Université de Strasbourg
Intelligence Artificielle
Licence 3 Informatique

Sujet 3 – Évaluer la contribution des attributs d'entrée dans un réseau de neurones artificiels

Modèle fondamental : réseau de neurones artificiels

A. Objectif

Pour ce sujet, votre objectif est de mettre en œuvre un réseau de neurones artificiels et d'explorer son comportement localement, en construisant une approximation interprétable de son processus décisionnel aux alentours de prédictions individuelles.

B. Travail à réaliser

1) Construire un réseau de neurones artificiel

En utilisant le code développé en TP, mettez en œuvre un réseau de neurones artificiel entièrement connecté de deux couches cachées de 16 et 8 unités respectivement, activées avec tanh, et entraîné sur votre ensemble d'entraînement en mini-batch.

2) Sélection d'instances

De votre ensemble de validation, sélectionnez trois instances correctement classées et trois instances incorrectement classées : choisissez des instances vous paraissant intéressantes, par exemple des instances de même classe mais prédites différemment par le modèle, etc.

3) Génération d'instances perturbées

Chaque instance sera une instance dont vous examinerez la contribution des attributs d'entrée à leur prédiction.

Pour chaque instance :

1. Générez 250 versions perturbées de l'instance en modifiant les attributs de l'instance par une petite quantité de bruit, en appliquant aléatoirement un bruit d'amplitude maximale de ± 10 % de la valeur originale de l'attribut.
2. Utilisez votre réseau de neurones pour prédire la classe de chaque instance perturbée.

À chaque instance sélectionnée est maintenant associé un petit jeu de données étiquetées qui va être utilisé pour entraîner un modèle interprétable approximant localement le réseau de neurones.

4) Entraînement de modèles interprétables locaux

Pour chaque instance, entraînez un modèle interprétable sur le jeu de données précédemment généré :

1. Vous pouvez choisir de mettre en œuvre une régression linéaire dont les paramètres sont entraînés par descente de gradient.
2. Alternativement, vous pouvez mettre en œuvre un arbre de décision, de profondeur maximale 4.

Dans l'un ou l'autre cas, les modèles doivent être mis en œuvre *from scratch* (si vous choisissez d'utiliser un arbre de décision, vous pouvez utiliser le modèle que vous avez développé lors des travaux pratiques).

5) Examen de la contribution des attributs

1. Interprétez le modèle local entraîné pour chaque instance sélectionnée à l'étape 2 pour évaluer le degré de contribution de chaque attribut à la prédiction réalisée par le réseau de neurones. Qu'est-ce qui constitue le « degré de contribution » d'un attribut de votre modèle local ?
2. Comparer ces contributions entre les instances correctement et incorrectement classées.
3. Proposez une visualisation pour soutenir votre analyse (e.g. des diagrammes en barres indicatives de la contribution des attributs).

6) Conclusion réflexive

1. Auriez-vous pu obtenir directement (i.e. sans passer par d'autres modèles) une estimation de la contribution des attributs des données aux décisions du réseau de neurones ?
2. En quoi les approximations locales vous permettent-elle de mieux comprendre comment votre réseau de neurones calcule une prédiction pour une entrée donnée ?
3. Quelles sont les forces et les faiblesses de cette approche locale pour évaluer la contribution des attributs aux décisions du modèle original ?

C. Items à évaluer

1. Extraire un batch de 4 instances du jeu de test, réaliser une prédiction avec le réseau de neurone et afficher les distribution de probabilité calculées.
2. En utilisant une matrice de confusion, indiquez quelle est la classe la plus difficile à prédire pour votre réseau de neurones.
3. En comparant l'exactitude d'entraînement et de validation de votre réseau de neurones, indiquez si le modèle a sur-appris (expliquez votre réponse).

4. Pour une instance donnée, montrez dans quelle mesure les instances perturbées donnent lieu aux mêmes prédictions (ou non), par le modèle original, que l'instance elle-même.

5. Montrez les instances que vous avez sélectionnez et pourquoi elles vous ont paru intéressantes

6. Montrez le modèle local construit pour une des instances sélectionnée et illustrez ses performances : permet-il d'approximer convenablement le comportement du réseau de neurones dans la localité de l'instance ?

7. Présentez un graphique montrant les contributions des attributs d'une instance pour sa prédiction.

8. Les contributions calculées par le modèle local seront-elles les mêmes si la génération des instances perturbées était différente (e.g. bruit de ± 20 % au lieu des 10 actuels) ?