

## Arbres de décision

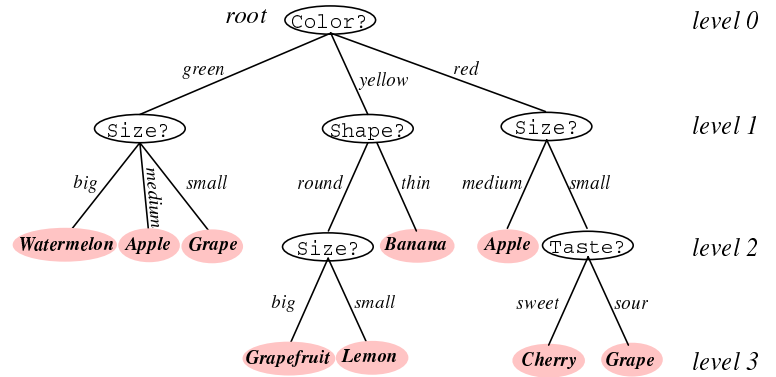
1

### • Objectif

- classification en utilisant une **séquence de questions** fermées
- les questions sont organisées dans un **arbre**

## Arbres de décision

2



## Arbres de décision

3

### • Avantages

- fonctionnent avec des données **non-métriques**
- invariabilité par **translation**, par **changement d'échelle**, par **transformation monotone des coordonnées**
- **interprétabilité**
- entraînement **efficace**
- classification **très efficace**

### • Désavantages

- **instabilité**

## Arbres de décision

4

### • Algorithmes de CART (arbres de classification et régression)

- **combien de découpages** par noeud?
- **quel attribut** faut-il tester à un noeud?
- **quand arrêter** de découper?
- si l'arbre est trop grand, **comment élaguer**?
- si une feuille est non-pure, **comment choisir la catégorie**?

## Arbres de décision

5

### • Nombre de découpages

- tous les arbres de décision peuvent être représentés par un **arbre de décision binaire**

### • Affectation de catégorie

- vote par **majorité**

## Arbres de décision

6

### • Sélection de test

- objectif: un **arbre simple** (rasoir d'Occam)
- choisir le découpage qui **augmente le plus la pureté**

• L'impureté du noeud

- fréquence de classe:  $\mu_j = \frac{\#\{\text{classe} = C_j\}}{n}$
- impureté d'entropie:  $i(N) = - \sum_j \mu_j \lg \mu_j$
- impureté de variance (deux catégories):  
 $i(N) = \mu_1 \mu_2$
- impureté de Gini:  $i(N) = \sum_{i \neq j} \mu_i \mu_j = 1 - \sum_j \mu_j^2$
- impureté de mauvaise classification:  
 $i(N) = 1 - \max_j \mu_j$

• Sélection de test

- chute d'impureté:  
 $\Delta i(N) = i(N) - \mu^{(s)} i(N^{(s)}) - (1 - \mu^{(s)}) i(N^{(d)})$
- approche gloutonne

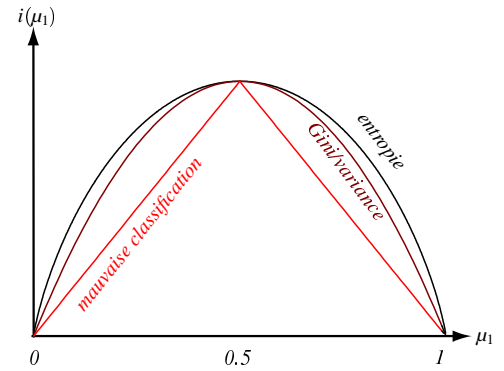
• Forme générale de la fonction

- découpage sur un attribut simple  $\rightarrow$  arbre monotétique
- découpage linéaire

• Élaguer

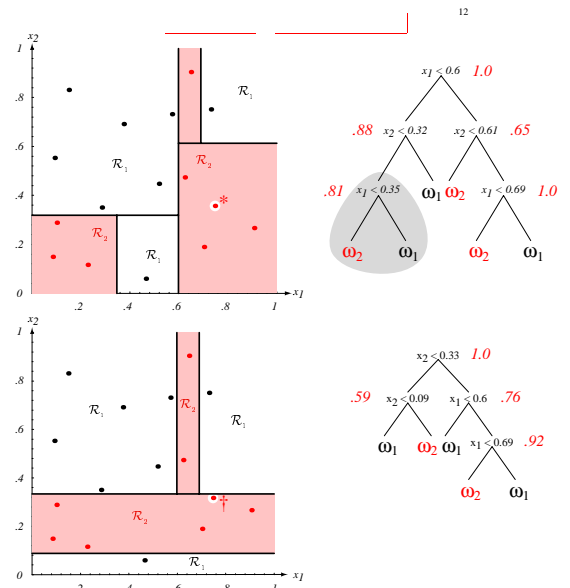
- l'effet d'horizon
- pousser l'arbre jusqu'à un point par feuille
- supprimer (unifier) les noeuds si le pureté ne diminue pas
- pas de validation croisée
- plus de calcul
- élaguer les règles pour simplifier la description

• L'impureté de noeud

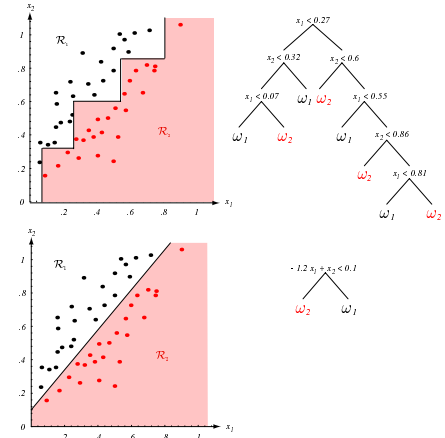


• Quand arrêter

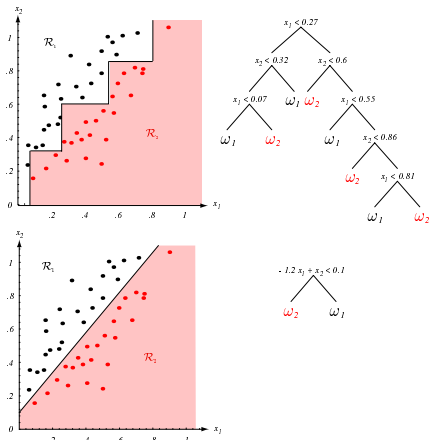
- un point par feuille: overfitting
- trop tôt: grande erreur d'entraînement
- technique générale: validation/validation croisée
- chute d'impureté < seuil
- nombre de points < seuil
- principe de MDL (minimum description length): minimiser  
 $\alpha \cdot \text{taille} + \sum_{\text{feuilles } N} i(N)$
- méthodes statistiques pour mesurer la signification de la réduction d'impureté



- Complexité:  $O(dn \lg n)$
- Choix de traits
- Arbres multivariés



Choix de traits



Arbres multivariés

