Explicabilité des réseaux de neurones sur graphes au niveau des embeddings

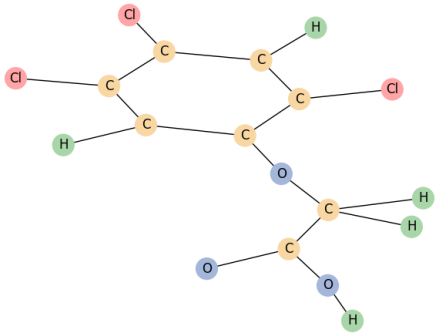
Elouan Vincent

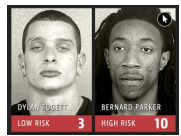
Référent: Marc Plantevit

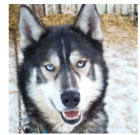
LRE, EPITA

10 janvier 2023

Graphes et GNN







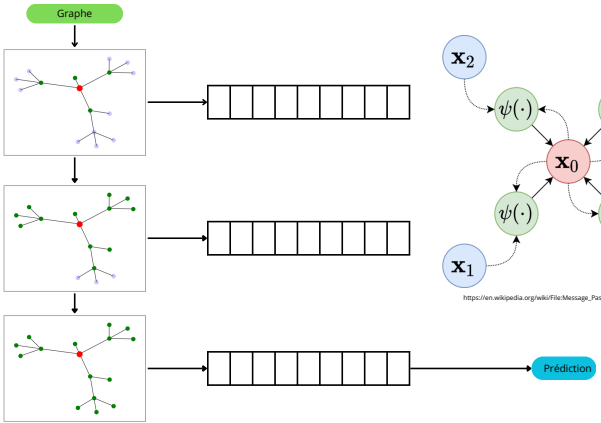
(a) Husky classified as wolf

(b) Explanation

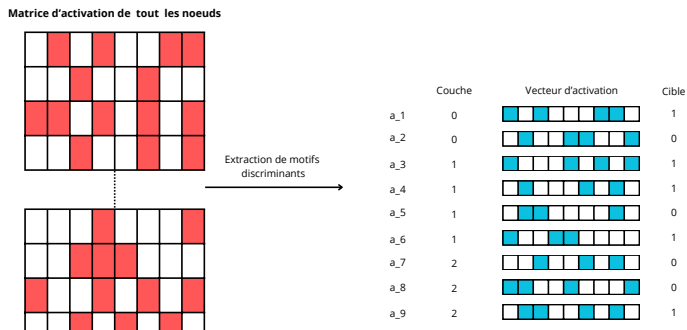
- Ethique, légale, debug, extraction de connaissances
- Peu de méthode sur les GNN

Yuan, H., Yu, H., Gui, S., and Ji, S. (2022). Explainability in graph neural networks: A taxonomic survey.

GNN - Comment ça marche ?



GNN - Règle d'activation



Veyrin-Forrer et al. "On GNN explanability with activation patterns" 2021

Objectifs

- On sait extraire des motifs d'activation intéressants appris par le modèle
- On arrive à extraire des sous-graphes importants pour chaque règle d'activation
- On veut pouvoir caractériser les *features* capturées par le modèle

On veut maintenant comprendre comment les règles interagissent entre elle, et comment elle participe aux prédictions du modèle

Shapley - Théorie

Principes

- Issu de la théorie des jeux.
- Calcul une contribution "équitable" d'un ensemble de joueurs à un résultat.

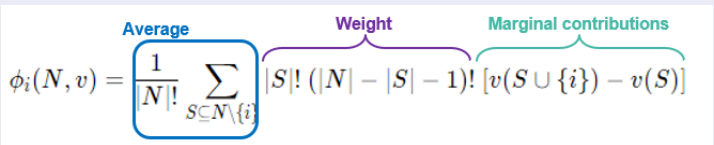
 Hart, S. (1989). Shapley Value. In: Eatwell, J., Milgate, M., Newman, P. (eds) Game Theory. The New Palgrave. Palgrave Macmillan, London.

Shapley - Théorie

Principes

- Issu de la théorie des jeux.
- Calcul une contribution "équitable" d'un ensemble de joueurs à un résultat.

Moyenne des contributions marginales

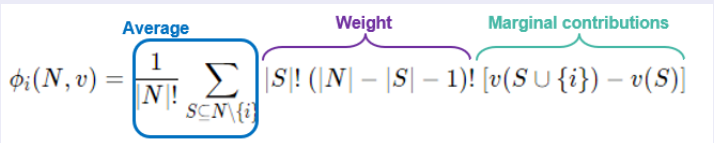


https://medium.com/the-modern-scientist/what-is-the-shapley-value-8ca624274d5a

Hart, S. (1989). Shapley Value. In: Eatwell, J., Milgate, M., Newman, P. (eds) Game Theory. The New Palgrave. Palgrave Macmillan, London.

Shapley - Théorie

Moyenne des contributions marginales



https://medium.com/the-modern-scientist/what-is-the-shapley-value-8ca624274d5a

- N un ensemble de joueur
- v une fonction de valeur
- S une coalition, un sous-ensemble de N
- Φ_i la contribution du joueur i

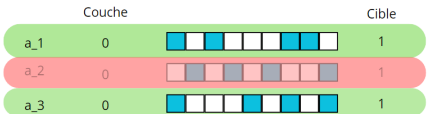
Shapley - Pour les règles d'activation

La valeur de Shapley permet de calculer de manière équitable un score de contribution par rapport à un résultat

Fonction de valeur

- Ensemble de joueurs N : les règles d'activations
- Un jeu pour chaque classe du jeu de donnée
- v, la prédiction moyenne du modèle sur le jeu de donnée sur la classe ciblée.

Shapley - Les coalitions



S = {a_1, a_3}

Masque couche 0

1	0	1	0	1	0	1	1	1

Deux options possibles :

- Désactiver en multipliant la sortie par 0.
- Remplacer par l'activation moyenne de cette composante dans le jeu de donnée.

SHAP - Importance de variable

Limite de la valeur de Shapley

- $2^{30} = 1\ 073\ 741\ 824$ coalitions possibles sur 30 règles.
- Besoin de l'approximer.

SHAP

- Méthode d'estimation de la valeur de Shapley
- Implémentée pour calculer l'importance des variables dans des données tabulaires pour une instance donnée

KernelSHAP

- Tire des coalitions au hasard, leur assigne un poids π, et entraine un modèle linéaire en fonction des prédictions du modèle pour les coalitions.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017)

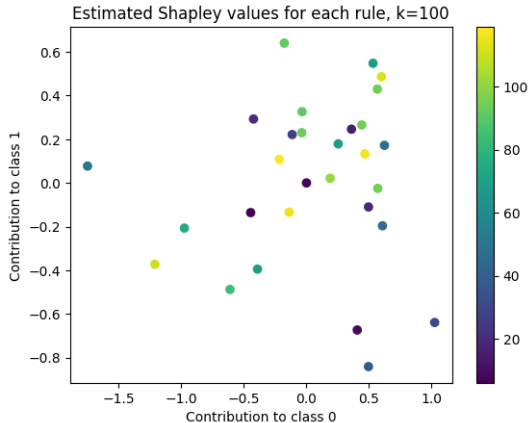
SHAP - Pour les règles d'activation

On garde la même fonction de valeur et méthode d'exclusion de joueur précédente.

Les étapes

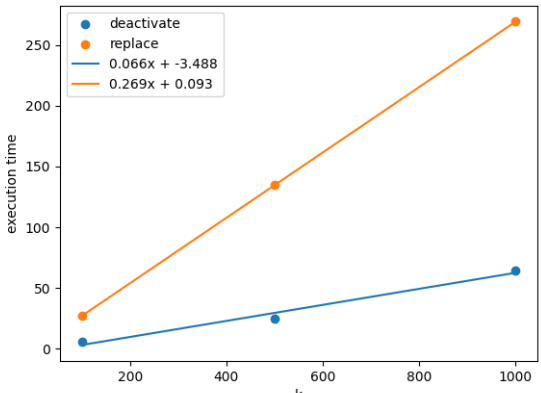
- Tirer k coalitions.
- Récupérer la prédiction pour S en utilisant le modèle perturbé.
- Calculer le poids de la coalition en utilisant π
- "Fit" le modèle linéaire pondéré
- Retourner les contributions des règles, qui sont les coéfficients du modèle linéaire

Résultats



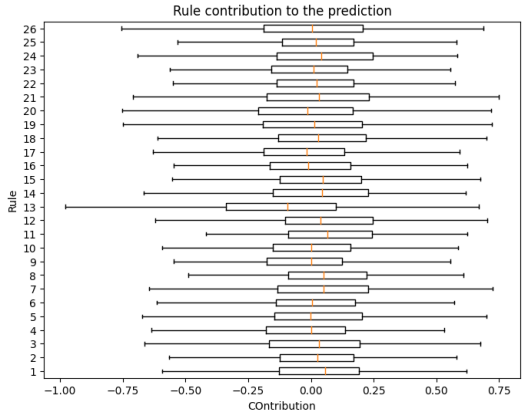
Contributions calculées par notre méthode sur les deux classes, le score d'InsideGNN en couleur

Résultats



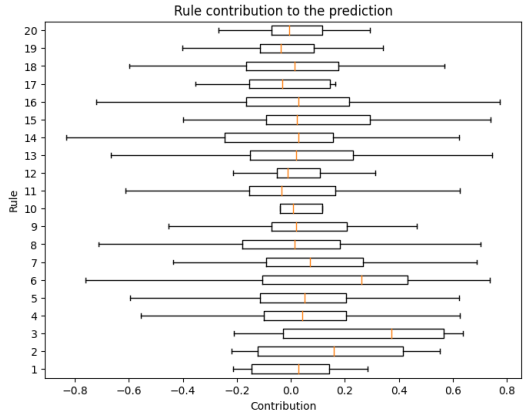
Temps d'execution des deux méthodes d'exclusion pour différentes valeurs de k

Résultats



Distribution des contributions des règles pour la prédiction de chaque instance

Résultats



Pour chaque règles, distribution des contribution à chaque instance où la règle est activée dans le modèle non-perturbé. *classe 0, k=100*

Conclusion

Contributions

- Adaptation de la valeur de Shapley sur les règles d'activation
- Implémentation de SHAP pour les règles d'activation

Perspectives

- Retravailler la fonction de valeur
- Essayer d'autres méthodes d'exclusion des joueurs
- Réduire le type des instances utilisées dans la fonction de valeur
- Reformuler le problème
- Optimiser le tirage des coalitions

Merci !

Des questions ?

Bibliographie

- Veyrin-Forrer et al. "On GNN explanability with activation patterns" 2021
- Hart, S. (1989). Shapley Value. In: Eatwell, J., Milgate, M., Newman, P. (eds) Game Theory. The New Palgrave. Palgrave Macmillan, London.
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017)
- Yuan, H., Yu, H., Gui, S., and Ji, S. (2022). Explainability in graph neural networks: A taxo- nomic survey.