

Réconcilier réseaux de neurones artificiels et interprétabilité

Benjamin Leblanc

16 mai 2023

Réconcilier réseaux de neurones artificiels et interprétabilité

Benjamin Leblanc

16 mai 2023

Nous définissons le degré d'interprétabilité

Nous définissons le degré d'interprétabilité d'un prédicteur

Nous définissons le degré d'interprétabilité d'un prédicteur par la capacité d'un non-expert

Nous définissons le degré d'interprétabilité d'un prédicteur par la capacité d'un non-expert à comprendre son processus de décision

Nous définissons le degré d'interprétabilité d'un prédicteur par la capacité d'un non-expert à comprendre son processus de décision en ne considérant que le modèle en lui-même.

Tâche: prédire le loyer mensuel d'un appartement dans la ville de Québec

Modèle linéaire

Arbre de décision

Tâche: prédire le loyer mensuel d'un appartement dans la ville de Québec

Modèle linéaire

$$\hat{y} = 70\$ + \\ 90\text{¢} \cdot \text{nombre de pieds carré} + \\ 58\$ \cdot \text{nombre de pièces}$$

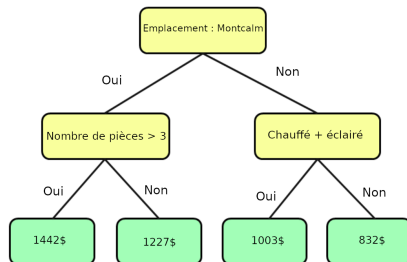
Arbre de décision

Tâche: prédire le loyer mensuel d'un appartement dans la ville de Québec

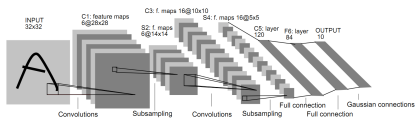
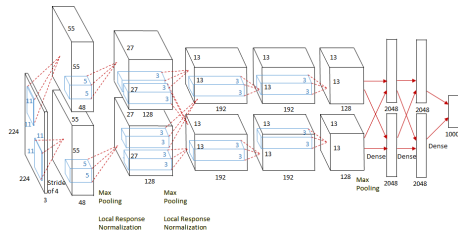
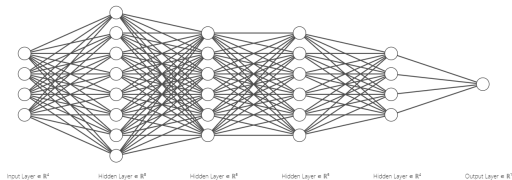
Modèle linéaire

$$\hat{y} = 70\$ + \\ 90\text{¢} \cdot \text{nombre de pieds carré} + \\ 58\$ \cdot \text{nombre de pièces}$$

Arbre de décision



Réseaux de neurones artificiels



Réseaux de neurones artificiels

Réseaux de neurones artificiels

- Ici : réseaux pleinement connectés

Réseaux de neurones artificiels

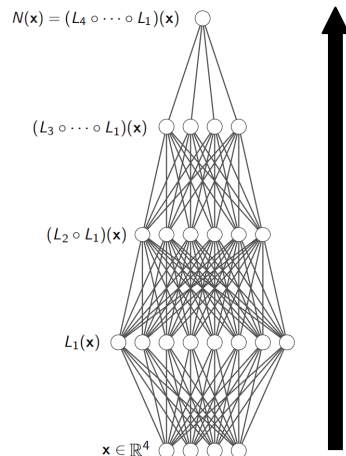
- Ici : réseaux pleinement connectés
- $B(\mathbf{x}) = (L_I \circ \dots \circ L_1)(\mathbf{x}) = L_I(\dots L_2(L_1(\mathbf{x})) \dots)$

Réseaux de neurones artificiels

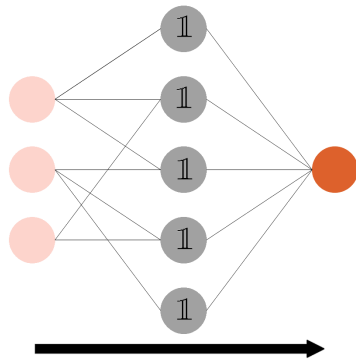
- Ici : réseaux pleinement connectés
- $B(\mathbf{x}) = (L_I \circ \dots \circ L_1)(\mathbf{x}) = L_I(\dots L_2(L_1(\mathbf{x})) \dots)$
- $L_k(\mathbf{x}) = g_k(\mathbf{W}_k \mathbf{x}) \quad \forall k \in \{1, \dots, I\}$

Réseaux de neurones artificiels

- Ici : réseaux pleinement connectés
- $B(\mathbf{x}) = (L_I \circ \dots \circ L_1)(\mathbf{x}) = L_I(\dots L_2(L_1(\mathbf{x})) \dots)$
- $L_k(\mathbf{x}) = g_k(\mathbf{W}_k \mathbf{x}) \quad \forall k \in \{1, \dots, I\}$

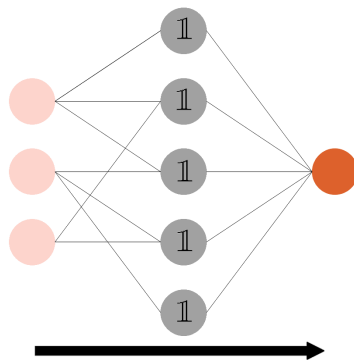


Réseaux de neurones simples



Réseaux de neurones simples

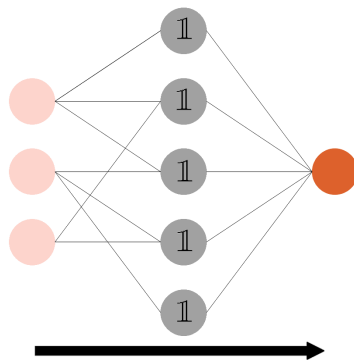
- Activations binaires (fonction *seuil*)



Réseaux de neurones simples

- Activations binaires (fonction *seuil*)

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{si } x \text{ est vrai,} \\ 0 & \text{sinon.} \end{cases}$$

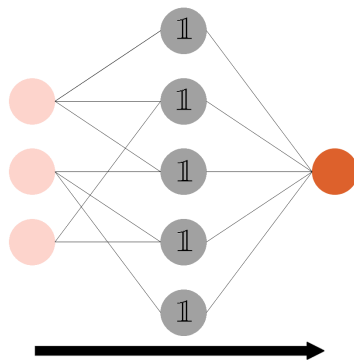


Réseaux de neurones simples

- Activations binaires (fonction *seuil*)

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{si } x \text{ est vrai,} \\ 0 & \text{sinon.} \end{cases}$$

- Peu profonds (1 couche cachée)

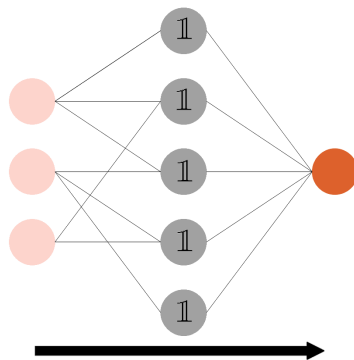


Réseaux de neurones simples

- Activations binaires (fonction *seuil*)

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{si } x \text{ est vrai,} \\ 0 & \text{sinon.} \end{cases}$$

- Peu profonds (1 couche cachée)
- Peu larges

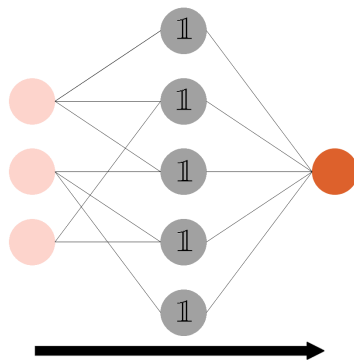


Réseaux de neurones simples

- Activations binaires (fonction *seuil*)

$$\mathbb{1}_{\{x\}} = \begin{cases} 1 & \text{si } x \text{ est vrai,} \\ 0 & \text{sinon.} \end{cases}$$

- Peu profonds (1 couche cachée)
- Peu larges
- Clairsemés (x2)

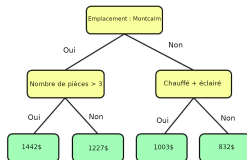


Réseaux de neurones artificiels simples

Tâche: prédire le loyer mensuel d'un appartement dans la ville de Québec

Modèle linéaire + **Arbre de décision** = **Réseau à activation binaire**

$$\hat{y} = 70\$ + \\ 90\text{¢} \cdot \text{nombre de pieds carré} + \\ 58\$ \cdot \text{nombre de pièces}$$



Réseaux de neurones artificiels simples

Tâche: prédire le loyer mensuel d'un appartement dans la ville de Québec

Modèle linéaire

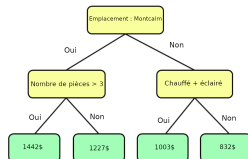
+

Arbre de décision

=

Réseau à activation binaire

$$\hat{y} = 70\$ + \\ 90\text{¢} \cdot \text{nombre de pieds carré} + \\ 58\$ \cdot \text{nombre de pièces}$$



$$\hat{y} = 800 + \\ 200 \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ 160 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ 110 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}}$$

Construire le réseau : BGN

$$\begin{aligned}\hat{y} = & 800 + \\ & 200 \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ & 160 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ & 110 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}}\end{aligned}$$

Caractéristiques

$$\begin{aligned}\hat{y} = & 800 + \\ & 200 \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ & 160 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ & 110 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}}\end{aligned}$$

Caractéristiques

- En vue d'attaquer des tâches de régression

$$\hat{y} = 800 + 200 \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + 160 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + 110 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 800 + 200 \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + 160 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + 110 \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 980\$$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 980\$ +$$

$$\mathbb{1}_{\{\text{nombre de pièces} > 4\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 920\$ + 180\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 920\$ + 180\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 840\$ + 220\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + 155\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 840\$ + 220\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + 155\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\hat{y} = 800\$ + 220\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + 160\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + 110\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)

$$\begin{aligned}\hat{y} = & 800\$ + \\ & 200\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ & 160\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ & 110\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}} + \\ & \dots\end{aligned}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)
- Garantie de convergence sur l'erreur en entraînement

$$\begin{aligned}\hat{y} = & 800\$ + \\ & 200\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ & 160\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ & 110\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}} + \\ & \dots\end{aligned}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)
- Garantie de convergence sur l'erreur en entraînement
- Architecture n'est pas fixe

$$\begin{aligned}\hat{y} = & 800\$ + \\ & 200\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ & 160\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ & 110\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}} + \\ & \dots\end{aligned}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)
- Garantie de convergence sur l'erreur en entraînement
- Architecture n'est pas fixe
- Fait usage d'une régularisation L1

$$\begin{aligned}\hat{y} = & 800\$ + \\ & 200\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ & 160\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ & 110\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}} + \\ & \dots\end{aligned}$$

Caractéristiques

- En vue d'attaquer des tâches de régression
- Algorithme glouton (inspiré d'Adaboost)
- Garantie de convergence sur l'erreur en entraînement
- Architecture n'est pas fixe
- Fait usage d'une régularisation L1
- Pas d'hyperparamètre (taux d'apprentissage, taille des *batch*, ...)

$$\begin{aligned}\hat{y} = & 800\$ + \\ & 200\$ \cdot \mathbb{1}_{\{\text{nombre de pièces} > 4\}} + \\ & 160\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1200\}} + \\ & 110\$ \cdot \mathbb{1}_{\{\text{nombre de pieds carré} > 1400\}} + \\ & \dots\end{aligned}$$

Testons l'*interprétabilité* de notre approche!

La tâche

Testons l'*interprétabilité* de notre approche!

La tâche

- Prédire la valeur d'une maison (mesuré en 1k USD)

Testons l'*interprétabilité* de notre approche!

La tâche

- Prédire la valeur d'une maison (mesuré en 1k USD)
- Caractéristiques retenues (parmi huit):

Testons l'*interprétabilité* de notre approche!

La tâche

- Prédire la valeur d'une maison (mesuré en 1k USD)
- Caractéristiques retenues (parmi huit):
 - L'âge median des maisons du même quartier (MedAge)

Testons l'*interprétabilité* de notre approche!

La tâche

- Prédire la valeur d'une maison (mesuré en 1k USD)
- Caractéristiques retenues (parmi huit):
 - L'âge median des maisons du même quartier (MedAge)
 - Le nombre médian de chambre à coucher des maisons du même quartier (MedBed)

Testons l'*interprétabilité* de notre approche!

La tâche

- Prédire la valeur d'une maison (mesuré en 1k USD)
- Caractéristiques retenues (parmi huit):
 - L'âge median des maisons du même quartier (MedAge)
 - Le nombre médian de chambre à coucher des maisons du même quartier (MedBed)
 - Le revenu median des ménages des maisons du même quartier (mesuré en 1k USD) (MedInc)

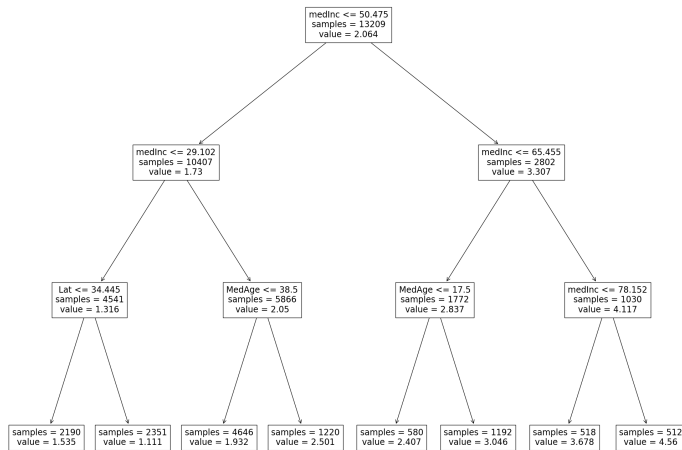
Testons l'*interprétabilité* de notre approche!

La tâche

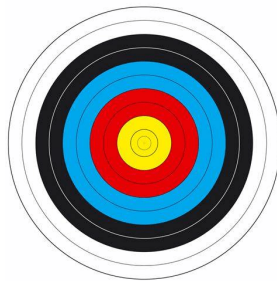
- Prédire la valeur d'une maison (mesuré en 1k USD)
- Caractéristiques retenues (parmi huit):
 - L'âge median des maisons du même quartier (MedAge)
 - Le nombre médian de chambre à coucher des maisons du même quartier (MedBed)
 - Le revenu median des ménages des maisons du même quartier (mesuré en 1k USD) (MedInc)

$$\begin{aligned} B(\text{MedInc}, \text{MedAge}, \text{MedBed}) = & \\ & 100 + \\ & 101 \cdot \mathbb{1}_{\{0.59 \cdot \text{MedAge} + \text{MedInc} > 63.56\}} + \\ & 127 \cdot \mathbb{1}_{\{0.08 \cdot \text{MedAge} + \text{MedInc} > 65.46\}} + \\ & 60 \cdot \mathbb{1}_{\{\text{MedInc} > 28.20\}} + \\ & 36 \cdot \mathbb{1}_{\{\text{MedBed} > 2.0\}} + \\ & 27 \cdot \mathbb{1}_{\{\text{MedAge} > 20.0\}} \end{aligned}$$

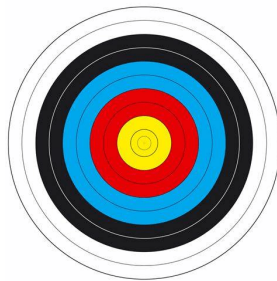
Testons l'interprétabilité de notre approche!



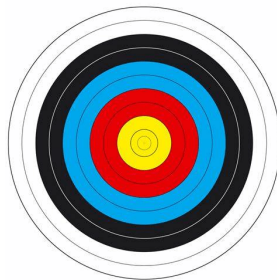
$$\begin{aligned} B(\text{MedInc}, \text{MedAge}, \text{MedBed}) = & \\ & 100 + \\ & 101 \cdot \mathbb{1}_{\{0.59 \cdot \text{MedAge} + \text{MedInc} > 63.56\}} + \\ & 127 \cdot \mathbb{1}_{\{0.08 \cdot \text{MedAge} + \text{MedInc} > 65.46\}} + \\ & 60 \cdot \mathbb{1}_{\{\text{MedInc} > 28.20\}} + \\ & 36 \cdot \mathbb{1}_{\{\text{MedBed} > 2.0\}} + \\ & 27 \cdot \mathbb{1}_{\{\text{MedAge} > 20.0\}} \end{aligned}$$



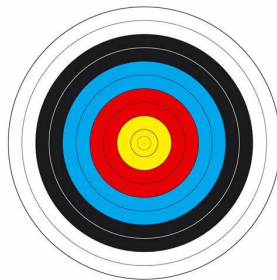
✓ Tâches multivariées



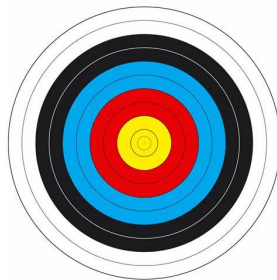
- ✓ Tâches multivariées
- ✓ Construire des réseaux de neurones profonds



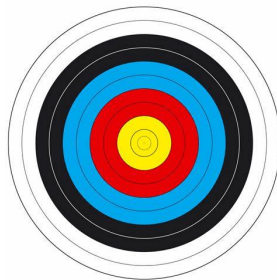
- ✓ Tâches multivariées
- ✓ Construire des réseaux de neurones profonds
- ✓ Méthode éprouvée sur des données de la vie réelle



- ✓ Tâches multivariées
- ✓ Construire des réseaux de neurones profonds
- ✓ Méthode éprouvée sur des données de la vie réelle
- ✓ Monotonie du prédicteur



- ✓ Tâches multivariées
- ✓ Construire des réseaux de neurones profonds
- ✓ Méthode éprouvée sur des données de la vie réelle
- ✓ Monotonie du prédicteur
- ✓ Quantifier la stabilité du prédicteur



Les **réseaux de neurones artificels** peuvent être **interprétables**...

Les **réseaux de neurones artificels** peuvent être **interprétables**...
... Lorsque entraînés avec ce but en tête!

Merci de votre attention :)

benjamin.leblanc.2@ulaval.ca

ArXiv : **A Greedy Algorithm for Building Compact Binary Activated Neural Networks**