ULaval **IFT&GLO**

Mini-Projet #3 (sur 20%)
à rendre le 19 août 2020 à 23h55mn

Professeur Brahim Chaib-draa *GLO-7050 : Apprentissage machine en pratique*

Les projets du cours sont les projets des versions du cours COMP 551, cours donné par les collègues de McGill, un grand merci à eux pour nous avoir donné l'autorisation de les utiliser.

Préambule

- **Ce mini-projet peut se faire en équipe de deux**. Toutefois, ceci nous vous empêche pas de discuter avec les autres étudiants qui suivent le cours. En aucun cas cependant vous ne devez reprendre le code et l'écrit d'autrui ; il vous est demandé d'élaborer les vôtres.
- Comme pour le Mini-Projet-2, vous allez soumettre votre travail sur MonPortail, et également à une nouvelle compétition Kaggle. Vous devez former des équipes d'au plus deux étudiants sur MonPortail ensuite former la même équipe sur Kaggle. L'inscription à la compétition Kaggle se fait au niveau de :
  `https://www.kaggle.com/t/409d0452000541599e3974a5660bcf33`
- Si vous "empruntez" des idées, méthodes, démarches ou autres, merci d'indiquer vos sources dans le rapport.
- Il vous est fortement suggéré d'argumenter et/ou justifier vos réponses.
- Après la date de remise, vous avez jusqu'à une semaine pour remettre votre travail avec une pénalité de 20%. Au delà le travail vaut 0.
- Vous êtes libres d'utiliser les librairies de votre choix, telles que PyTorch, Tensorflow, Keras, FastAI, ... etc.
- Si vous avez des questions concernant le travail, merci de passer par le Forum, en posant clairement vos questions.
- **Seules les données fournies pour la compétition (MNIST et MNIST Modifié) peuvent être utilisées pour l'entrainement des modèles supervisés. Vous ne devez pas utiliser d'autres données externes. Par contre, l'augmentation des données est permise, les modèles pré-entrainés sur ImageNet sont aussi autorisés.**

1 Objectif du mini-projet

Dans ce mini-projet, l'objectif est de participer à un challenge de prédiction par analyse d'image. La tâche est basée sur l'ensemble de données de MNIST (https ://en.wikipedia.org/wiki/MNIST_-database). Le MNIST original contient des chiffres manuscrits de 0 à 9 et l'objectif est de déterminer quel chiffre est présent dans une image. Pour ce projet, vous allez travailler avec un ensemble de données MNIST modifié. Dans lequel, les images contiennent trois chiffres et l'objectif est de sortir le chiffre ayant la valeur numérique la plus élevée. Chaque exemple est représenté par une matrice

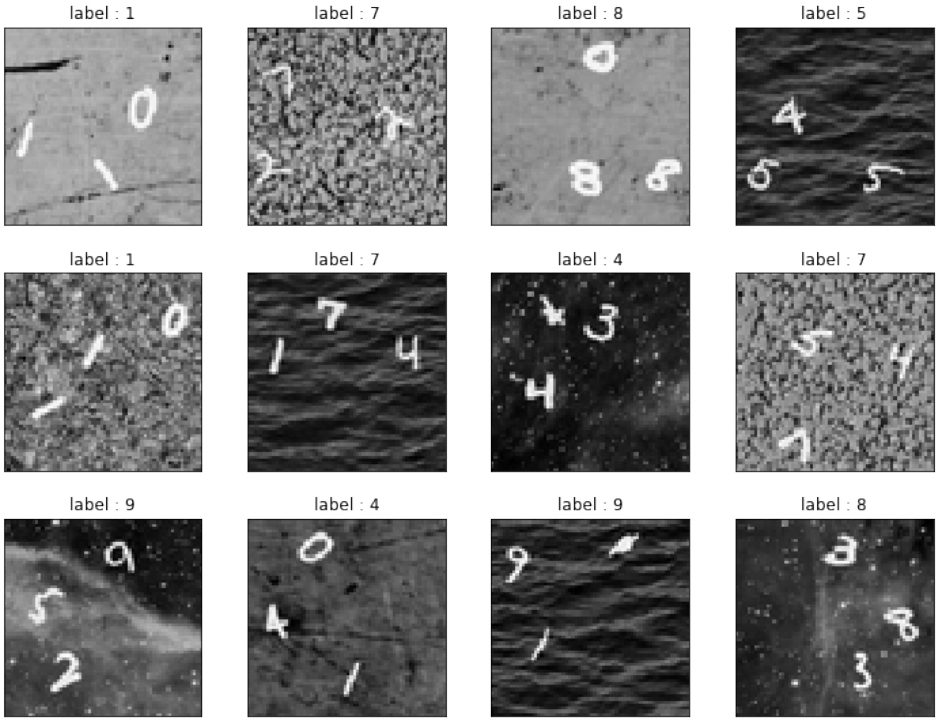

FIGURE 1 – Exemple d'images du dataset MNIST modifié. Par exemple, l'étiquette pour l'image en haut à gauche est 1, tandis que celle pour l'image en bas à droite est 8.

2D d'intensité de pixels (c'est-à-dire que les images sont en niveaux de gris et non en couleurs). Des exemples sont illustrés dans la figure 1. Remarquez qu'il s'agit d'une tâche de classification supervisée : chaque image a une étiquette associée (c'est-à-dire le chiffre de l'image avec la valeur numérique la plus élevée) et votre objectif est de prédire cette étiquette.

2 Tâches à accomplir

Vous devez concevoir et valider un modèle de classification supervisée pour effectuer la prédiction sur MNIST modifié. Il n'y a aucune restriction sur votre modèle, sauf qu'il doit être écrit en Python 3.x. Comme pour les mini-projets précédents, vous devez rédiger un rapport détaillant votre approche, vous devez donc développer un pipeline de validation complet et idéalement fournir une justification / motivation pour vos decisions de conception. **Vous êtes libre de développer un modèle unique ou d'utiliser un ensemble de modèles ; il n'y a pas de restrictions strictes à ce niveau**.

3 Déliverables

3.1 Code et compte-rendu (rapport.pdf)

Vous devez soumettre deux fichiers distincts à Monportail (en utilisant les noms de fichiers et leur type comme décrit ci-dessous) :

1. **code.zip** : collection de fichiers .py, .ipynb et autres fichiers de support pour le code. **Il doit être possible pour les assistants du cours de reproduire tous les résultats reflétés dans votre rapport ainsi que votre meilleure soumission Kaggle.** Vous devez soumettre un README détaillant les packages utilisés ainsi que les instructions pour exécuter le code.
2. **rapport.pdf** : Compte-rendu du projet (en pdf) ne devant pas excéder 5 pages et dont les détails sont donnés ci-dessous.

3.2 Le compte-rendu (rapport.pdf)

Chaque équipe doit soumettre un compte rendu du projet ne dépassant pas cinq pages (simple interligne, 10 pt police ou plus grand ; des pages supplémentaires peuvent être utilisées pour les références / le contenu bibliographique et les annexes).

Il est fortement recommandé d'utiliser LaTeX pour la rédaction, et la fonction bibtex pour les citations. Pour la mise en forme il vous est suggéré de suivre le style (`https://nips.cc/Conferences/2019/PaperInformation/StyleFiles`).

Vous êtes libre de structurer le rapport comme bon vous semble ;

Ci-dessous sont des lignes directrices et des recommandations générales, mais il ne s'agit que d'**une suggestion de structure**

Résumé (100-250 mots) Résumez la tâche du projet et vos principales conclusions.

Introduction (5+ phrases) Résumez la tâche du projet, l'ensemble de données et votre plus important résultats. Cela devrait être similaire au résumé mais plus détaillé.

Travaux connexes (4+ phrases) Résumez la littérature pertinente des années écoulées.

Ensemble de données et configuration (3+ phrases) Décrire très brièvement l'ensemble de données (et ce qui est en lien) ainsi que les méthodes de prétraitement. Remarque : vous n'avez pas besoin de vérifier explicitement que les données satisfont aux exigences i.i.d. (ou l'une des autres hypothèses formelles de classification linéaire).

Approche proposée (7+ phrases) Décrivez brièvement votre modèle (ou les différents modèles développés, s'il y en a plusieurs), en fournissant des citations si nécessaire. Si vous utilisez ou construisez sur un modèle existant basé sur des travaux publiés précédemment, il est essentiel que vous citiez ce travail. Discutez le choix et l'implémentation de vos modèles. Incluez toute décision concernant la répartition train/validation, les stratégies de régularisation, toute astuce d'optimisation, le choix d'hyperparamètres, etc. Il n'est pas nécessaire de fournir le détail des modèles que vous utilisez, mais vous devez fournir, pour chaque modèle, au moins quelques phrases concernant le background et ce qui vous a motivé à le choisir.

Résultats (7+ phrases, éventuellement avec des figures ou des tableaux) Fournissez les résultats de votre approche (par exemple, précision sur l'ensemble de validation, temps d'exécution). Vous devez indiquer dans cette section la précision sur l'ensemble de test de votre meilleur modèle (le score que vous avez eu sur Kaggle), mais la plupart de vos résultats doivent être sur votre ensemble de validation (ou de validation croisée).

Discussion et conclusion (3 phrases et plus) Résumez les principales conclusions que vous tirez d'un tel projet et éventuellement des orientations pour des travaux futurs.

État des contributions (1 à 3 phrases) Indiquez la répartition de la charge de travail.

Evaluation

Comme pour le mini-projet-2, ce mini-projet est noté sur 100 points, et la répartition des points est la suivante :

- **Performances (50 points)**
  - Votre note sera calculée selon votre performance sur un ensemble de test. Le calcul de la note est une interpolation linéaire entre la performance d'un modèle aléatoire, d'une référence TA (score obtenu par l'assistant) et du deuxième meilleur score de la compétition. Les trois premières équipes de la compétition reçoivent toutes la note complète (50 points). - Ainsi, si par exemple, X indique votre précision sur l'ensemble de test, R indique la précision du modèle aléatoire, B indique la précision du deuxième meilleur groupe, et T indique le score de l'assistant (TA Baseline), votre score sera calculé comme suit :

$$\text{points} = 50 * \begin{cases} 0 & \text{if } X < R \\ \frac{X-R}{T-R} * 0.75 & \text{if } X > R \text{ and } X \leq T \\ \frac{X-T}{B-T} * 0.25 + 0.75 & \text{if } X > T \text{ and } X \leq B \\ 1 & \text{if } X > B \end{cases}$$

    L'équation peut paraître compliquée, mais l'idée de base est la suivante :
    - La référence B représente le score nécessaire pour obtenir plus de 0% sur la compétition, la référence de l'assistant T représente le score nécessaire pour obtenir au moins 75% sur la compétition, et B le score de la 2ème meilleure équipe représente le score nécessaire pour obtenir 100%.
    - Si votre score se situe entre R et T, alors votre note est une interpolation linéaire entre 0% et 75%.
    - Si votre score se situe entre T et le score du 2ème meilleur groupe B, alors votre note est une interpolation linéaire entre 75% et 100% des points de la compétition.
  - **De plus, l'équipe avec le meuilleur score recevra un bonus de 5 points, l'équipe avec le 2ème meuilleur score recevra un bonus de 3 points et l'équipe avec le 3ème meuilleur score recevra un bonus de 2 points**
- **Qualité de la rédaction et de la méthodologie proposée (50 points)**

Comme pour les mini-projets précédents, votre projet sera jugé en fonction de sa qualité scientifique :

- Rapport sur toutes les expérimentations ?
- La méthodologie que vous proposez est-elle techniquement fondée ?
- À quel point vos expériences sont-elles détaillées/rigoureuses/approfondies ?
- Votre rapport décrit-il clairement la tâche sur laquelle vous travaillez, le dispositif expérimental, les résultats et les figures (par exemple, n'oubliez pas les étiquettes des axes et les légendes des figures, n'oubliez pas d'expliquer les figures dans le texte).
- Le rapport est-il bien organisé et cohérent ?
- Le rapport est-il clair et exempt d'erreurs grammaticales et de fautes de frappe ?
- Votre rapport comprend-il une discussion adéquate des travaux et des citations ?

Qualité du compte rendu et méthodologie proposée (50 points). Comme pour les mini-projets précédents votre rédaction sera jugée en fonction de sa qualité scientifique en rapport avec les questions suivantes (incluses mais non limitée à) :

- Avez-vous rapporté sur toutes les expériences et comparaisons requises ?
- La méthodologie que vous proposez est-elle techniquement valable ?
- Dans quelle mesure vos expériences sont-elles détaillées / rigoureuses / étendues ?
- Votre rapport décrit-il clairement la tâche sur laquelle vous avez travaillé, la configuration expérimentale, les résultats et les figures (par exemple, n'oubliez pas les axes et les légendes sur les figures, n'oubliez pas d'expliquer les chiffres dans le texte).
- Votre rapport est-il bien organisé et cohérent ?
- Votre rapport est-il clair et exempt d'erreurs grammaticales et de fautes de frappe ?
- Votre rapport comprend-il une discussion adéquate des travaux et citations connexes ?

4 Remarques finales

Vous devez faire preuve d'initiative, de créativité, de rigueur scientifique, de pensée critique, et avoir de bonnes compétences en communication. Vous n'êtes pas obligé de vous limiter aux exigences énumérées ci-dessus - n'hésitez pas à y aller au-delà et explorer plus loin.

Vous pouvez discuter des méthodes et des problèmes techniques avec des membres d'autres équipes, mais vous ne pouvez pas partager de code ou de données avec d'autres équipes. Toute équipe ayant triché (par exemple, utiliser des données externes à la compétition, utiliser des ressources sans références appropriées) sur le code, les prédictions ou le rapport écrit recevra un score de 0 pour toutes les composantes du projet.

Règles spécifiques à la compétition Kaggle :

- Ne pas tricher ! Vous devez soumettre un code qui permet de reproduire le score de votre solution soumise sur Kaggle.
- Les données de la compétition sont basées sur un ensemble de donnée public. Vous ne devez pas tenter de tricher en cherchant des informations sur l'ensemble de test. Les soumissions dont la précision et/ou les prédictions sont douteuses seront signalées et la note 0 sera adressée.
- Ne faites pas plus d'une équipe pour votre groupe (par exemple, pour faire plus de soumissions). Si tel est le cas, une note de 0 vous sera donnée pour la création intentionnelle de nouveaux groupes dans le but de faire plus soumissions Kaggle.