Arbres de classification / régression + Boosting.

I. Décisions par arbres (Et régression)

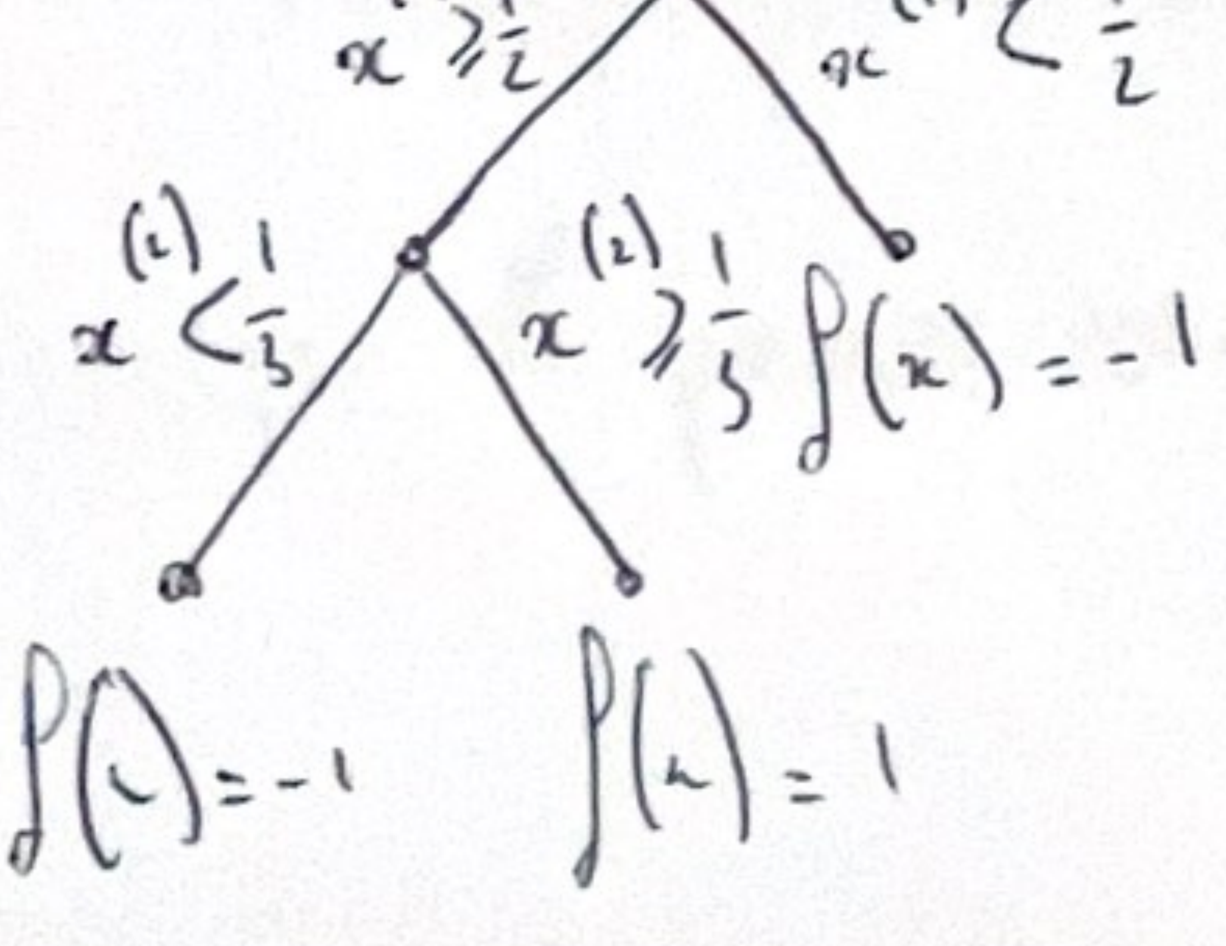
Contexte: Trouver $\hat{f}$ tq $\hat{f}$ rend petite l'erreur empirique

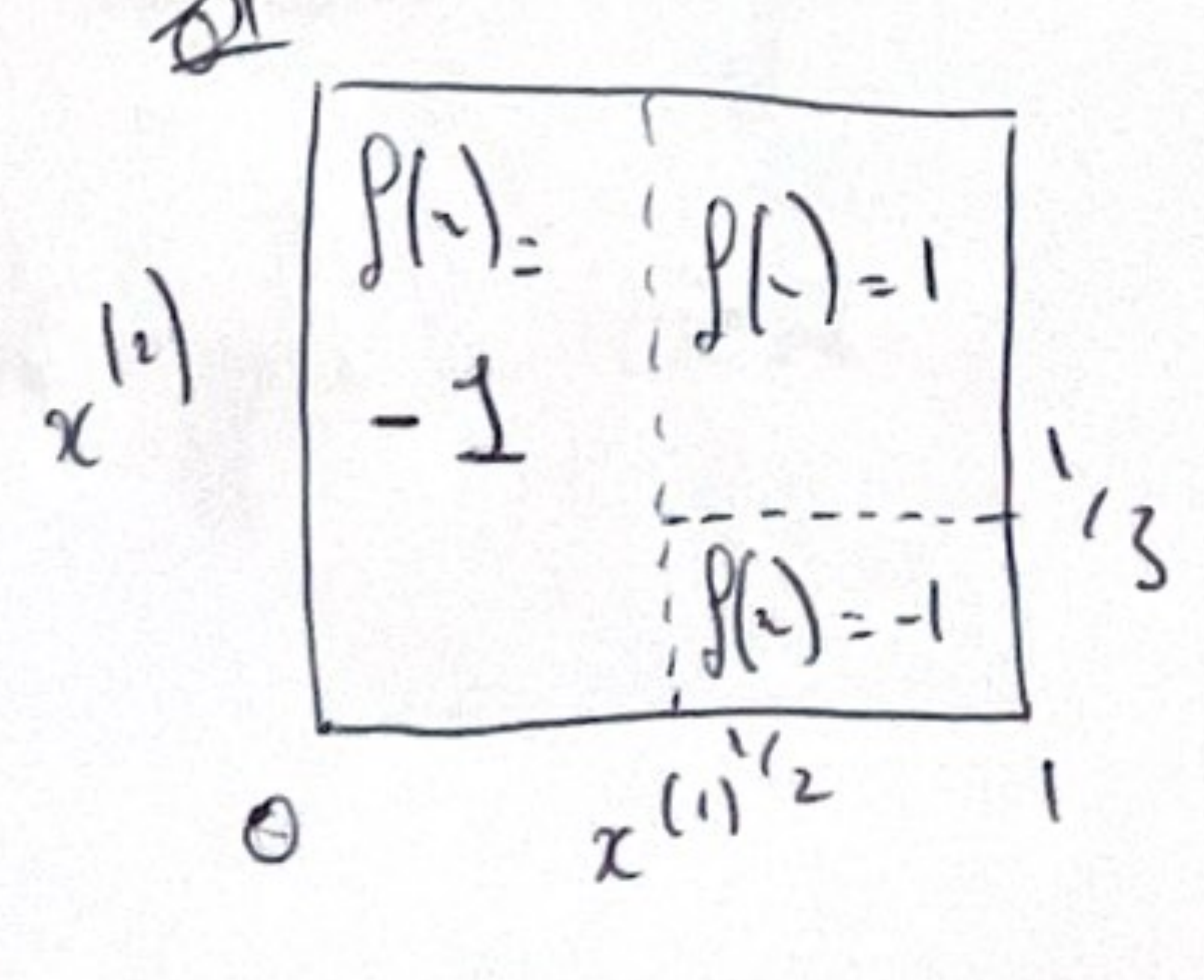
$$\hat{R}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \hat{f}(x_i))$$

- Classification : $\ell(y,z) = \mathbb{1}_{\mathbb{R}_-}(yz) \quad y \in \{-1,1\}$
- Régression : $\ell(y,z) = (y-z)^2$

Chercher $\hat{f}$ sous la forme d'un "algorithme" se représentant sur un arbre.

Exemple $x = (x^{(1)}, x^{(2)}) \in [0,1]^2$, $y \in \{-1,1\}$





Cet arbre est-il unique?

<u>Avantage principal</u>: Interprétabilité !

<u>Formalisation du problème</u> :

- $x = (x^{(1)}, \dots, x^{(d)})$ où $x^{(1)}, \dots, x^{(d)}$ sont des variables explicatives.

- $y \in \begin{cases} \{1, \dots, k\} & \text{pour un problème de } \cancel{\text{régression}} \text{ classification à } k \text{ classes} \\ \in \mathbb{R}^{d'} & \text{pour un problème de régression.} \end{cases}$

<u>définition</u>: Une fonction f se représente sous la forme d'un arbre si $f(x)$ peut être calculé avec un graphe de calculs représentable sur un arbre tel que

- Chaque noeud interne a exactement deux fils.
- Le critère qui détermine si on doit partir dans le fils droit ou gauche porte sur une variable explicative unique
- Chaque feuille encode une valeur de y.

<u>Remarque</u>: Le premier noeud est appelé la racine.

Définition : Une division pour un noeud est dite admissible si aucun des fils qu'elle engendre n'est $\emptyset$.

↳ On élimine les divisions triviales.

II. Entraînement : Choix des divisions

Il reste à expliquer :

- Comment ~~on choisit~~ choisir de bonnes divisions
- Quoi faire dans le cas d'une feuille
- Expliquer quand on doit s'arrêter ou créer une feuille.

Quand s'arrêter :

Un noeud devient une feuille si :

- Il est homogène (tous les ys des xs concernés par le noeud sont les mêmes
- (ou) Si le nombre de (x,y)s concernés par ce noeud est trop faible (hyperparamètre)
- (ou) Si le noeud est trop profond (hyperparamètre).

Quoi faire en cas de feuille

- En régression : Prédire la moyenne empirique des ys concernés par la feuille

- En classification : Prédire la classe majoritaire parmis les ys concernés par la feuille.

Comment séparer un noeud en deux fils :

Soit D (pour "discrepancy") une f mesure de non-uniformité.

Pour une variable explicative J,

$D(\{g_1, \dots, g_n\})$ mesure à quel point les points sont différents.

Alors, pour un noeud N, l'algorithme cherche à décomposer N sous forme de deux fils tq

$$(F_{gauche}, F_{droit}) \in \underset{\substack{(F_g', F_d') \\ \neq\emptyset \ \neq\emptyset}}{\operatorname{argmin}} \underbrace{D(F_g') + D(F_d')}_{\text{Essayer de maximiser la cohésion au sein de chaque fils.}}$$

III. Mesures de ~~[illegible]~~ non-uniformité

1) Pour la ~~classification~~ régression

Pour la ~~classification~~ régression, on utilise la variance empirique

$$D(\{y_1, \dots, y_{n'}\}) = \sum_{i=1}^{n'} (y_i - \bar{y})^2$$

où $\bar{y} = \sum_{i=1}^{n'} y_i$

2) Pour la classification

Pour la classification, on construit $(p_1, \dots, p_6)$ le vecteur des probabilités empiriques des classes au noeud N

Il existe deux mesures de non-uniformité (au choix)

Entropie $D = -\sum_i p_i \log(p_i)$ conv : $0 \times \log(0) = 0$

$\Rightarrow$ maximale pour $(p_1, \dots, p_6) = \left(\frac{1}{6}, \dots, \frac{1}{6}\right)$

$\Rightarrow$ minimale si $\exists i$ tq $p_i = 1$.

Mesure la quantité d'information.

Coefficient de Gini

$$D = |N| \sum_{i=1}^{\ell} p_i (1-p_i)$$

IV Boosting - Random Forests

L'idée générale du boosting est qu'il peut être plus intéressant de combiner des estimateurs aux performances modestes plutôt que d'essayer d'en construire un unique.

1) Bagging

L'idée générale du bagging est de

1) Pour b allant de 1 à B
   - Sampler avec remplacement un sous-jeu de données.
   - Entraîner $\hat{f}_b$ sur ce sous-jeu de données.

2) Combiner $\{f_1, \dots, f_b\}$ en un estimateur $\hat{f}$
   - En faisant la moyenne empirique pour de la régression
   - En ~~les~~ retournant la classe majoritaire pour de la classification.

2) Random Forests

L'algorithme de Random Forest est le résultat de la méthode de bagging appliquée avec les arbres de classification et de régression.

<u>Exercice</u> Quel est la prédiction aux points $(x^{(1)}, x^{(2)}) = (0,2 ; 0,4)$

$(0,8 ; 0,3)$

$(0,5 ; 0,5)$

pour de la classification si les arbres de la forêt sont