UNIVERSITÉ SORBONNE PARIS NORD

Projet : Implémentation de LIME

Local Interpretable Model-agnostic Explanations

Master 1 Informatique
membres : - **Tesnime Ben Salem.**
-**Melissa HACHEMI.**

Date : Novembre 2025

Table des matières

1. Introduction
2. Méthodologie
3. Résultats et Analyses
4. Comparaison avec LIME officiel
5. Conclusion

1. Introduction

1.1 Contexte

L'intelligence artificielle moderne repose de plus en plus sur des modèles complexes (Random Forest, réseaux de neurones, gradient boosting) qui atteignent d'excellentes performances prédictives mais souffrent d'un problème majeur : leur **manque d'interprétabilité**. Ces modèles "boîtes noires" rendent leurs décisions opaques, ce qui pose des problèmes éthiques, légaux et pratiques, notamment dans des domaines sensibles comme la santé, la justice ou la finance.

LIME (Local Interpretable Model-agnostic Explanations), proposé par Ribeiro et al. en 2016, répond à ce défi en fournissant des **explications locales** : plutôt que de comprendre le modèle globalement, LIME explique pourquoi une prédiction spécifique a été faite pour une instance donnée.

1.2 Objectifs du projet

Ce projet vise à :

1. **Implémenter** l'algorithme LIME from scratch en Python
2. **Comprendre** les principes mathématiques sous-jacents (échantillonnage, pondération, régression locale)
3. **Analyser** l'influence des hyperparamètres (σ, n_samples) sur la qualité des explications
4. **Comparer** notre implémentation avec la bibliothèque officielle LIME
5. **Évaluer** les forces et limites de cette approche d'explainability

1.3 Principe de LIME

LIME repose sur une idée simple mais puissante : **approximer localement un modèle complexe par un modèle linéaire interprétable**

Réponse des questions

Quel dataset avez-vous choisi et pourquoi?

On a choisi le dataset California Housing car il est propre, réaliste et bien adapté à un problème de régression, ce qui permet de tester efficacement les performances d'un modèle prédictif.

Combien de features avez-vous ?	8
Quelle est la variable cible (y)?	MedHouseVal
Quelle est la taille de votre ensemble de test ?	4128
Quel score R2 obtenez-vous sur le test ?	R^2= 0.7955449233851258
Quelle est l'erreur MSE ?	MSE=0.2679197680430621

Pourquoi ce modèle est-il considéré comme une "boîte noire"?

Le Random Forest est considéré comme un modèle "boîte noire" parce qu'il est composé de nombreux arbres de décision entraînés sur des sous-échantillons et des sous-ensembles de variables. Les prédictions résultent de la moyenne ou du vote majoritaire de tous ces arbres, ce qui rend le processus de décision global très difficile à interpréter.

Pourquoi x∗ doit venir du test set et pas du train set?

Si x* vient du train set, le modèle a déjà "vu" cette instance pendant l'entraînement. L'explication pourrait refléter un sur-apprentissage plutôt que la vraie logique de généralisation du modèle.LIME vise à expliquer comment le modèle fait ses prédictions sur de nouvelles données.

Qu'est-ce qu'un "voisin" dans le contexte de LIME?

Un voisin dans LIME est un point créé en modifiant légèrement les caractéristiques de l'instance à expliquer, afin d'étudier comment le modèle change localement.

Pourquoi utiliser des perturbations gaussiennes?

LIME utilise des perturbations gaussiennes pour générer des voisins réalistes autour de l'instance à expliquer. La distribution normale permet de créer des points proches mais variés, cohérents avec la structure des données, ce qui stabilise l'explication locale.

Pourquoi doit-on prédire sur TOUS les voisins?

On doit prédire sur tous les voisins car :

On a besoin des labels y_hat_i = h(z_i) pour entraîner le modèle linéaire local g
Chaque voisin contribue à l'apprentissage du modèle linéaire avec un poids différent
Plus on a de points, meilleure sera l'approximation linéaire locale

Pourquoi les voisins proches ont-ils plus de poids?

Fidélité locale : LIME cherche à expliquer le comportement du modèle autour de x*. Les points proches de x* sont plus représentatifs du comportement local du modèle que les points éloignés.

σ (sigma) est l'écart-type de la gaussienne utilisée par LIME pour pondérer les voisins.

Que se passe-t-il si σ est très grand? Très petit?

σ très grand : tous les voisins ont un poids similaire → on explique un comportement plus global (moins fidèle localement)

σ très petit : seuls les voisins très proches comptent → risque de sur-apprentissage, moins stable

Que signifie un R2 < 0.7?Que faut-il faire?

Le R^2 mesure la fidélité du modèle linéaire local g par rapport au modèle complexe

Un $R^2 < 0.7$ signifie qu'on doit être prudent dans l'interprétation des coefficients !

$R^2 = 1.0$: Le modèle linéaire explique parfaitement les prédictions de h localement $R^2 = 0.7$: Le modèle linéaire capture 70% de la variance $R^2 < 0.7$: Mauvaise approximation linéaire

Pourquoi utilise-t-on une régression linéaire et pas un autre modèle?

1. Interprétabilité
2. Principe de LIME : "Local"
3. Stabilité et simplicité

Quelle feature est la plus importante pour cette prédiction

MedInc (revenu médian) est la feature la plus importante avec un coefficient de +0.0907. Interprétation :

Si MedInc augmente de 1 unité, la prédiction augmente de 0.0907 C'est la feature avec l'impact absolu le plus grand : |0.0907| > |0.0584| > |0.0502|...

Le R2 est-il satisfaisant (> 0.7)?

Les valeurs changent à chaque exécution à cause de l'aléatoire dans LIME. Le R^2 est-il satisfaisant (> 0.7) ? Réponse : NON, $R^2 = 0.5311 < 0.7$ Ce que cela signifie :

Le modèle linéaire local ne capture que 50% de la variance des prédictions du modèle complexe L'approximation linéaire est médiocre dans ce voisinage Les coefficients peuvent être imprécis ou instables

Comment σ influence-t-il les explications?

σ petit (0.25) : Explication très locale, seuls les voisins immédiats comptent. Plus précis mais instable.

σ grand (2.0) : Explication plus globale, les voisins éloignés gardent du poids. Plus stable mais moins fidèle au comportement local.

Quel σ donne le meilleur R^2 ?

σ = 0.25 (car le modèle complexe nécessite une approximation très locale. Les faibles R^2 avec σ grand indiquent que cette région de l'espace n'est pas bien approximable par un modèle linéaire global.)

Comparaison avec LIME officiel

Notre R^2 trop élevé obtenu avec un noyau très étroit peut indiquer un sur-apprentissage local (overfitting). Notre explication est très précise pour les points extrêmement proches de l'instance, mais elle peut être instable (les coefficients changent radicalement si l'instance bouge un peu) et moins généralisable au voisinage plus large.

Dans LIME officiel R^2= 42%, score de **fidélité locale** de l'explication est faible. Le modèle linéaire local de LIME ne parvient à expliquer qu'environ **46.52\%** des variations des prédictions générées par votre modèle Random forest dans le voisinage de l'instance expliquée.et suggère que le comportement du modèle de forêt aléatoire est très **non-linéaire** autour de cette instance.

Après avoir modifier les paramètres kernel width et nombre de voisin plus grand on a reussi a baissé le r^2 de notre model .

Conclusion

Ce projet nous a permis de :

1. **Implémenter** LIME de manière fonctionnelle et fidèle à l'article original
2. **Comprendre** profondément les mécanismes d'explicabilité locale
3. **Analyser** l'impact des hyperparamètres (σ, n_samples) sur la qualité
4. **Identifier** les limites pratiques de LIME (R^2 faible, instabilité)
5. **Comparer** avec l'implémentation officielle

Limites observées:

1. **Instabilité aléatoire :** Les perturbations gaussiennes introduisent du bruit
2. **Choix du noyau :** Le noyau RBF est un choix arbitraire
3. **Approximation linéaire:** Échoue si le modèle est très non-linéaire localement, R^2 faible = signal d'alarme
4. **Dépendance aux hyperparamètres:**σ, n_samples doivent être ajustés manuellement.

Ce projet nous a permis de développer une compréhension profonde de l'explicabilité en machine learning, compétence devenue essentielle dans un contexte où la confiance et la transparence des systèmes IA sont cruciales.