Mondrian Forests

HMMA 308 : Apprentissage Statistique

KANDOUCI Walid

https://github.com/WalidKandouci/HMMA308--Mondrian-Forests

Université de Montpellier

Sommaire

Introduction

Arbres de Mondiran

Application

Conclusion

Introduction

Rappel forêts aléatoires:

- Algorithme bagging
- Bon résultat sur de vrai données
- simples à mettre en œuvre

Introduction

Forêts Mondrian:

- Nouvelle classe des forêts aléatoires
- Efficaces
- Offre meilleure estimation d'incertitude que les forêts aléatoires

Arbres de Mondrian

L'approche:

- chaque noeud j a exactement un noeud parent, sauf pour un noeud racine distingué ϵ qui n'a pas de parents
- chaque noeud j est le parent d'exactement zéro ou deux noeuds enfants : (le noeud de gauche "$left(j)$" et le noeud de droite "$right(j)$")

Arbres de Mondrian

- (T, δ, ξ) un arbre de décision
- parent (j) : le parent du noeud j
- $N(j)$: 1 l'indice de nos données d'apprentissage au point $j(N(j) = \{n \in \{1, \ldots, N\} : x_n \in B_j\})$
- $\mathcal{D}_{N(j)} = \left\{ \boldsymbol{X}_{N(j)}, Y_{N(j)} \right\}$: les caractéristiques et les étiquettes des points de données d'apprentissage au noeud j
- $\cdot \ell_{jd}^x$ et u_{jd}^x : les bornes inférieurs et supérieurs de nos donnes d'apprentissage au noeud j le long de la dimension d
- $\cdot B_j^x = \left(\ell_{j1}^x, u_{j1}^x \right] \times \ldots \times \left(\ell_{jD}^x, u_{jD}^x \right] \subseteq B_j$: le plus petit rectangle qui entoure les données d'apprentissage au noeud j

Arbres de Mondrian

L'algorithme:

Algorithm 1 SampleMondrianTree($\lambda, \mathcal{D}_{1:n}$)

Initialisation: $T = \emptyset$, leaves$(T) = \emptyset, \delta = \emptyset, \xi = \emptyset, \tau = \emptyset, N(\epsilon) = \{1, 2, \ldots, n\}$
SampleMondrianBlock$(\epsilon, \mathcal{D}_{N(\epsilon)}, \lambda)$

Algorithm 2 SampleMondrianBlock$(j, \mathcal{D}_{N_{(j)}}, \lambda)$

Add j to T
For all d, set $\ell^x_{jd} = \min(\boldsymbol{X}_{N(j),d}), u^x_{jd} = \max(\boldsymbol{X}_{N(j),d})$
Sample E from exponential distribution with rate $\sum_d (u^x_{jd} - \ell^x_{jd})$
if $\tau_{parent\,(j)} + E < \lambda$ **then**
  Set $\tau_j = \tau_{\text{parent}\,(j)} + E$
  Sample split dimension δ_j, choosing d with probability proportional to $u^x_{jd} - \ell^x_{jd}$
  Sample split location ξ_j uniformly from interval $[\ell^x_{j\delta_j}, u^x_{j\delta_j}]$
  Set $N(\text{left}(j)) = \left\{n \in N(j) : \boldsymbol{X}_{n,\delta_j} \leq \xi_j\right\}$ and $N(\text{ right }(j)) = \left\{n \in \mid N(j) : \boldsymbol{X}_{n,\delta_j} > \xi_j\right\}$
  SampleMondrianBlock $\left(\text{left}(j), \mathcal{D}_{N(\text{left}(j))}, \lambda\right)$
  SampleMondrianBlock $\left(\text{ right } (j), \mathcal{D}_{N(\text{ right }(j))}, \lambda\right)$
else Set $\tau_j = \lambda$ and add j to leaves (T)

Arbres de Mondrian

Les arbres de Mondrian diffèrent des arbres de décision comme suit:

- Les divisions sont échantillonnées indépendamment des $Y_{N(j)}$
- Chaque noeud j est associé avec un temps de division τ_j
- λ contrôle le nombre totale des divisions
- la division représenté par un noeud interne j ne tient que dans B_j^x et non pas B_j

Application

Fonction "*random_mondrian*"

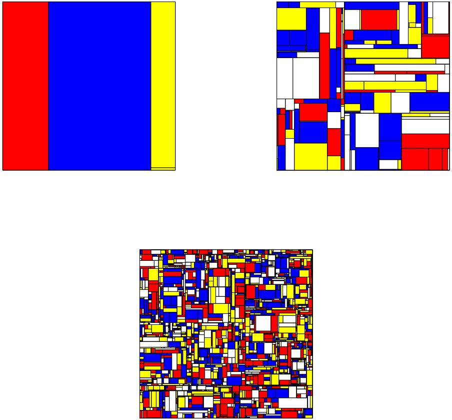

Figure: Exemple générateur aléatoire Mondrian (budget=2,10,50)

Conclusion

- Forêts de Mondrian = Processus de Mondrian + Forêts aléatoires
- Peut fonctionner en mode batch ou en mode en ligne
- Meilleur estimation d'incertitude que les forêts aléatoires