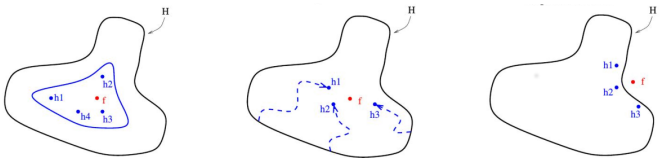
Arbres de Décision et Méthodes d'Ensemble

1. Introduction à l'Apprentissage Supervisé

L'objectif est de trouver une approximation $h(x)$ de la fonction inconnue $f(x)$ pour prédire y à partir des données x.

Problèmes

- **Statistique** : Ensemble d'entraînement insuffisant.
- **Computationnel** : Recherche locale sous-optimale.
- **Représentationnel** : La vraie fonction f peut être hors du domaine H de ce que l'on est capable de modéliser.



Solution : Apprentissage par Ensemble

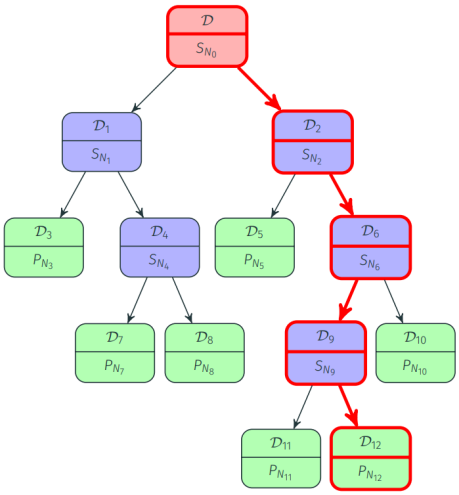
- Créer plusieurs modèles h_t.
- Les combiner avec une règle (ex. : vote majoritaire).
- Maximiser simultanément la **précision** et la **diversité** des modèles.

2. Arbres de Décision

Structure

Un arbre se compose de :

- **Nœud racine** : contient toutes les données d'entraînement D.
- **Nœuds internes** : tests sur des caractéristiques.
- **Feuilles** : prédisent les classes ou les probabilités d'appartenances au classes.



Hugo Tondenier

3. Construction d'un Arbre

Étapes

1. Initialiser le nœud racine avec les données D.
2. Décider de scinder ou non un nœud.
   - Pour décider de scinder un nœud, on évalue l'impureté actuelle du nœud et on la compare à l'impureté moyenne des sous-nœuds obtenus en appliquant différentes règles de séparation (recherche exhaustive variable par variable). La scission est effectuée si elle réduit significativement l'impureté.
   - Une mesure du gain d'impureté est utilisée pour quantifier cette amélioration :
   $$\text{Gain d'impureté} = \text{Impureté actuelle} - \sum_k \left(\frac{|D_k|}{|D|} \times \text{Impureté}(D_k) \right),$$
   où D_k représente les données dans le sous-nœud k.
3. Choisir une règle de séparation / de mesure de l'impureté (Gini, entropie).
   - **Indice de Gini** : L'indice de Gini mesure la probabilité qu'un échantillon sélectionné aléatoirement soit mal classé si son étiquette est assignée aléatoirement selon la distribution des classes dans le nœud. Il est défini comme :
   $$\text{Gini}(D) = 1 - \sum_{i=1}^{C} p_i^2,$$
   où p_i est la proportion d'échantillons de la classe i dans D.
   - **Entropie** : L'entropie mesure le degré de désordre ou d'incertitude dans les classes. Elle est définie comme :
   $$\text{Entropie}(D) = -\sum_{i=1}^{C} p_i \log_2(p_i),$$
   où p_i est la proportion d'échantillons de la classe i dans D.
   - **Choix entre Gini et Entropie** :
     - L'indice de Gini est plus rapide à calculer et favorise les scissions qui maximisent une classe dominante.
     - L'entropie est plus sensible aux déséquilibres dans les probabilités des classes, ce qui peut permettre une meilleure séparation dans des cas spécifiques.
4. Attribuer une prédiction aux feuilles (classe majoritaire ou probabilités lissées).

Critères d'Arrêt

- Limite de profondeur.
- Nombre minimal d'instances dans un nœud.
- Gain d'impureté insuffisant.

4. Forêts d'Arbres de Décision

Pourquoi ?

- Les arbres de décision seuls sont instables et sur-apprennent.
- Les forêts réduisent la variance en combinant plusieurs arbres.

Principe

1. Construire plusieurs arbres indépendants (**Bagging**, **Random Forests**).
2. Combiner leurs prédictions par vote majoritaire ou moyenne pondérée.

Avantages

- Réduction de la variance et meilleure généralisation.
- Robustesse face aux variations des ensembles d'entraînement.

Hugo Tondenier