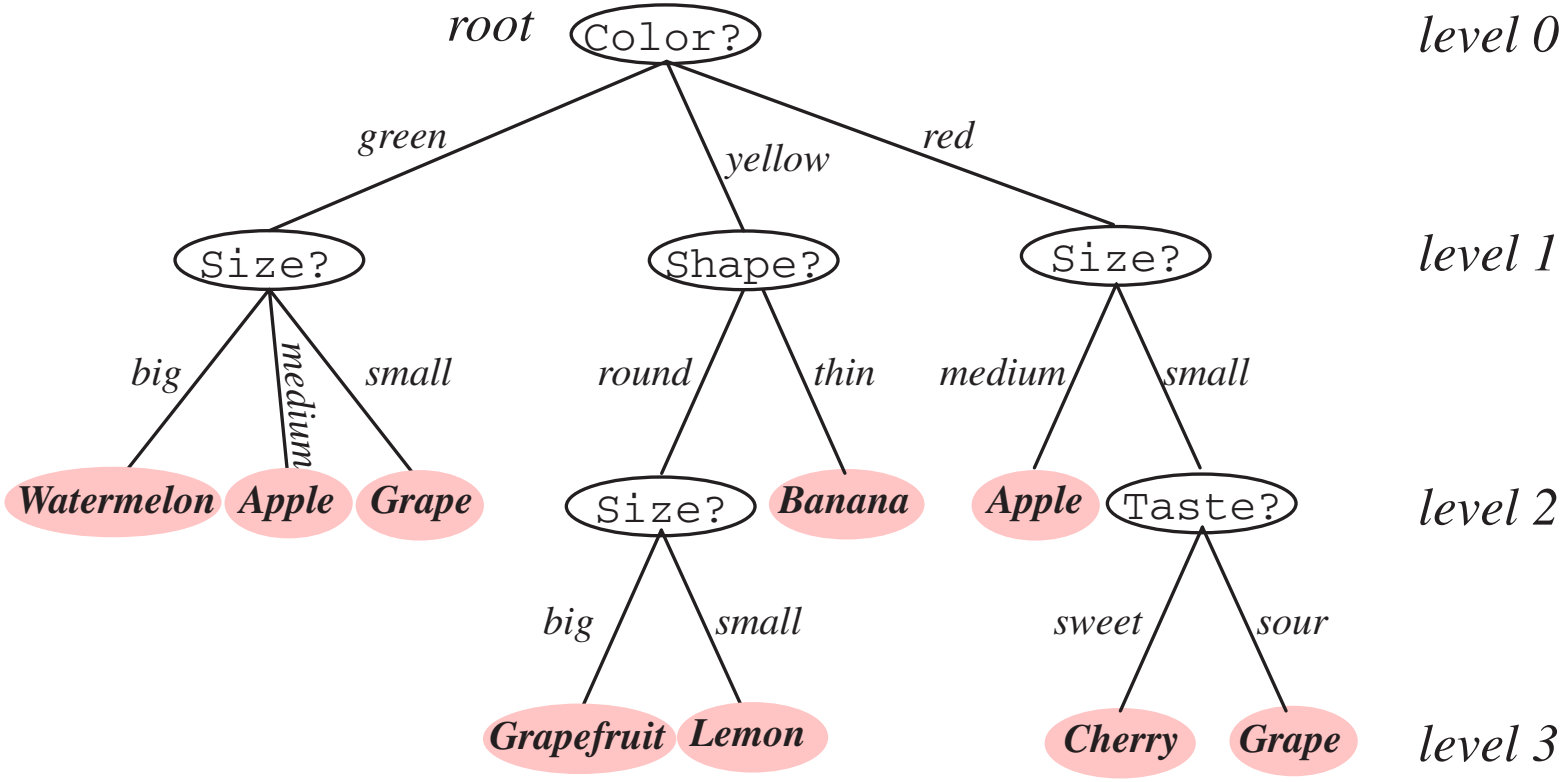
Arbres de décision

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

Arbres de décision



Arbres de décision

- Avantages
  - fonctionnent avec des données non-métriques
  - invariabilité par translation, par changement d'échelle, par transformation monotone des coordonnées
  - interpretabilité
  - entraînement efficace
  - classification très efficace
- Désavantages
  - instabilité

Arbres de décision

- 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

- 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

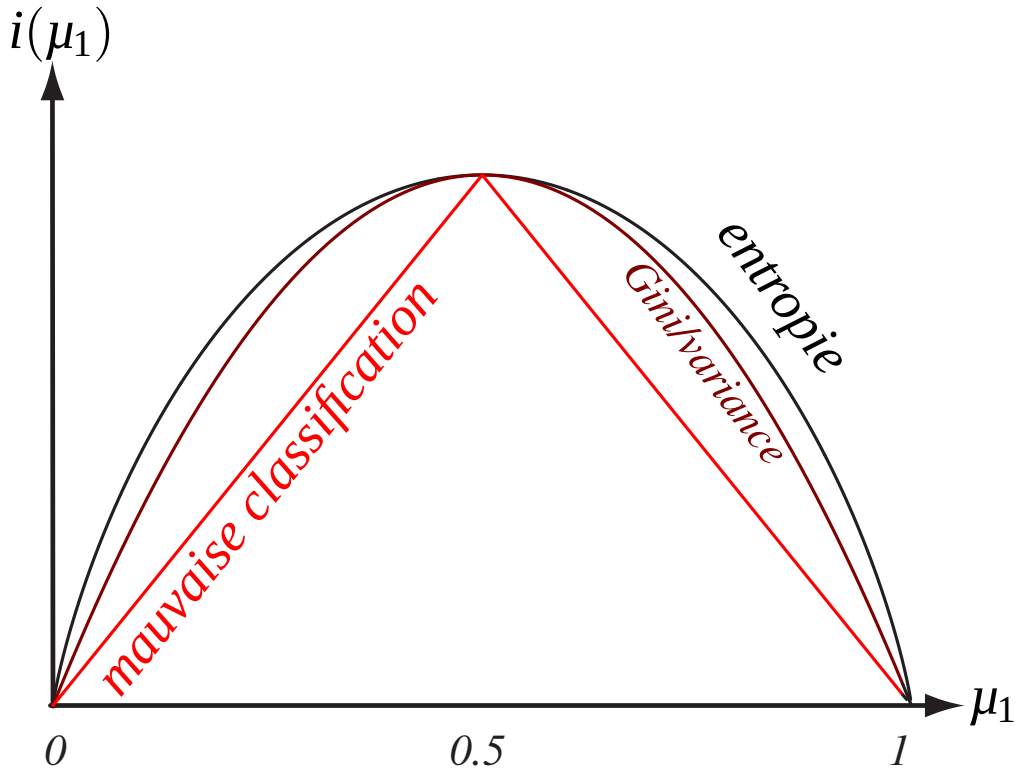
- Sélection de test
  - objectif: un arbre simple (rasoir d'Occam)
  - choisir le découpage qui augmente le plus la pureté

Arbres de décision

- 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:
    $$i(N) = 1 - \max_j \mu_j$$

Arbres de décision

- L'impureté de noeud



Arbres de décision

- Sélection de test
  - chute d'impureté:

$$\Delta i(N) = i(N) - \mu^{(g)} i(N^{(g)}) - (1 - \mu^{(g)}) i(N^{(d)})$$

  - approche gloutonne

- Forme générale de la fonction
  - découpage sur un attribut simple $\longrightarrow$ arbre monotétique
  - découpage linéaire

Arbres de décision

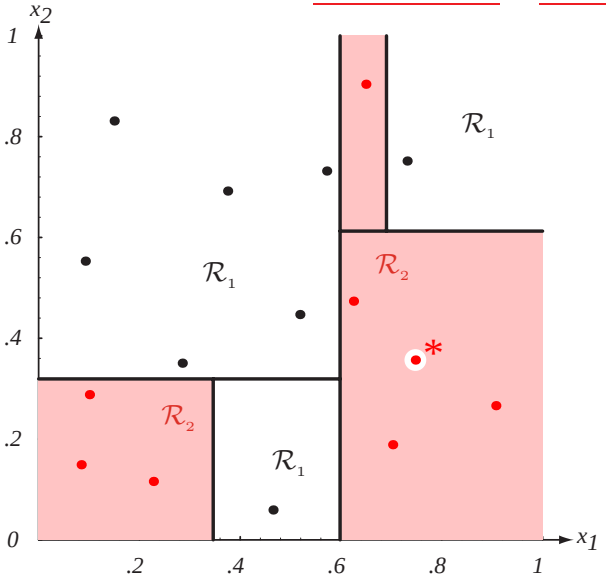
- Quand arrêter
  - un point par feuille: overfitting
  - trop tôt: grande erreur d'entraînement
  - technique générale: validation/validation croisée
  - chute d'impureté $<$ seuil
  - nombre de points $<$ seuil
  - principe de MDL (minimum description length): minimiser

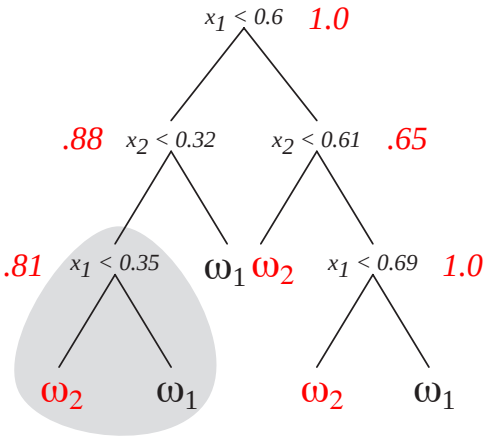
$$\alpha \cdot taille + \sum_{feuilles\ N} i(N)$$

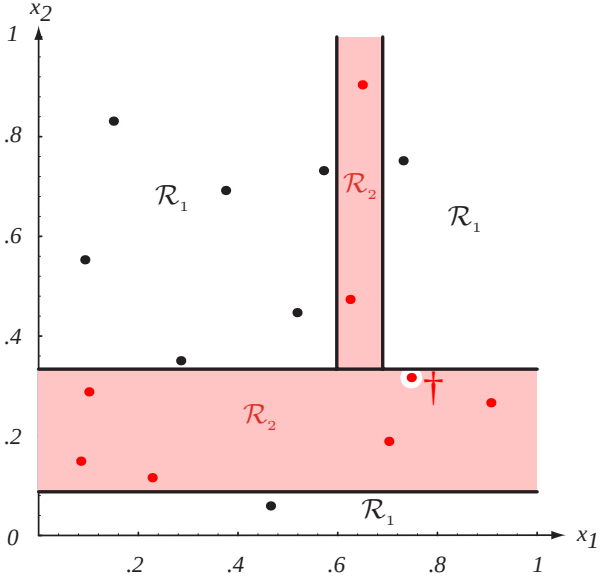
  - méthodes statistiques pour mesurer la signification de la réduction d'impureté

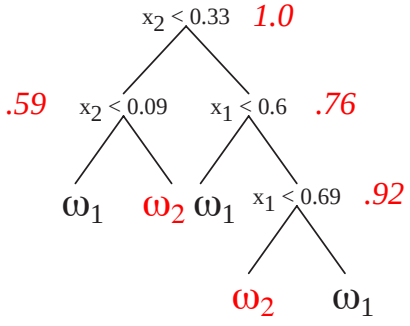
Arbres de décision

- 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


x_2
1
.8
.6
.4
.2
0
.2
.4
.6
.8
1
x_1
$\mathcal{R}_1$
$\mathcal{R}_2$
$\mathcal{R}_1$
$\mathcal{R}_2$
$\mathcal{R}_1$
*

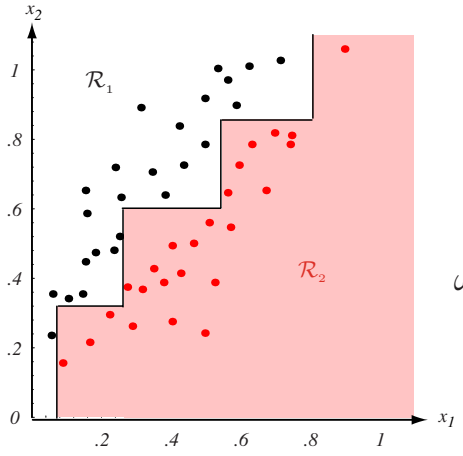

$x_1 < 0.6$ 1.0
.88
$x_2 < 0.32$
$x_2 < 0.61$
.65
.81
$x_1 < 0.35$
ω_1
ω_2
$x_1 < 0.69$
1.0
ω_2
ω_1
ω_2
ω_1

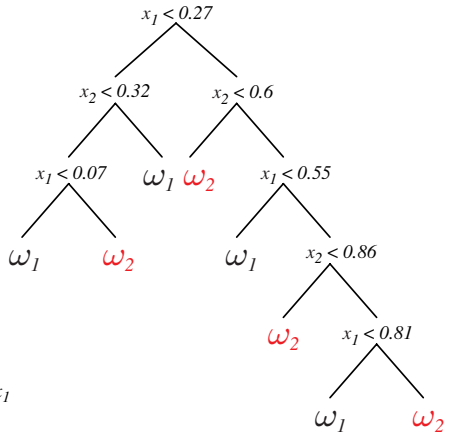

x_2
1
.8
.6
.4
.2
0
.2
.4
.6
.8
1
x_1
$\mathcal{R}_1$
$\mathcal{R}_2$
$\mathcal{R}_1$
$\mathcal{R}_2$
$\mathcal{R}_1$
†

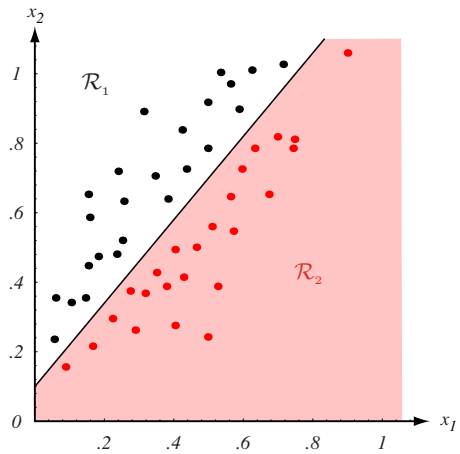

$x_2 < 0.33$ 1.0
.59
$x_2 < 0.09$
$x_1 < 0.6$
.76
ω_1
ω_2
ω_1
$x_1 < 0.69$
.92
ω_2
ω_1

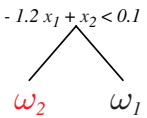
Arbres de décision

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

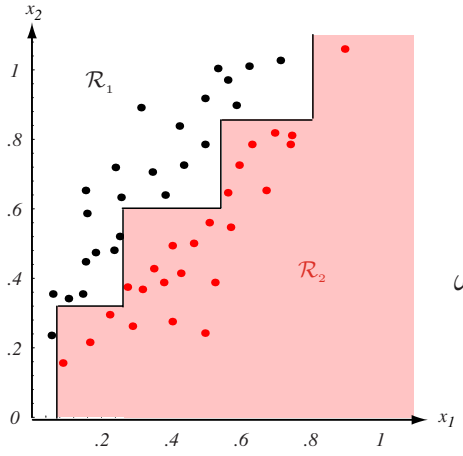

x_2
1
.8
.6
.4
.2
0
$\mathcal{R}_1$
$\mathcal{R}_2$
.2
.4
.6
.8
1
x_1

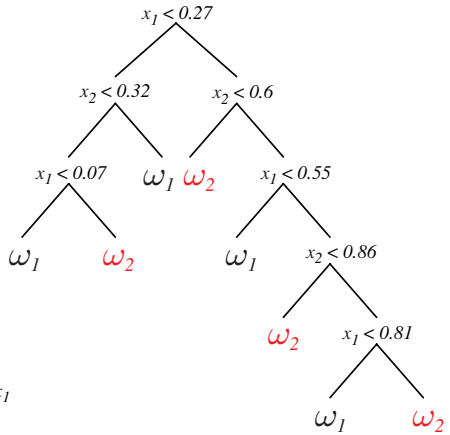

$x_1 < 0.27$
$x_2 < 0.32$
$x_2 < 0.6$
$x_1 < 0.07$
ω_1
ω_2
$x_1 < 0.55$
ω_1
ω_2
ω_1
$x_2 < 0.86$
ω_2
$x_1 < 0.81$
ω_1
ω_2

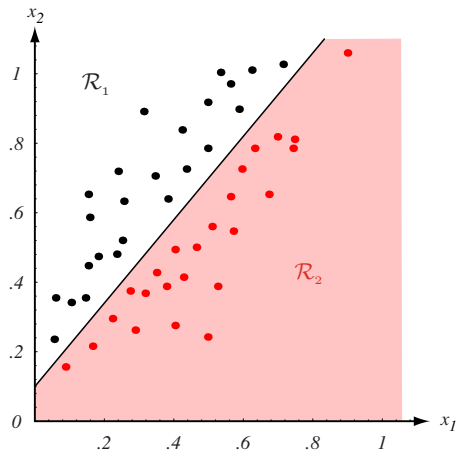

x_2
1
.8
.6
.4
.2
0
$\mathcal{R}_1$
$\mathcal{R}_2$
.2
.4
.6
.8
1
x_1

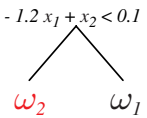

$-1.2\,x_1 + x_2 < 0.1$
ω_2
ω_1

Choix de traits









Arbres multivariés