HPC-AI : Introduction au Machine Learning **Projet numérique**

Date de rendu : 12 février 2021

Dans ce projet, vous allez prendre part à une compétition Kaggle In Class. Il s'agit de prédire combien de fois un article en ligne est partagé, en utilisant des variables qui décrivent cet article (mots clés, sujet, longueur, etc.)

Kaggle [1] est une plateforme dédiée à la science des données, qui propose de nombreuses compétitions de prédiction.

Pour accéder à la compétition, cliquez l'URL suivante :

`https://www.kaggle.com/t/ca3872e42ebd4ee1bd8bd3ce226060bd`

Vous trouverez sur la page de la compétition les données, une description détaillée des variables, et une explication de la métrique d'évaluation.

Pour ce cours, vous allez participer à la compétition seul ou seule. Vous pouvez par contre bien sûr échanger entre vous au cours du projet.

Le projet sera noté sur un rapport de deux pages (sans compter figures et tableaux), ainsi que sur la qualité de vos soumissions finale au leaderboard. Le tout est dû le **vendredi 12 février 2020 à 23h59**.

Vous pouvez utiliser un notebook Jupyter mais votre code ne sera pas pris en compte dans votre note.

Comment rendre votre rapport ? Envoyez votre rapport,

- sous la forme d'un unique **fichier PDF**
- de maximum deux pages (sans compter figures et tableaux)
- par email à `chloe-agathe.azencott@mines-paristech.fr`
- **avec pour objet** le texte : `[HPC-AI] Rapport Machine Learning`.

Votre rapport peut être **en français ou en anglais**.

Contenu du rapport

1. Prétraitement [2 pts]

Commencez par visualiser vos variables (par exemple sous forme d'histogrammes). Pensez-vous avoir besoin de les centrer-réduire ? De leur appliquer une transformation type passage au log ? Souhaitez-vous toutes les conserver ?

Décrivez dans votre rapport les choix que vous avez fait en les justifiant.

1. `https://www.kaggle.com/about`

2. Sélection de modèle [6 pts]

Mettez en place une validation croisée sur les données pour sélectionner le(s) meilleur(s) hyperparamètre(s) :

- d'une régression logistique avec régularisation ridge ;
- d'au moins deux autres approches de classification non-linéaires de votre choix (SVM, forêts aléatoires, etc.).

Dans votre rapport, incluez :

- Le code utilisé pour sélectionner le(s) meilleur(s) hyperparamètre(s) de la régression logistique avec régularisation ridge ;
- La liste des approches que vous avez évaluées, y compris les valeurs de leurs hyperparamètres ;
- pour la ou les métriques de performance de votre choix, la performance en validation croisée :
  - d'une régression logistique non régularisée ;
  - d'une régression logistique avec régularisation ridge ;
  - d'au moins deux approches non-linéaires.

  Vous pouvez utiliser une table ou une figure.

3. Réduction de dimension [6 pts]

Choisissez au moins une méthode de réduction de dimension et reprenez la section 2 mais en utilisant cette fois la version réduite des données. Comparez vos résultats à ceux obtenus à la section 2.

4. Modèle final [2 pts]

Utilisez les résultats obtenus aux sections 2 et 3 pour déterminer les deux modèles que vous souhaitez utiliser afin d'obtenir une performance optimale. Justifiez votre choix dans le rapport.

Entraînez ces modèles sur l'intégralité de vos données étiquetées. Soumettez les prédictions faites par ces modèles au leaderboard.