

Rétro-ingénierie, maintenance et évol.

Sujet 1 – ML Flow – Gestion des expériences

Léo KITABDJIAN
Thomas FARINEAU
Ludovic BAILET
Mohamed MAHJOUB

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Question Générale

Comment MLFlow permet de gérer la traçabilité ainsi que la reproductibilité des expériences?

“Reproducibility in machine learning involves duplicating the ML procedure described in a paper or tutorial and achieving the same outcomes as the original creation.”

Deepchecks

Sous-questions qui en découlent

- Quelles sont les informations des expériences tracées par MLFlow ?
- Comment sont-elles stockées ?
- A quel point MLFlow garantit-il une reproductibilité des expériences ?

Moyens envisagés pour y répondre

Utilisation de la documentation

Documentation MLFlow détaille le suivi des informations d'expérience.

Explique méthodes de stockage des données par MLFlow.

Assure compréhension de la reproductibilité garantie par MLFlow.

Propose un tutoriel à reproduire pour expérimenter et avoir des métriques à comparer

Identifie divers cas d'utilisation pratiques de MLFlow.

<https://mlflow.org/docs/latest/index.html>

Utilisation de projets open-source sur GitHub

Projets GitHub offrant des exemples concrets d'intégration MLFlow.

Code source disponible qui facilite la compréhension pratique de MLFlow.

Démos GitHub démontrent l'usage de MLFlow en situation réelle.

Contributions communautaires enrichissent connaissances sur MLFlow.

<https://github.com/topics/mlflow-projects>

Limites?

Documentation MLFlow détaillée, mais complexe, peut être compliquée à comprendre.

- Suppose un niveau de compétence avancé en IA, posant des défis aux utilisateurs (nous) moins expérimentés dans ce domaine.

Projets GitHub avec une qualité souvent inégale et une documentation de leur fonctionnement très peu ou pas existante, nécessitant une sélection assez dure, car les projets ne sont pas égaux sur ça.

- Complexité des projets GitHub pouvant affecter négativement la qualité et la fiabilité des productions dérivées.
- Accès impossible au stockage des expériences sur les repos existants

Validation de nos résultats

- Assurer la cohérence des résultats sur plusieurs exécutions, malgré l'aspect hasardeux de l'IA.
 - Reproduire les résultats du tutoriel en utilisant les paramètres enregistrés par MLFlow, issu de la documentation ou d'un projet Git existant avec des résultats publics.

Sur les cas concrets d'utilisation :

- Réussir à faire fonctionner un projet GitHub utilisant MLFlow.
- Reproduire les expériences des projets qui indiquent leurs résultats dans leur fichier README.

Où en sommes nous?

- Analyse du code de projets GitHub disponibles pour essayer de mettre en œuvre ce que la documentation indique en termes de démarche à suivre pour la mise en place de MLFlow
 - <https://github.com/at0m-b0mb/MLflow-TensorFlow-Image-Classification-Guide>
 - https://github.com/VineetKT/genre_classification
 - <https://github.com/nasserboan/mlflow-pydata-talk>
- Lecture approfondie de la documentation pour répondre à la plupart de nos sous-questions:
 - Données stockées lors des expériences
 - Emplacement de ces données
 - Reproduire des résultats d'expériences
 - Traçabilité

Merci de votre attention

A dark blue diagonal gradient bar that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.