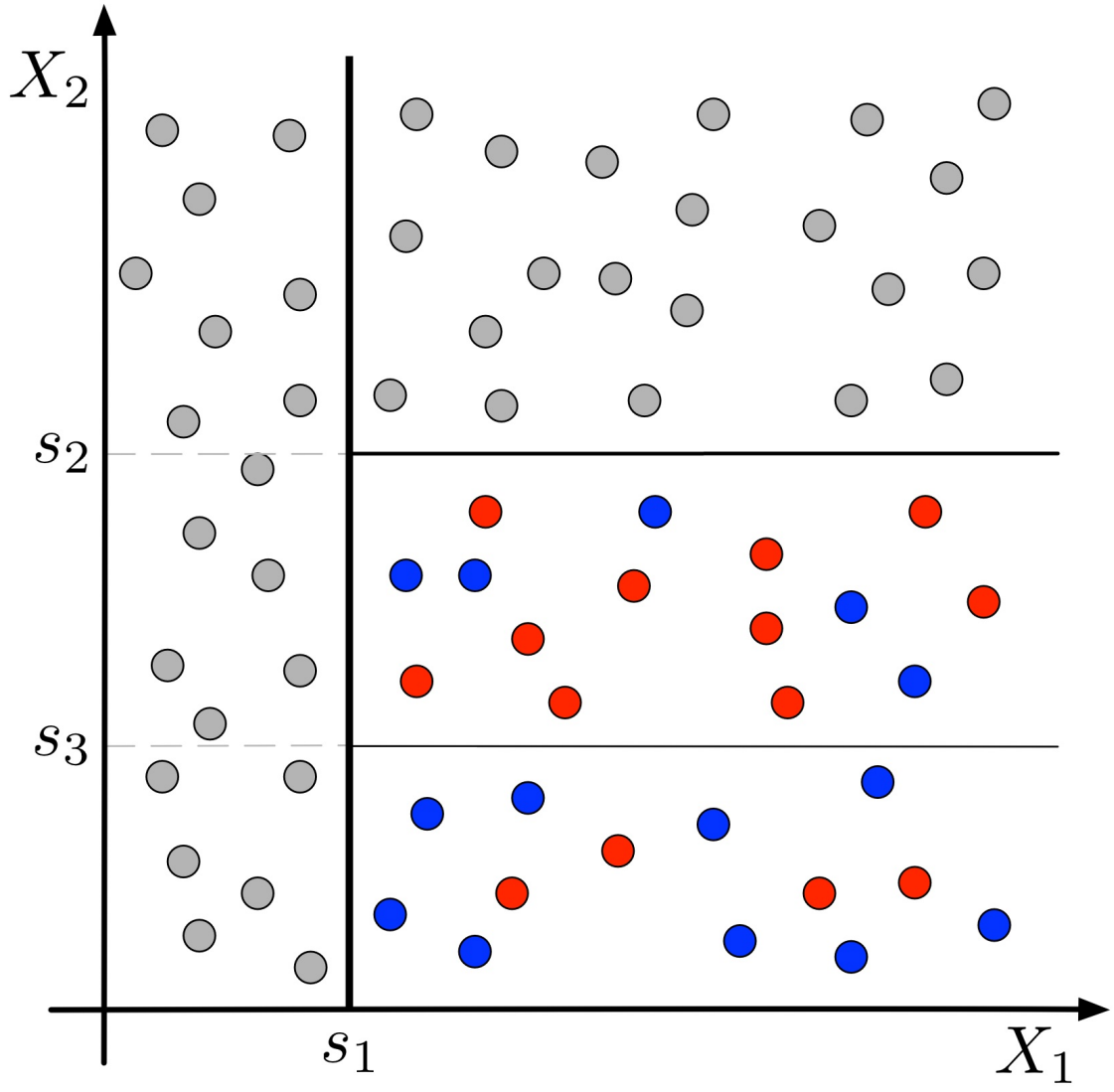
Exemple (Classification) VIII



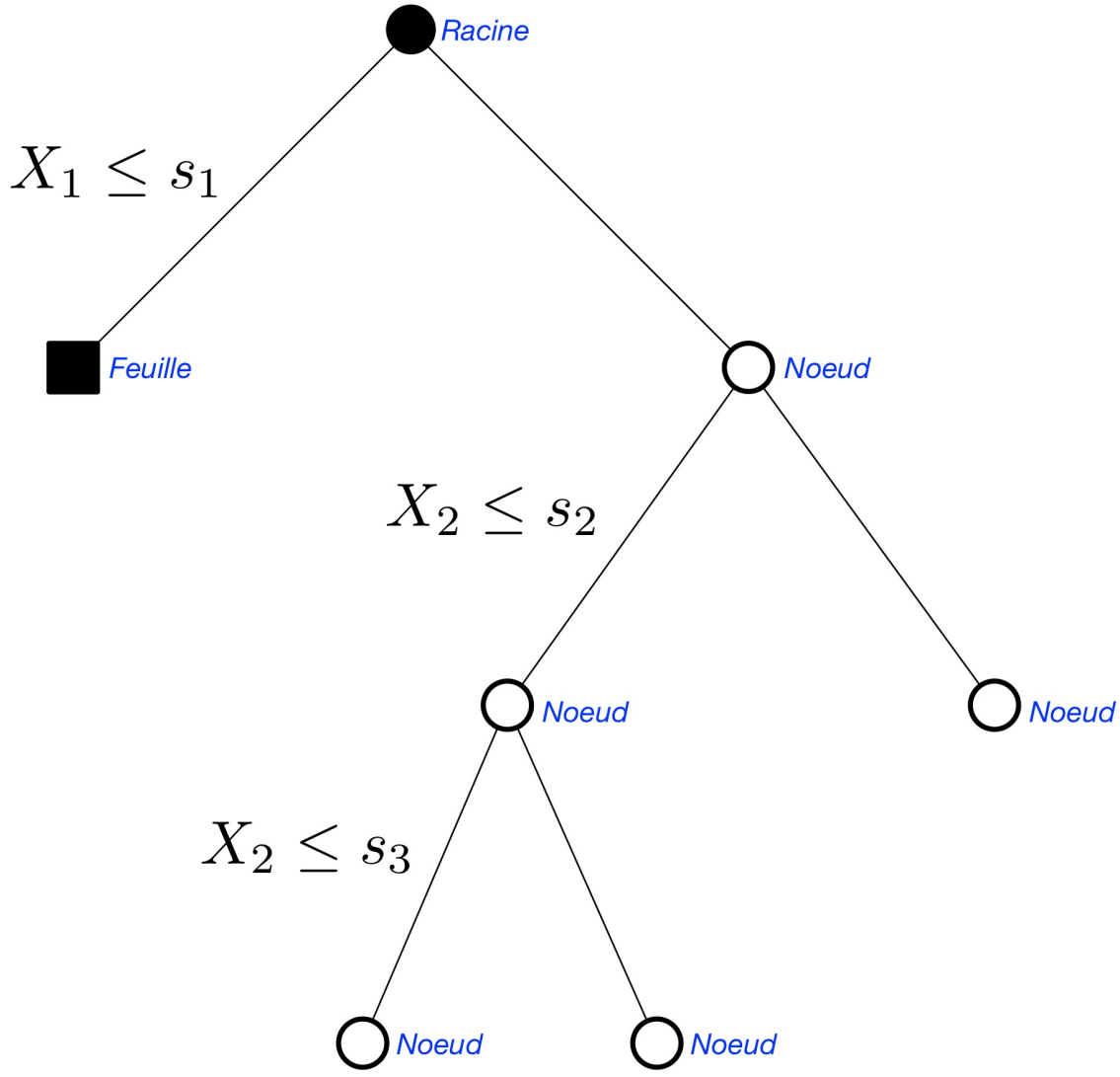
Exemple (Classification) IX



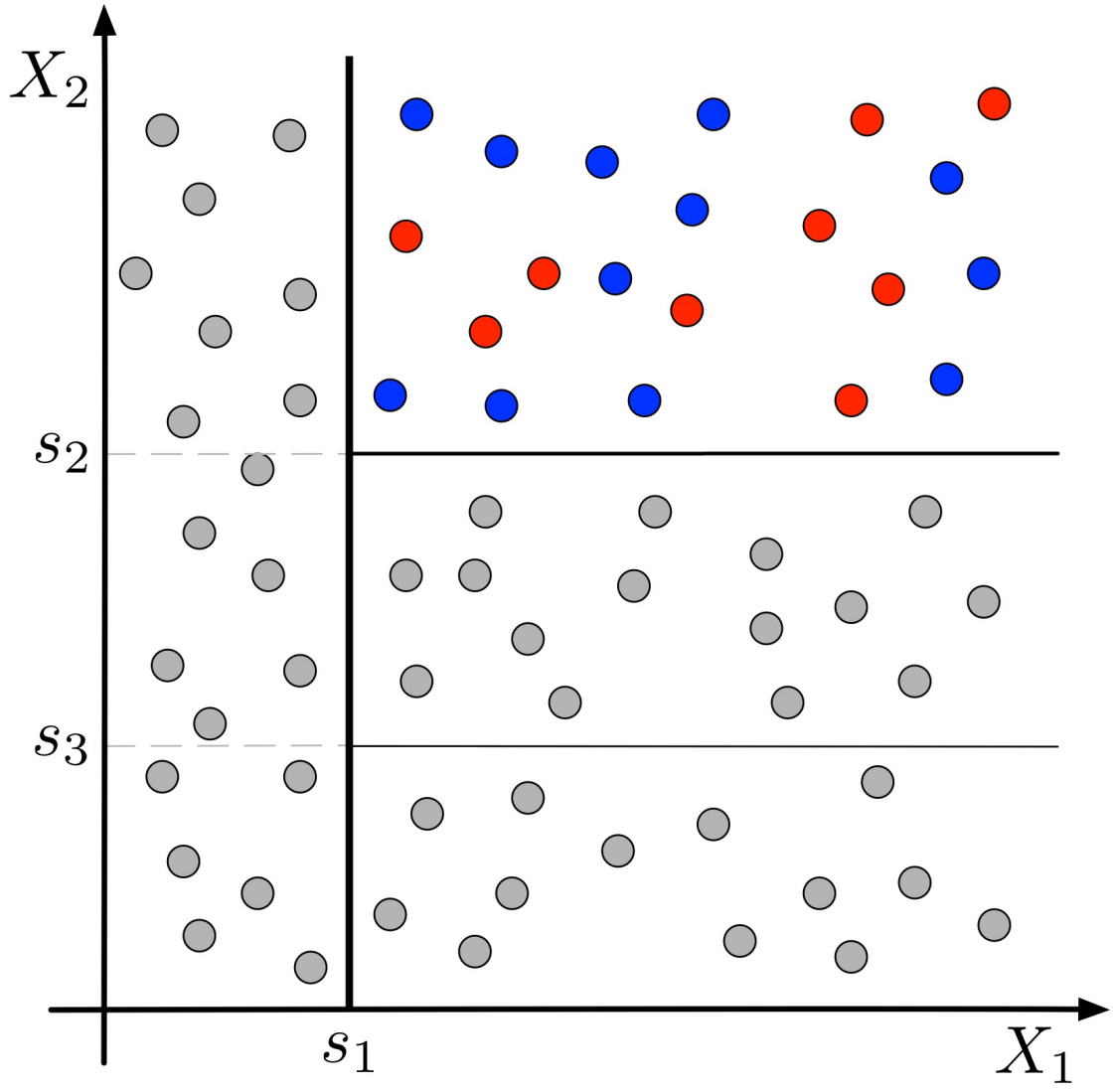
Exemple (Classification) X



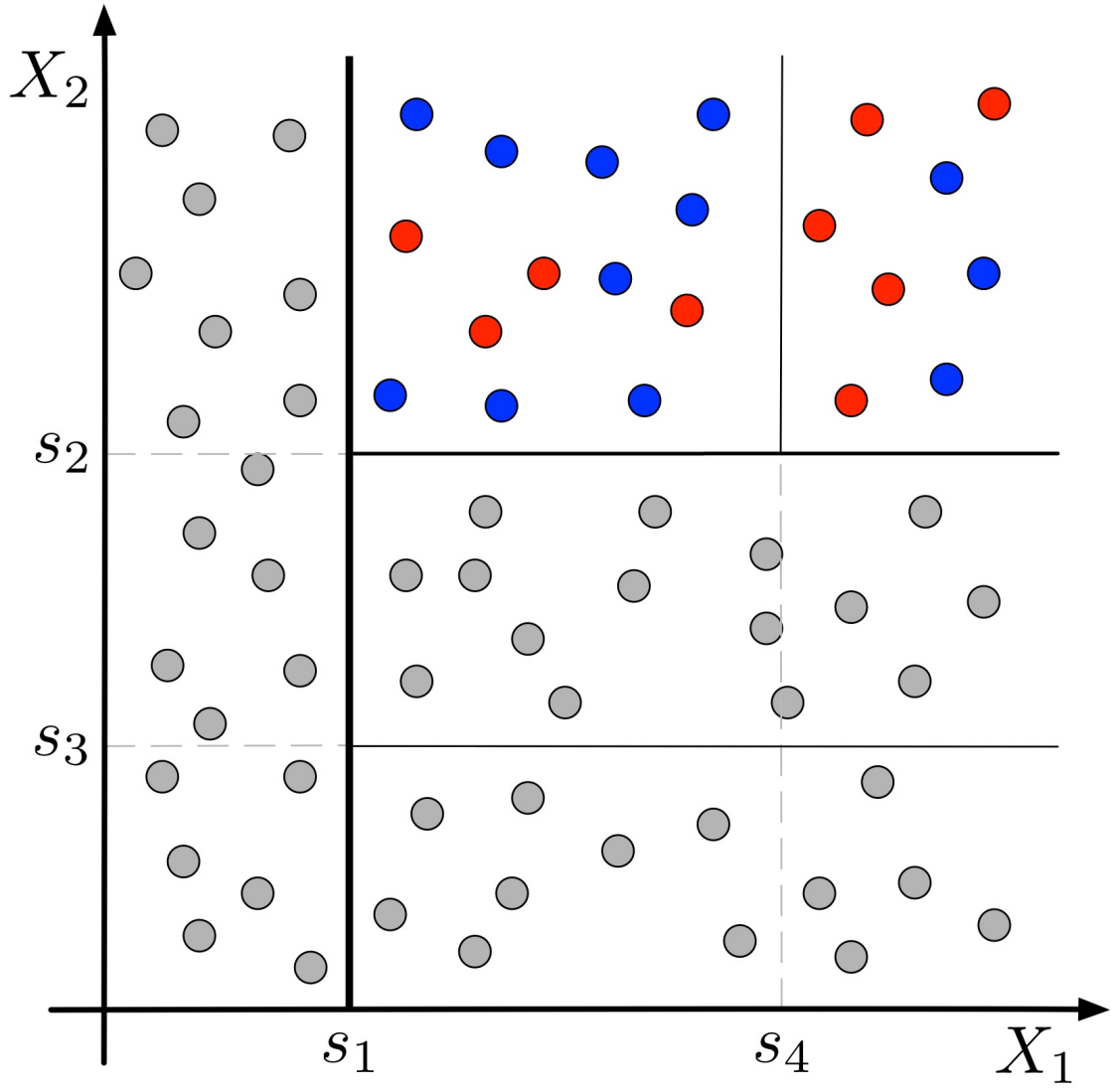
Exemple (Classification) XI



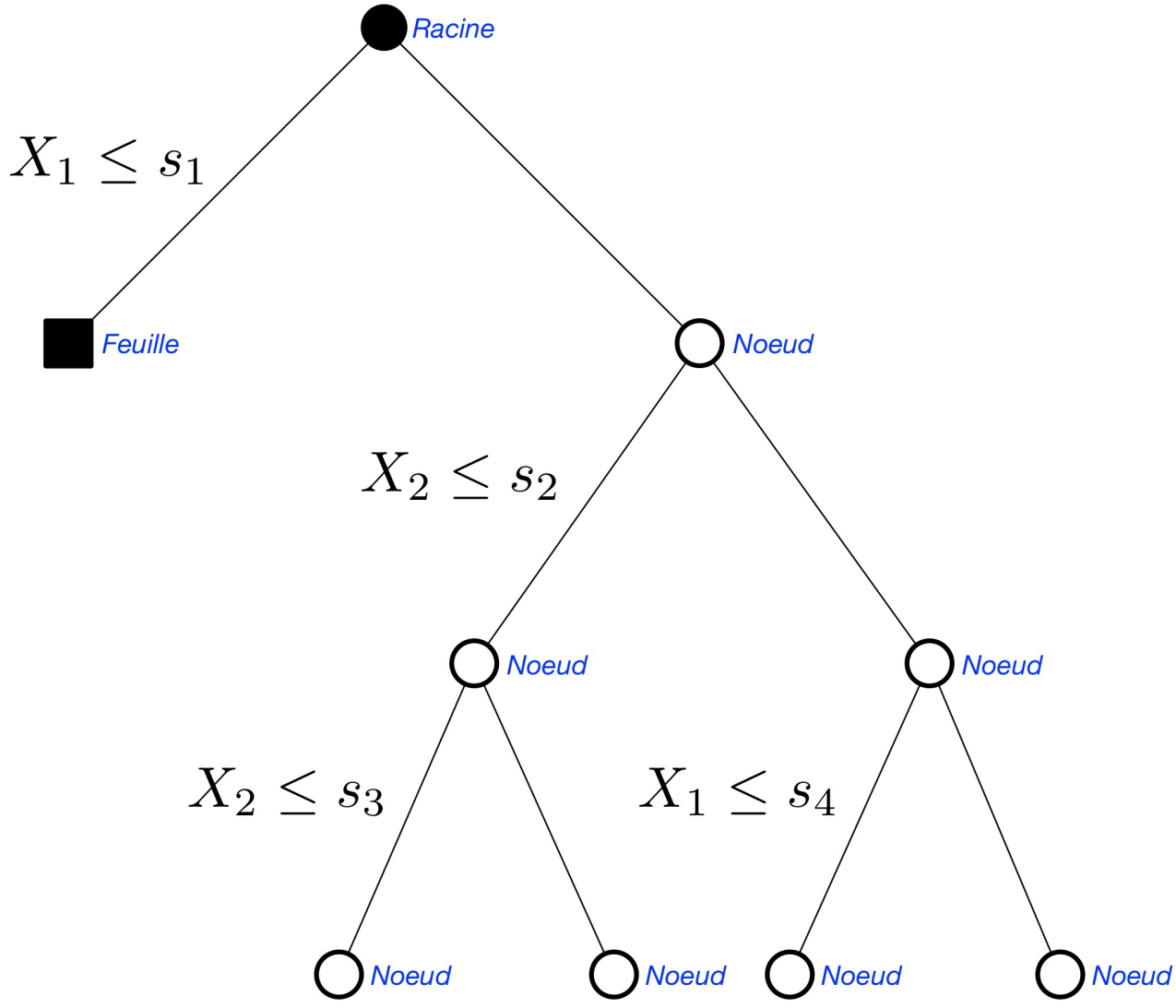
Exemple (Classification) XII



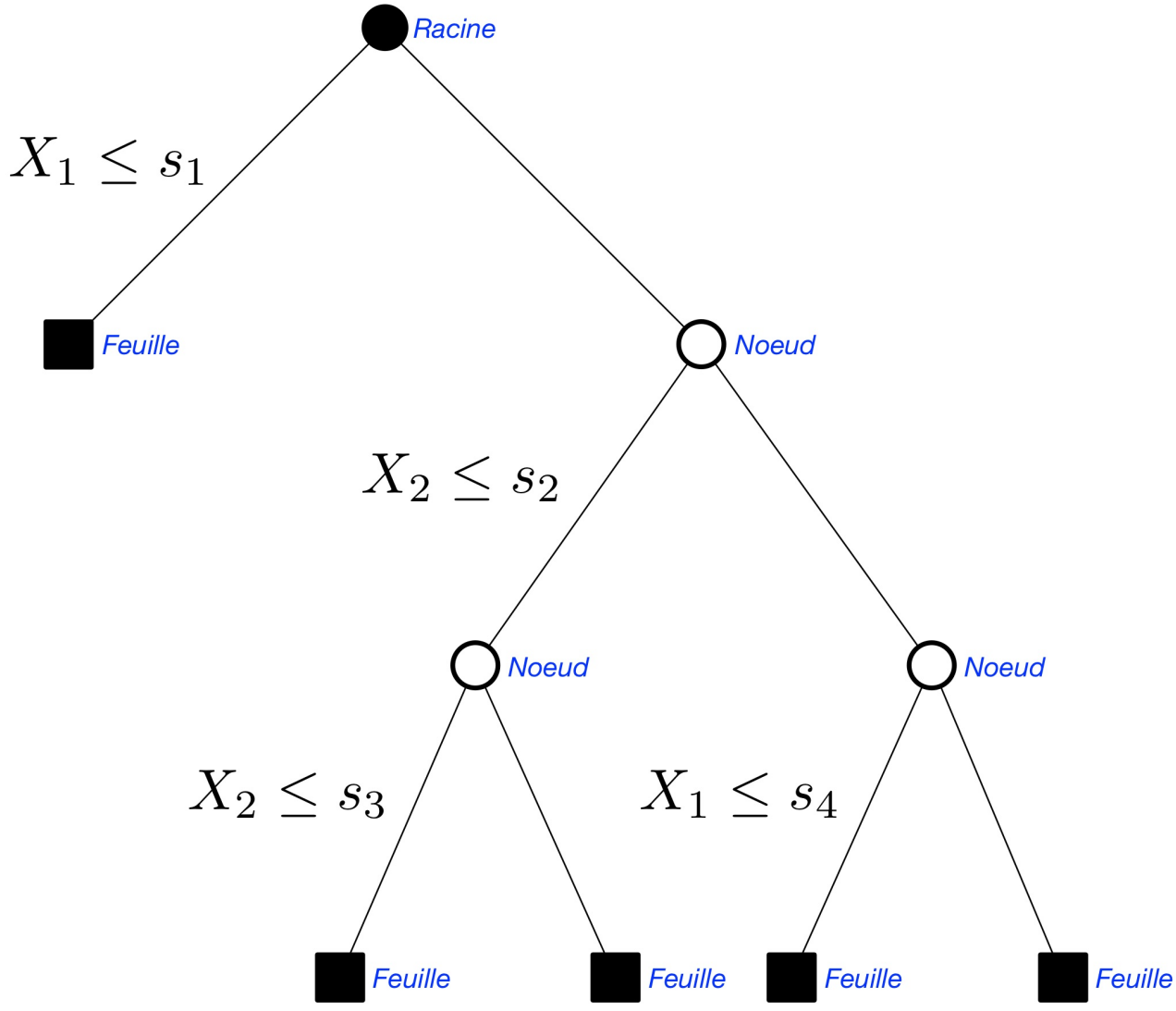
Exemple (Classification) XIII



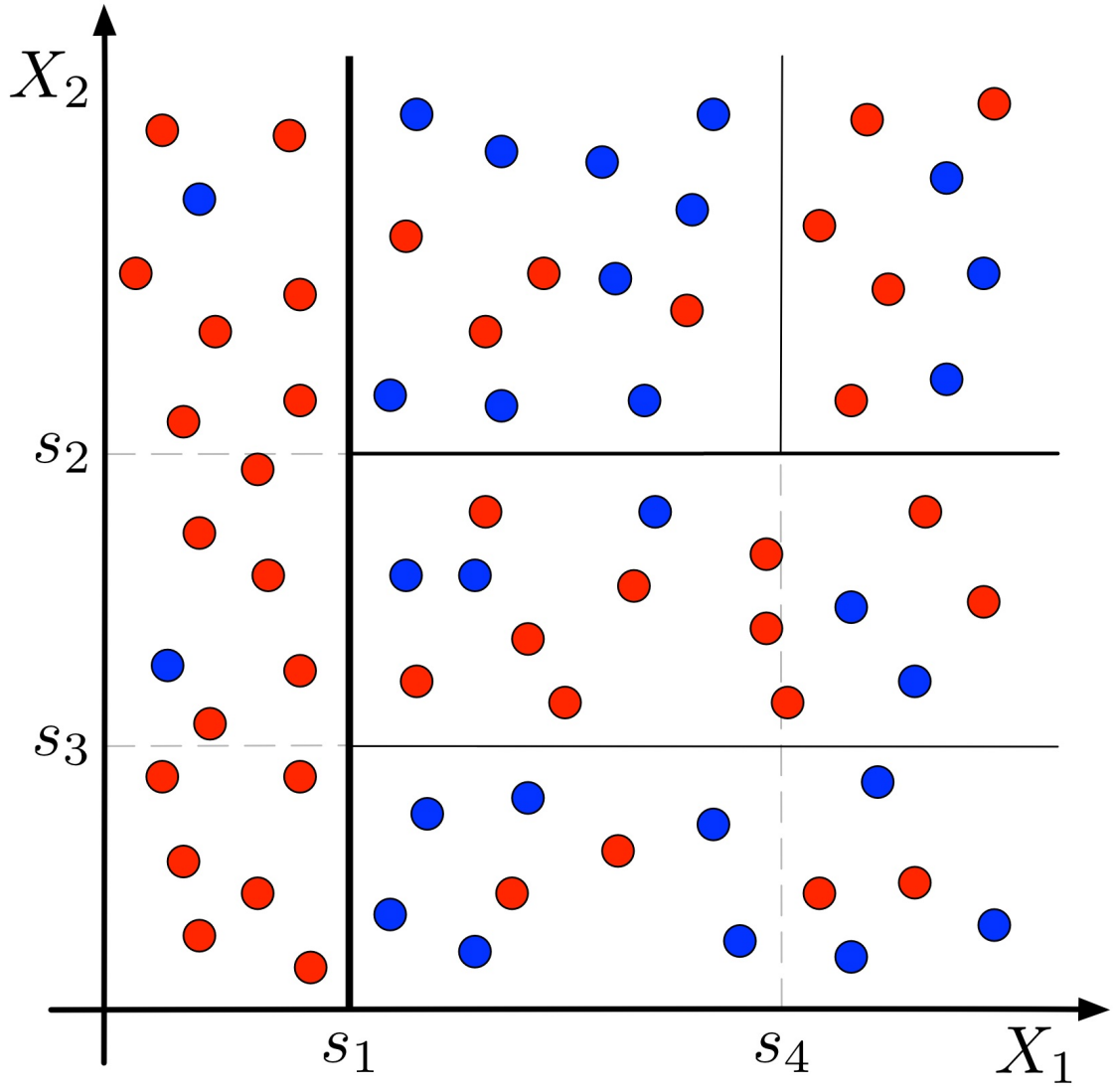
Exemple (Classification) XIV



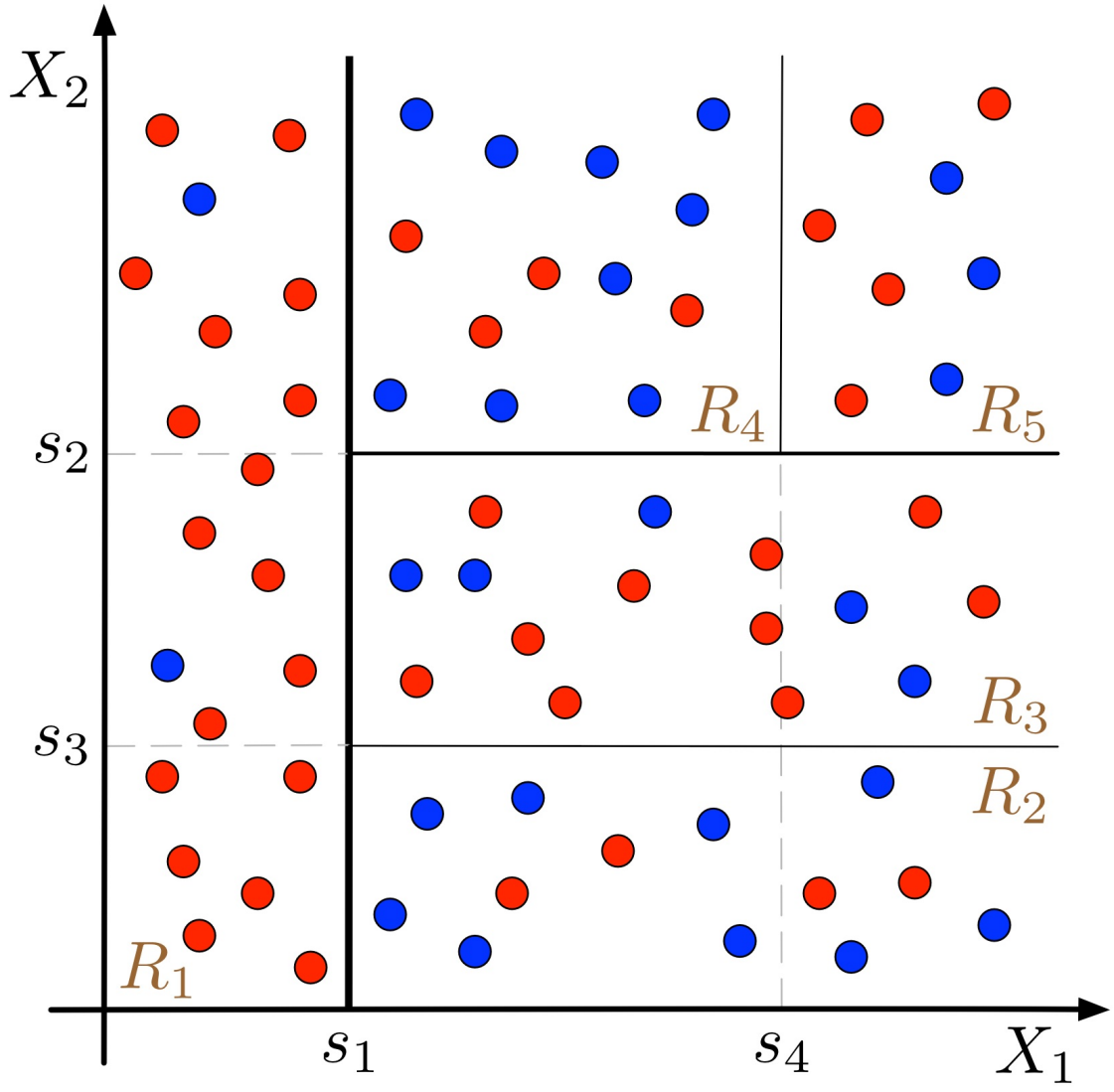
Exemple (Classification) XV



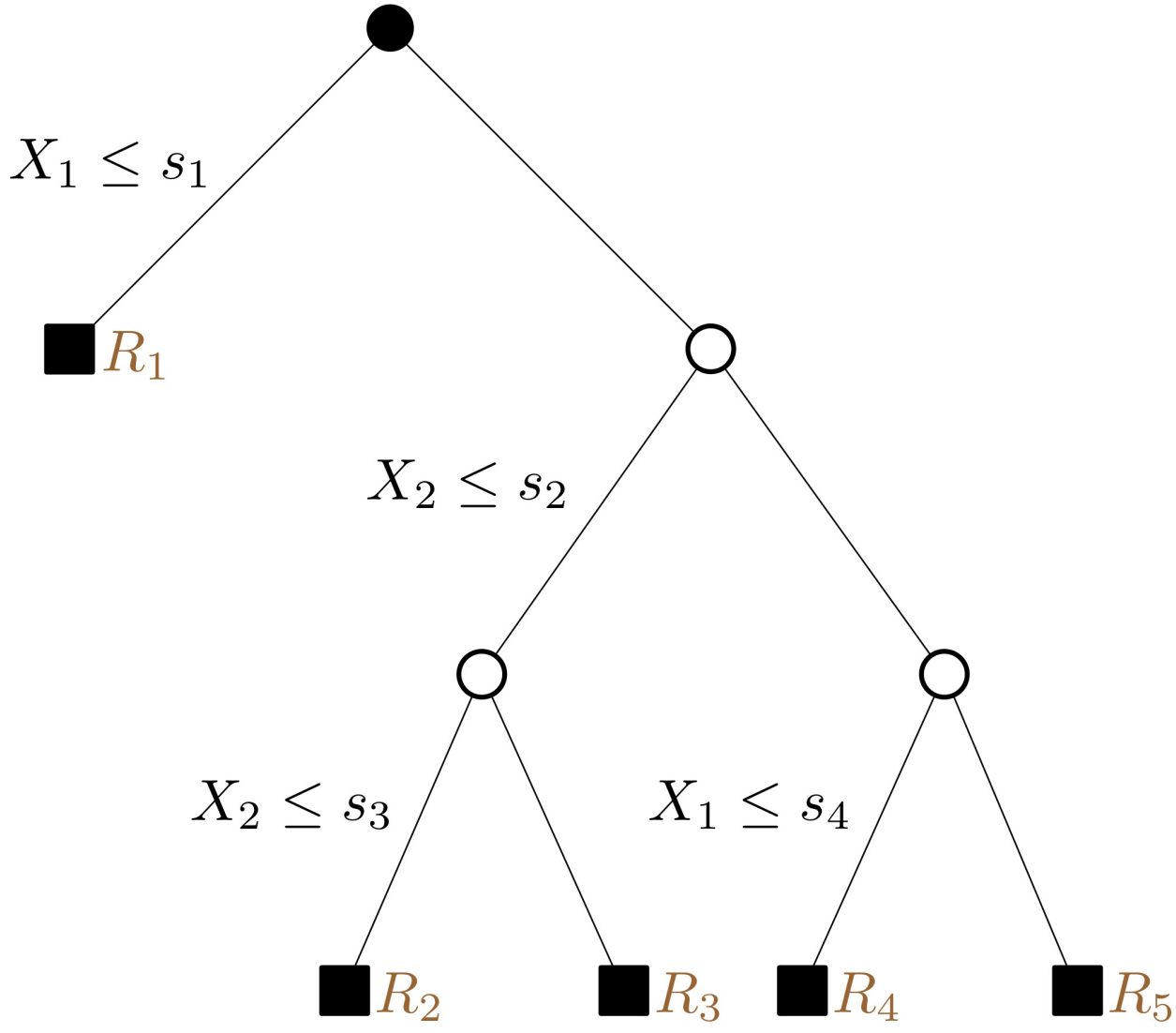
Exemple (Classification) XVI



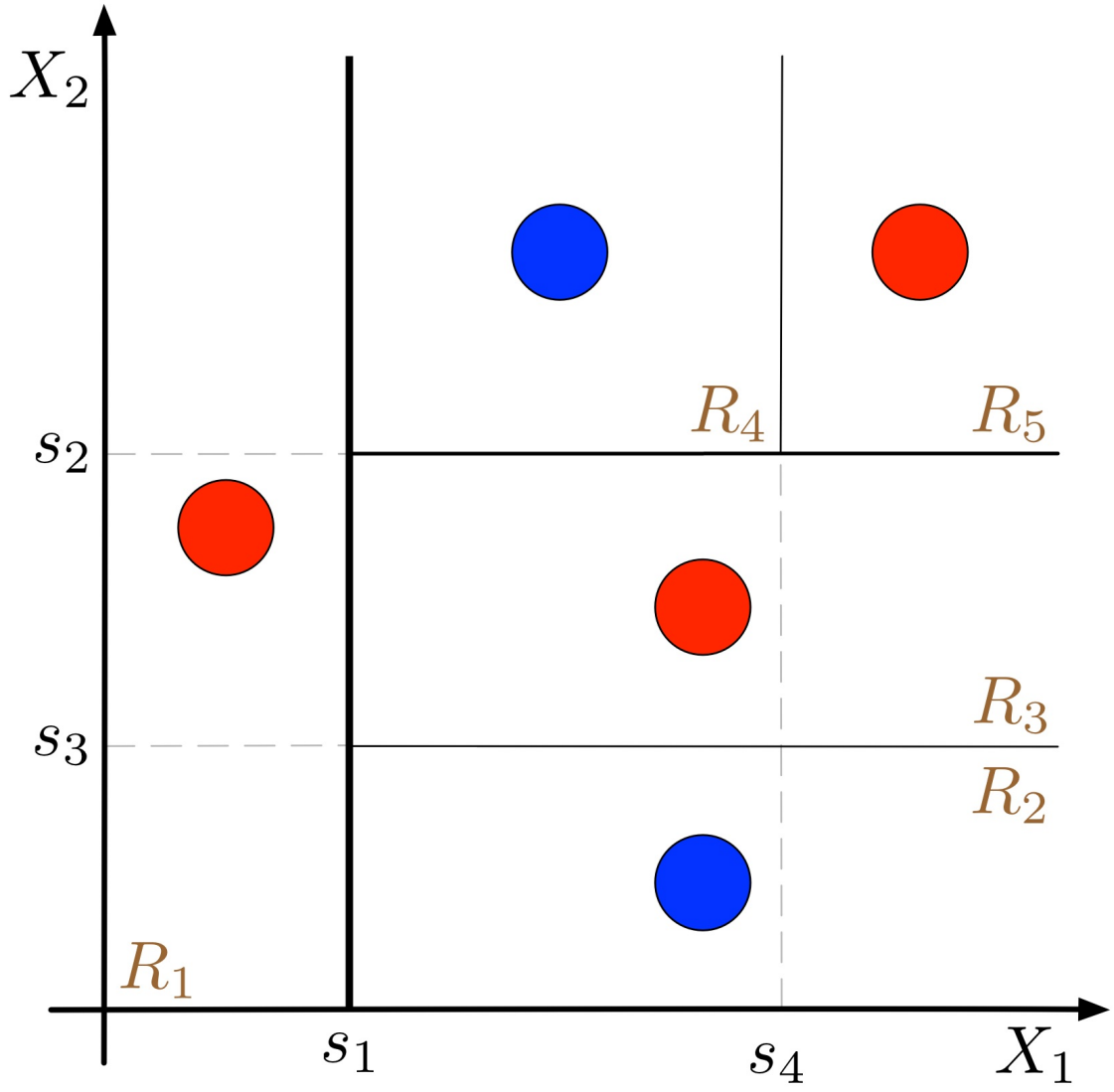
Exemple (Classification) XVII



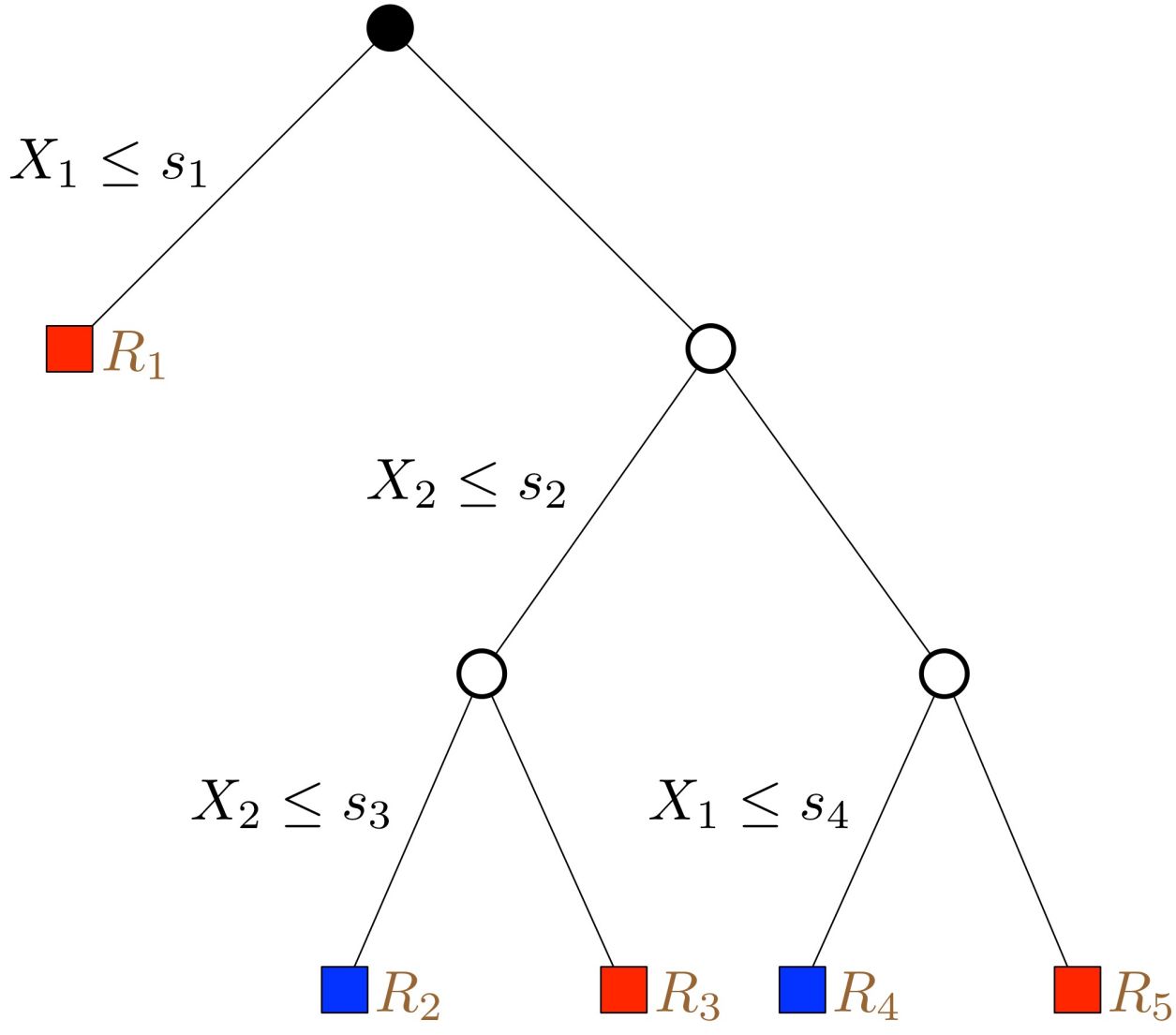
Exemple (Classification) XVIII



Exemple (Classification) XIX



Exemple (Classification) XX



Plan

Le nombre de divisions possibles

- Pour une variable quantitative : $\ell - 1$ où $\ell \leq n$ est le nombre de valeurs distinctes prises par la variable.
- Pour des variables ordinales : $\ell - 1$ où ℓ est le nombre de modalités.
- Pour des variables nominales : $2^{\ell} - 1$ où ℓ est le nombre de modalités.

Cas des covariables qualitatives

- L'algorithme tend à favoriser les variables avec beaucoup de modalités.
- Il est recommandé de réduire le nombre de modalités en fusionnant certaines catégories.

Enjeux techniques

- Choisir la meilleure division pour chaque variable.
- Déterminer la meilleure variable séparatrice.
- Décider qu'un noeud est une feuille.
- Estimer un modèle de prévision dans chaque pavé.

Sur-apprentissage

- Un arbre trop grand sur-apprend les données.
- Un arbre trop petit risque de ne pas apprendre suffisamment.

Données considérées

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\dots,n\}} \ .$$

  On note :

$$d_n = (x_i, y_i)_{i \in \{1,\dots,n\}} \ .$$

- On considère dans la suite que :
  - $X = (X_1, \dots, X_p)^\top$:
    Les p covariables peuvent être quantitatives ou qualitatives.
  - $Y \in \{1, \dots, K\}$ dans le cas de la classification supervisée.
  - $Y \in \mathbb{R}$ dans le cas de la régression.

Plan

Objectif

Classifier une variable qualitative Y à K classes :

$$\{1, \ldots, K\} \ ,$$

à partir de p covariables :

$$(X_1, \ldots, X_p) \ .$$

On considère des covariables quantitatives dans la suite.

Prévision de la feuille

- Soit une partition en M pavés $\{R_1, \ldots, R_M\}$.
- Soit C_m la classe du m-ème pavé.
- Pour $k \in \{1, \ldots, K\}$, on estime $\mathbb{P}(C_m = k)$ par :

$$\widehat{p}_k^m = \frac{1}{\text{Card}\{x_i \in R_m\}} \sum_{x_i \in R_m} \mathbb{1}_{y_i = k} \ .$$

- La classe prévue pour le m-ème pavé est la classe la plus présente dans ce pavé :

$$\widehat{C}_m = \arg \max_{k \in \{1, \ldots, K\}} \widehat{p}_k^m \ .$$

Critère de division

- Afin de déterminer la meilleure partition binaire, on utilise un algorithme « greedy » (gourmand).
- Considérons la partition binaire pour la j-ème covariable et le point seuil s :

$$R_1(j,s) = \{X / X_j \leq s\} ,$$
$$R_2(j,s) = \{X / X_j > s\} .$$

- On choisit la variable j et le seuil s qui minimisent (par exemple) l'erreur de classification sur les 2 pavés ainsi constitués :

$$\min_{j \in \{1,\ldots,p\}} \min_{s} \left[\left(1 - \widehat{p}^1_{\widehat{C}_1}\right) + \left(1 - \widehat{p}^2_{\widehat{C}_2}\right) \right] .$$

- Une fois cette division optimale déterminée, on réitère l'étape de division sur les deux pavés obtenus, et ainsi de suite.

Fonction d'hétérogénéité

- On a utilisé ici une fonction d'hétérogénéité (« impurity ») i particulière : l'erreur de classification.
- Plus généralement, une fonction d'hétérogénéité doit vérifier :
  - i est minimale, et égale à 0, pour les configurations avec une seule classe :

$$\begin{gathered}(1,0,0,\ldots,0)\\(0,1,0,\ldots,0)\\\vdots\\(1,0,0,\ldots,0)\ .\end{gathered}$$

  - i est maximale pour la configuration equi-répartie :

$$\forall i \in \{1,\ldots,K\} : p_i = \frac{1}{K}\ .$$

Exemples de fonctions d'hétérogénéité

Pour un pavé R_m :

- Erreur de classification :

$$i\left(R_m\right) = 1 - \widehat{p}^m_{\widehat{C}_m} \ .$$

- Indice de Gini :

$$i\left(R_m\right) = \sum_{k=1}^{K} \widehat{p}^m_k \left(1 - \widehat{p}^m_k\right) \ .$$

- Entropie de Shannon :

$$i\left(R_m\right) = -\sum_{k=1}^{K} \widehat{p}^m_k \ \ln\left(\widehat{p}^m_k\right) \ .$$

Retour sur l'algorithme de division I

- Afin de déterminer la meilleure partition binaire, on utilise un algorithme « greedy » (gourmand).
- Considérons la partition binaire pour la j-ème covariable et le point seuil s :

$$R_1(j,s) = \{X / X_j \leq s\} ,$$
$$R_2(j,s) = \{X / X_j > s\} ,$$

- On choisit la variable j et le seuil s qui minimisent (par exemple) l'erreur de classification sur les 2 pavés ainsi constitués :

$$\min_{j \in \{1,\ldots,p\}} \min_{s} \left[\left(1 - \widehat{p}^1_{\widehat{C}_1}\right) + \left(1 - \widehat{p}^2_{\widehat{C}_2}\right) \right] .$$

- Une fois cette division optimale déterminée, on réitère l'étape de division sur les deux pavés obtenus, et ainsi de suite.

Retour sur l'algorithme de division II

- Afin de déterminer la meilleure partition binaire, on utilise un algorithme « greedy » (gourmand).
- Considérons la partition binaire pour la j-ème covariable et le point seuil s :

$$R_1(j,s) = \{X / X_j \leq s\} ,$$
$$R_2(j,s) = \{X / X_j > s\} ,$$

- On choisit la variable j et le seuil s qui minimisent la fonction d'hétérogénéité sur les 2 pavés ainsi constitués :

$$\min_{j \in \{1,\ldots,p\}} \min_{s} \left[i\left(R_1(j,s)\right) + i\left(R_2(j,s)\right) \right] .$$

- Une fois cette division optimale déterminée, on réitère l'étape de division sur les deux pavés obtenus, et ainsi de suite.

Plan

Prévision de la feuille

- Soit une partition en M pavés $\{R_1, \ldots, R_M\}$.
- Soit c_m la valeur (constante) du m-ème pavé.
- On considère ici la fonction de régression suivante :

$$m(x) = \sum_{m=1}^{M} c_m \mathbb{1}_{x \in R_m} \ .$$

- Avec le critère des moindres carrés, on obtient la fonction de prévision suivante :

$$\widehat{m}(x) = \sum_{m=1}^{M} \widehat{c}_m \mathbb{1}_{x \in R_m} \ .$$

où :

$$\widehat{c}_m = \frac{1}{\text{Card}\,\{x_i \in R_m\}} \sum_{x_i \in R_m} y_i \ .$$

Critère de division

- Afin de déterminer la meilleure partition binaire, on utilise un algorithme « greedy » (gourmand).
- Considérons la partition binaire pour la j-ème covariable et le point seuil s :

$$R_1(j,s) = \{X / X_j \leq s\} ,$$
$$R_2(j,s) = \{X / X_j > s\} ,$$

- On choisit la variable j et le seuil s qui minimisent (par exemple) l'erreur suivante sur les 2 pavés ainsi constitués :

$$\min_{j \in \{1,\ldots,p\}} \min_{s} \left(\sum_{x_i \in R_1(j,s)} (y_i - \widehat{c}_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \widehat{c}_2)^2 \right) .$$

- Une fois cette division optimale déterminée, on réitère l'étape de division sur les deux pavés obtenus, et ainsi de suite.

Fonction d'hétérogénité

On a utilisé ici la fonction d'hétérogénité :

$$i\left(R_m\right) = \sum_{x_i \in R_m(j,s)} \left(y_i - \widehat{c}_m\right)^2 .$$

Plan

Critères d'arrêt

- Diminution de l'hétérogénéité inférieure à un seuil.
- Nombre de points inférieur à un seuil (e.g 5).
- Tests d'hypothèses.
- Croissance complète puis élagage.

Elagage

1. Construire l'arbre maximal à l'aide d'une procédure forward. A chaque étape, trouver la meilleure division et s'arrêter lorsque toutes les feuilles contiennent moins d'une nombre fixé de points (communément entre 1 et 5) ou ont les mêmes sorties.

2. Créer une suite imbriquée d'arbres, de complexité décroissante.

3. Elaguer les branches inutiles (déterminer l'arbre optimal).

Coût de la complexité

- Pour un arbre $\mathcal{T}$, avec un nombre de feuilles $|\mathcal{T}|$, on considère :

$$C(\mathcal{T}) = \sum_{i=1}^{|\mathcal{T}|} i(R_i) + \alpha |\mathcal{T}| \ .$$

- On désigne par $\mathcal{T}_{max}$ l'arbre maximal.
- Afin de déterminer l'arbre optimal $\mathcal{T}_{opt} \subset \mathcal{T}_{max}$ qui minimise C, on élague récursivement la feuille la plus faible (au sens de l'hétérogénéité i).
- Dans CART, l'hyper-paramètre α est choisi par validation croisée.

Avantages et inconvénients

- Avantages :
  - Aucune hypothèse sur la loi et sur la fonction de lien.
  - Facilité d'implémentation.
  - Représentation graphique agréable (à utiliser avec parcimonie).
- Inconvénients :
  - Nécessité de disposer de gros jeux de données.
  - Uniquement des séparations horizontales ou verticales.
  - Pas d'interactions entre les variables.
  - Instabilité : un échantillon légèrement différent peut produire un arbre différent.

Le coin R

On peut utiliser plusieurs packages, parmi lesquels :

- Le package `rpart` :
  - La fonction de base est `rpart`.
  - Les fonction `plot.rpart` et `text.rpart` permettent de visualiser l'arbre.
  - Pour élaguer l'arbre, on peut s'appuyer sur les fonctions `plotcp` et `printcp`, et utiliser l'option `cp` de la commande `rpart` (« cp » pour *complexity parameter*).
- Le package `caret` :
  - La fonction de base est `train`.
  - On utilise des options correspondant à CART, par exemple : `method="rpart"`.

Références

Breiman, L., J. Friedman, C. J. Stone et R. Olshen. 1984, *Classification and regression trees*, Taylor & Francis.