Cours - Informatique - Arbre binaire

Gustave Cortal

Principe de l'arbre

Soit une population P d'individus décrits par p variables explicatives $X^1, ..., X^p$. Chaque individu est caractérisé par une réalisation x du vecteur aléatoire $X = (X^1, ..., X^p)$, et par une variable Z à expliquer, qui est qualitative (**arbre de décision**) ou quantitative (**arbre de régression**).

Le modèle est un arbre binaire, càd une séquence de nœuds ayant 0 ou 2 fils. Chaque nœud interne (non terminal) de l'arbre est associé une variable explicative et un test sur cette variable. Chaque feuille (nœud terminal) correspond à une région de l'espace d'entrée à laquelle est associée une valeur particulière de la variable Z à expliquer.

Nous décrivons l'algorithme CART pour la construction de l'arbre. On suppose ici que les variables explicatives sont quantitatives.

Exemple introductif

Prédire si un individu est un triangle ou un carré à partir de deux variables explicatives quantitatives X^1 et X^2.

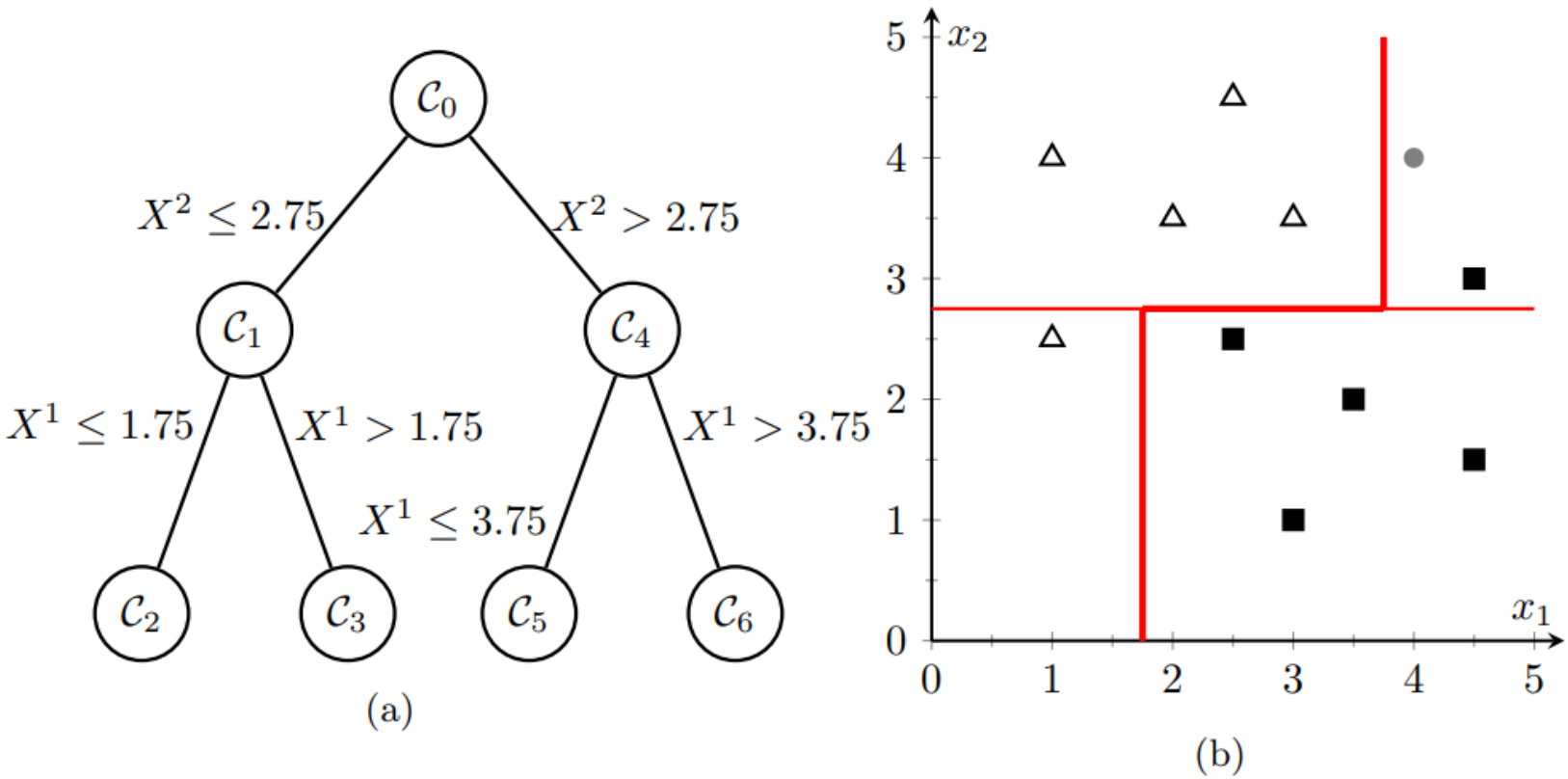


Figure: Arbre de classification (a) et partition correspondante de l'espace d'entrée et des données d'apprentissage (b)

Avantages et inconvénients

- Utilisé pour des problèmes de prédiction d'une variable quantitative (arbre de régression) ou qualitative (arbre de décision)
- Peut gérer des variables explicatives quantitatives et qualitatives
- Interpretabilité de la prédiction : représenter par une série de tests sur les variables explicatives

L'inconvénient est que les performances sont souvent inférieures aux autres méthodes d'apprentissage statistique. On peut améliorer les performances avec des méthodes ensemblistes comme la forêt aléatoire, mais on sacrifiera l'interprétabilité des prédictions.

Construction de l'arbre : phase de développement

L'objectif est de déterminer une **partition de l'espace d'entrée en régions**, chaque région correspondant à un sous-ensemble d'apprentissage *homogène* au sens de la variable Z à prédire.

Il faut trouver des régions de l'espace dans lesquelles les individus d'apprentissage sont issus de la même classe (pour l'arbre de décision) ou dans lesquelles les valeurs prises par la variable Z sont proches (pour l'arbre de régression).

Construction **incrémentale** : tant qu'une région contient des exemples d'apprentissage *hétérogènes*, elle est séparée en deux pour former deux sous-ensembles d'exemples les plus purs possibles.

$\rightarrow$ Nécessité de définir un **critère d'impureté** qui permet d'évaluer la qualité d'une séparation en mesurant l'homogénéité des sous-ensembles d'apprentissage.

Critère d'impureté

Notons $p = (p_1, ..., p_g)$ le vecteur contenant les proportions de chacune des g modalités de Z dans un ensemble considéré. Une mesure classique est l'*indice de Gini* G :

$$G(p) = \sum_{k=1}^{g} p_k(1 - p_k) \tag{1}$$

Lorsqu'une seule modalité de Z est présente dans l'ensemble considéré, $G(p) = 0$. $G(p)$ est maximal lorsque toutes les modalités sont en proportions égales.

Il faut trouver la séparation (couple variable-seuil) minimisant l'indice de Gini, donc maximisant l'homogénéité.

Construction de l'arbre : régularisation (1)

On cherche un arbre qui soit capable de bien classer l'ensemble d'apprentissage, tout en restant simple pour éviter le sur-apprentissage et donc pouvoir généraliser à de nouvelles données.

Stratégie **pré-élagage** : arrêter la croissance de l'arbre lorsque l'effectif d'une feuille est trop faible ou que la pureté est jugée suffisante.

Construction de l'arbre : régularisation (2)

Stratégie **post-élagage** : développer l'arbre complet puis élaguer certaines branches.

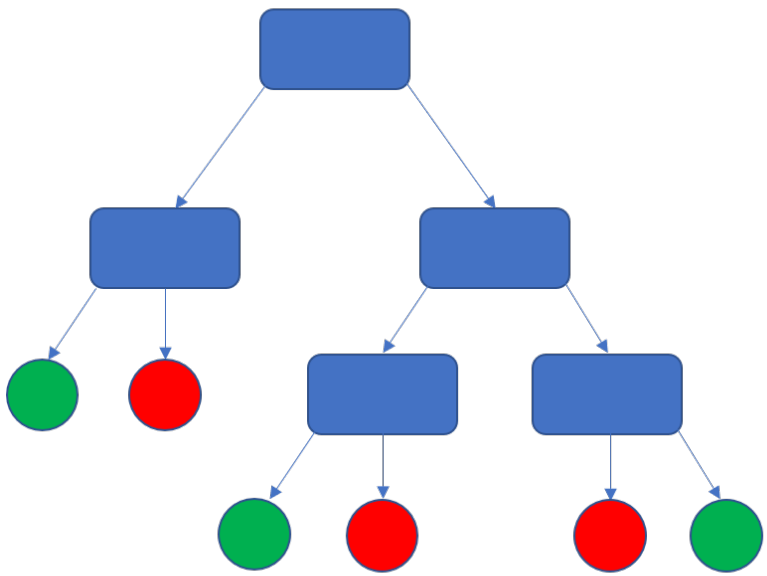
$$\eta(A) = \mathit{Precision}(A) + \lambda \mathit{Complexité}(A) \tag{2}$$

Precision(A) estime la précision d'un arbre A sur l'ensemble d'apprentissage (taux d'erreur d'apprentissage en décision ou somme des carrés des résidus en régression) et *Complexité* indique la complexité (nombre de feuilles de A). Le paramètre λ règle le compromis entre la précision et la complexité de l'arbre.

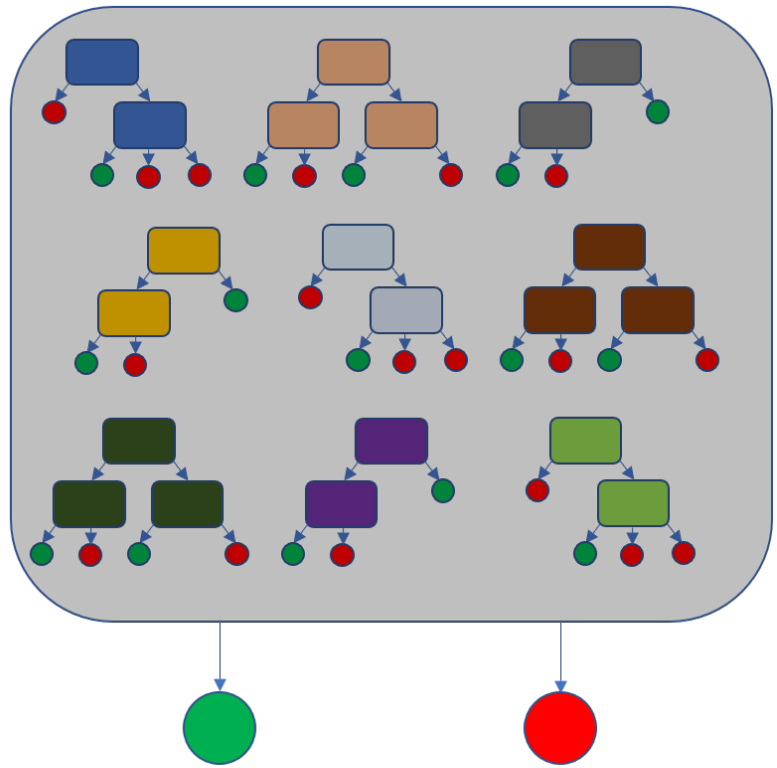
Si $\lambda = 0$, alors l'arbre est complet. Plus λ augmente, plus l'arbre se simplifie jusqu'à atteindre la racine.

Forêts aléatoires

Les forêts aléatoires sont un ensemble d'arbres entrainés sur un sous-ensemble des données d'apprentissage et un nombre réduit de variables explicatives. La prédiction est un vote majoritaire sur les classes prédites par les arbres.

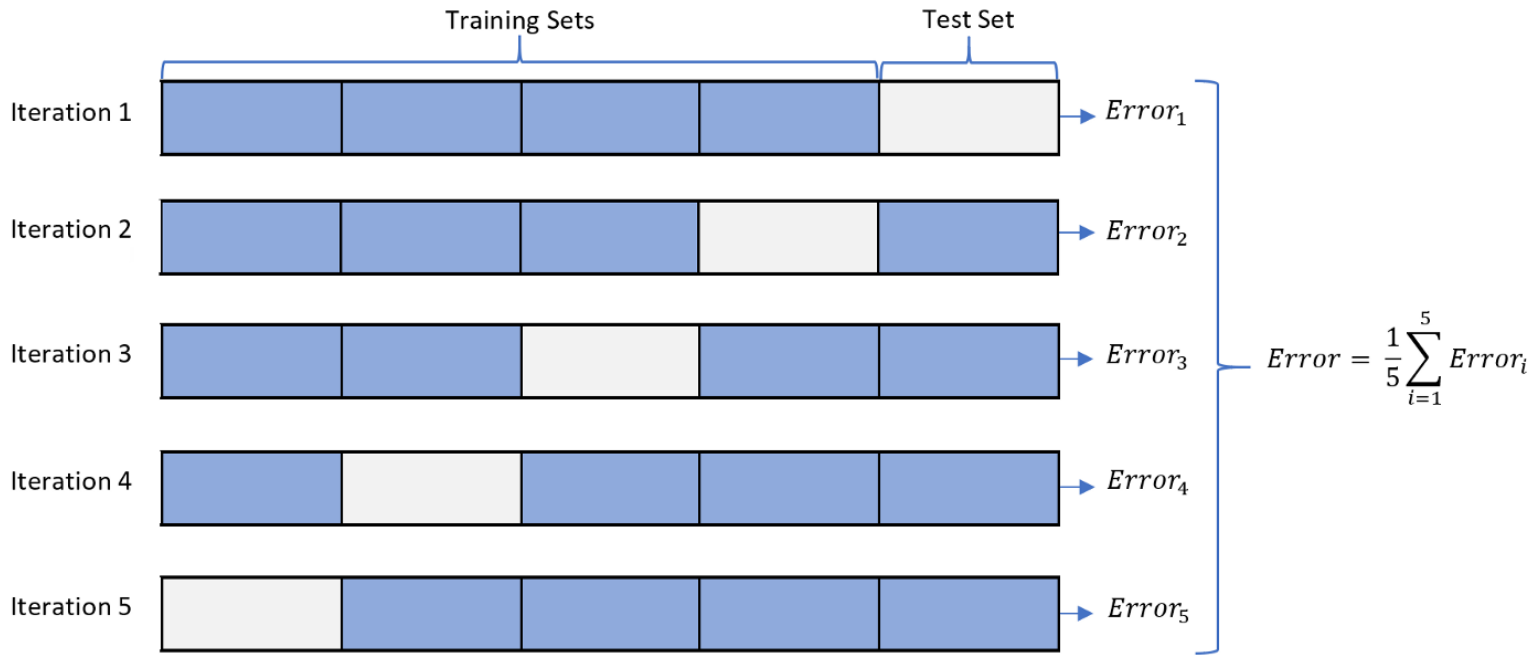


Decision Tree



Random Forest

Validation croisée



https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b