Mini Projet **Deep Learning**
décembre 2020 **Durée: jusqu'au 01/02/2021 à 23h59**

Objectif: Modèle profond pour la classification de données avec classes déséquilibrées.

- Le projet doit être réalisé en binôme. Merci de bien préciser les nom et prénom de chaque membre du binôme et de mettre en copie votre binôme lors de l'envoi des scripts et de votre rapport.
- Les résultats de vos analyses doivent être commentés et reprendre les différents notions vues en cours.
- Vous devez me faire parvenir votre projet (scripts et rapport séparés) pour le 01/02/2021 à 23h59 (*severine.affeldt@u-paris.fr*).

---

1. **Tutoriel en ligne**

   Pour réaliser ce mini-projet, vous devez dans un premier temps suivre le tutoriel *Classification on imbalanced data* disponible sur le site de `tensorflow`.

   Prenez le temps:
   - de bien connaître les données
   - de bien comprendre les différentes métriques d'évaluation
   - d'essayer d'autres modèles deep que le modèle *baseline* proposé
   - de comprendre l'intérêt du critère AUC

   Afin de réduire le problème du déséquilibre des classes, ce tutoriel propose une approche d'*oversampling*, qui consiste à augmenter le nombre d'instances de la classe minoritaire. Familiarisez-vous avec cette approche, et également avec d'autres approches d'*oversampling* disponibles aux liens ci-dessous. Pour remédier au déséquilibre, il est également possible de faire de l'*undersampling*, c'est-à-dire de réduire le nombre d'instances de la classe majoritaire.

   - Resampling strategies for imbalanced datasets
   - Undersampling and oversampling imbalanced data
   - Techniques to deal with imbalanced data

2. **Données pour ce mini-projet**

   Pour ce mini-projet, vous analyserez les données suivantes:
   - Credit fraud
   - Bank marketing
   - Employee attrition

3. **Réalisation attendue pour chaque jeu de données**

   Pour chacun des jeux de données ci-dessus, vous devrez réaliser les analyses suivantes. Appuyez-vous sur le tutoriel de *tensorflow* et aidez-vous des exemples d'utilisation des approches de sampling proposés aux points précédents.

1. Préparation des données
   (a) Explorez les données et proposez les pré-traitements adéquats en les justifiant. Donnez un aperçu général clair des données pré-traitées.
   (b) Préparez les données en trois jeux: entraînement, validation et test.

2. Entraînement *simple*
   (a) Entraînez un modèle deep *baseline* et faites son évaluation.
   (b) Proposez un modèle deep plus élaboré et faites son évaluation.
   (c) Résumez clairement vos évaluations.

3. Entraînement avec pondération des classes
   (a) Calculez le poids des classes.
   (b) Entraînez votre modèle deep baseline et faites son évaluation.
   (c) Entraînez votre modèle deep plus élaboré (modèle que vous avez précédemment proposé) et faites son évaluation.
   (d) Résumez clairement vos évaluations.

4. Entraînement avec oversampling
   (a) Utilisez différents approches d'oversampling pour mieux équilibrer vos données.
      - Random oversampling
      - SMOTE (Synthetic Minority Oversampling Technique)
      - ADASYN (Adaptive Synthetic)
   (b) Entraînez votre modèle deep *baseline* et faites son évaluation.
   (c) Entraînez votre modèle deep plus élaboré (modèle que vous avez précédemment proposé) et faites son évaluation.
   (d) Résumez clairement vos évaluations.

5. Entraînement avec undersampling
   (a) Utilisez différents approches d'undersampling pour modifier le jeu de données.
      - Random undersampling
      - Tomek links
   (b) Entraînez votre modèle deep *baseline* et faites son évaluation.
   (c) Entraînez votre modèle deep plus élaboré (modèle que vous avez précédemment proposé) et faites son évaluation.
   (d) Résumez clairement vos évaluations.

6. Entraînement avec une approche hybrid de resampling
   (a) Utilisez différentes approches hybrides pour modifier le jeu de données.
      - Random oversampling + Tomek links
      - SMOTE + Tomek links
   (b) Entraînez votre modèle deep *baseline* et faites son évaluation.
   (c) Entraînez votre modèle deep plus élaboré (modèle que vous avez précédemment proposé) et faites son évaluation.
   (d) Résumez clairement vos évaluations.

7. Conclusions générales
   - Comparez et résumez clairement les évaluations de vos modèles (baseline & élaboré) sans poids ni sampling, avec pndération et pour les différentes approches de sampling.