Réalisation et analyse des tests

Sébastien AMALLA

Contexte des tests en machine learning

La complexité structurelle

- Un modèle de ML est composé d'objets mathématiques potentiellement très complexes
- Grand nombre de paramètres rendant l'analyse manuelle quasi impossible
- Interactions non-linéaires entre les variables d'entrées
- Grande dépendance au fine tuning des hyperparamètres
- Difficulté à anticiper le comportement des modèles pour de nouvelles variables

Effet "boîte noire"

- Difficulté à comprendre les mécanismes internes de décision
- Manque de transparence dans les processus internes
- Difficulté de compréhension pour les non-experts
- Limitation de la traçabilité des décisions automatisées
- Obstacle à l'audit et à la justification des résultats / des décisions

Online learning

- Le modèle apprend au fil de l'arrivée des données
- Chaque nouvelle data peut mettre à jour le modèle
- Le comportement du modèle change avec le temps
- Évolution en temps réel possible, réactivité élevée
- Risque de dérive non détectée
- Surveillance continue obligatoire

Bach learning

- Modèle entraîné sur un jeu de données fixes, préparées à l'avance
- Modèle figé une fois entraîné
- Pour intégrer de nouvelles données, il faut le réentraîner entièrement
- Apprentissage hors-ligne, lent et coûteux
- Meilleur contrôle de la qualité des données, de la robustesse, du déploiement
- Moins de risque de dérives automatiques
- Incapable de s'adapter en temps réel

Rôle des tests pour le online learning

- Surveillance permanente en continue
  - Performances, baisses de qualité
  - Comparaison à un modèle de référence (baseline)
  - Identifier l'impact des nouvelles datas sur la prédiction
- Surveillance des datas :
  - Changements dans la distribution statistique
  - Données corrompues
  - Biais récents
- Sécurité face aux attaques : tentatives d'empoisonnement des données
- Traçabilité des évolutions
- Tester tout le système, pas uniquement le modèle (pipelines, ...)

Rôle des tests pour le batch learning

- Garantir la performance du modèle avant le déploiement
  - Précision, rappel, F1-score, toute autre métrique
- Comparer plusieurs modèles ou configurations
- Vérifier la capacité de généralisation
- Tester de manière approfondie :
  - Données incomplètes, sensibilité aux variations dans les datas
  - Mesurer la stabilité de la prédiction
  - Vérifier la résistance face aux données atypiques
- Valider la qualité et cohérence des données
- Assurer la reproductibilité et la traçabilité

Quand interviennent les tests

- Tout au long du cycle de vie de l'IA
- Phase de conception : Faisabilité, analyse de risque, définition des KPI, validation des hypothèses métier sur les données
- Collecte des données : Qualité (complétude, exactitude), détection des biais statistiques, représentativité, pipelines de nettoyage et de transformation
- Phase d'entraînement : Détection de l'overfitting, validation croisée, tests statistiques
- Phase d'évaluation : Tests sur les jeu de test, robustesse, stress, sécurité, biais, équité, conformité réglementaire
- Déploiement : Test d'intégration, performance, rollback, sécurité, surveillance en continu

Limites des tests traditionnels

- Inadéquations des tests logiciels classiques pour de l'IA
- Les TU supposent un comportement déterministe
- Absence de résultats attendus clairement définis
- Faible couverture fonctionnelle des comportements réels du modèle
- Scénarios réels difficiles à anticiper
- Cas limites rarement observés dans les données
- La qualité d'un résultat est souvent subjectif
- Diagnostique de la cause d'une erreur parfois impossible

Nouvelles approches pour les tests IA

- Approche basée sur les data (data-centric)
  - Qualité, représentativité, biais, pipelines
- Performances
  - Set de test, cross-validation, métriques, généralisation
- Robustesse
  - Données bruités, incomplètes, extrêmes, aberrantes, stabilité des prédictions
- Recours aux tests statistiques et probabilistes
  - Tests statistiques de comparaison de modèle, validation par intervalles de confiance, tests de significativité des performances
  - Basés sur des seuils probabilistes

Nouvelles approches pour les tests IA

- Approche par explicabilité (XAI-driven testing)
  - Analyse de l'influence des variables, dépendances illégitimes, cohérence,
- Approche de sécurité
  - Empoisonnement des données, fuites d'information, défenses
- Approche éthique et réglementaire
  - Transparence, traçabilité, RGPD, AI Act)
- Approches continues
  - Monitoring des performances en temps réel
  - Concept drift : Nouvelles tendances dans les datas, évolution du contexte des datas, saisonnalité
  - Shadow testing : entraînement d'un deuxième modèle "dans l'ombre", reçoit les mêmes entrées que le modèle en prod, sert de comparaison

Qualité des données

- Valeurs manquantes, aberrantes, doublons, ...
- Biais liés aux données
- Risque de sous-représentation de certaines populations
- Nécessiter de tester plusieurs jeux de données
- Validation des données en amont des tests modèles
- Réflexion sur les différents contextes d'utilisation

Tests de sensibilité au bruit

- Ajouter un léger bruit gaussien sur une variable numérique (<5%)
- Quantifier l'écart de prédiction du modèle entre avant et après l'ajout de ce bruit
  - Objectif : vérifier que la prédiction ne change pas drastiquement pour de petites variations dans les données d'entrée
- Recommencer pour chaque feature
- Documenter l'influence du bruit sur chaque feature
- Tester l'ajout de bruit sur différents ensembles de features, puis sur le dataset global
- Quantifier la sensibilité du modèle au bruit

TP

- Faire une prédiction, puis bruiter une feature numérique
  - Appliquer successivement 1%, 3%, 5%, 10% et 20% de son écart-type en bruit
- Calculez la métrique de performance pour chaque niveau (RMSE, F1-score, ...)
- Tracer la courbe de l'évolution de la MSE par rapport au pourcentage bruit : "Noise sensitivity curve"
- Recommencer pour les trois features les plus corrélées