# IFT3395/6390 Projet de session

Professeur: Aaron Courville Aux: Philippe Lacaille, Tegan Maharaj

31 octobre 2017

## 1 Résumé du projet

Le but de ce projet est de vous faire dévélopper des comptences et intuitions de recherche en comparant la performance et les caractèristiques de differents algorithmes d'apprentissage automatique sur differents ensembles de données. Vous devrez vous trouver une équipe, soumettre une proposition de projet, rédiger un article décrivant vos resultats, faire un poster, et présenter vos résultats à vos pairs.

## 2 Instructions

- Formez une équipe de 2 à 3 personnes (le forum de Studium peut faciliter cela).
- Choisissez 3+ algorithmes (classifieurs, regression) et 2+ ensembles de données. Quelques algorithmes et ensembles de données vous sont suggerés dans la section 3, mais vous êtes libres d'en choisir d'autres. On vous encourage à choisir des données et / ou des méthodes qui vous intéressent et que vous êtes motivé à explorer.
- Soumettez une proposition de projet d'une page, **à remettre dans deux semaines (le 14 novembre)**. La proposition doit contenir:
  1. Un titre (de votre choix), noms des membres d'équipe, code de cours, date
  2. Un court résumé qui explique la motivation de vos choix, les phénomènes et problèmes qui vous intéressent
  3. La liste des algorithmes choisis, avec une brève description de chacun
  4. La liste des bases de données choisis, avec une brève description de chacune
- Rédiger un article de 6 pages dans le format NIPS **à remettre à la fin de la session** (DATE À DETERMINER)
  1. Décrire votre approche (ensembles de données et algorithmes choisis), avec votre motivation, de facon similaire à la proposition de projet.
  2. Évaluez chaque algorithme sur chaque ensemble de données, i.e. au moins 6 expériences. On vous encourage fortement d'en faire plus, e.g. pour adjuster des hyper-paramètres, essayer différentes techniques de régularisation, etc. N'oubliez pas d'inclure toutes vos expériences dans votre rapport (en annexe au besoin).
  3. Affichez vos résultats en terme de taux d'erreur (dans un tableau) ainsi que visuellement (courbe d'apprentissage, frontières de décisions ou toute autre figure afin d'expliquer vos résultats).
  4. Comparez la performance et les caractéristiques de chacun des algorithmes choisis sur les ensembles de données. Faites des recommandations afin de sélectionner des algorithmes, et notez vos observations sur vos ensembles de données. Référez vous à la section 4 pour des suggestions et questions de recherche.
  5. Incluez une section "Répartition et remerciments", avec une description des contributions de chaque membre de l'équipe, et remerciments s'il y a lieu.
- Préparer un poster (format A0), résumant vos résultats, ainsi que vos observations et recommendations. **Vous présenterez ce poster à vos pairs à la fin de la session** (DATE À DETERMINER).

# 3 Algorithmes et bases de données suggérés

## 3.1 Algorithmes

- Bayes Naïf (estimateur de densité de votre choix)
- Forêt d'arbres décisionnels
- Machine à vecteurs de supports
- Perceptron multicouche
- Réseau convolutionnel

## 3.2 Bases de données

- MNIST (http://yann.lecun.com/exdb/mnist/)
- CIFAR-10 (https://www.cs.toronto.edu/~kriz/cifar.html)
- Télémarketing bancaire (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)
- Qualité des vins (http://www3.dsi.uminho.pt/pcortez/wine/)
- Prédiction du salaire (http://archive.ics.uci.edu/ml/datasets/Adult)
- Analyse des sentiments (http://ai.stanford.edu/~amaas/data/sentiment/)

# 4 Suggestions et questions de recherche

À titre d'exemple/inspiration, voici plusieurs questions que vous pouvez explorer dans votre article.

- Théoriquement, lequel de vos algorithmes est plus propice au sur-apprentissage? Observez-vous ces resultats?
- Quelles sont les hypothèses implicitement reflétées dans vos algorithmes? Lesquelles en sont des "bonnes"?
- Comparez la performance (taux d'erreur, log-vraisemblance, etc.) des différents algorithmes sur chaque base de données. Est-ce qu'un des algorithmes performe systématiquement mieux que les autres? Pourquoi?
- Explorez vos données, faites une analyse statistique de vos données en affichant la moyenne, variance, et ce, potentiellement selon différents traits caractéristiques. Explorez la visualisation des traits selon les classes. Est-ce que certaines de ces propriétés vous aident à expliquer la performance des différents algorithmes? Comparez vos statistiques entre les ensembles de données, est-ce que certaines bases de données sont plus balancées que d'autres? Est-ce que cela est désirable?
- Quelles propriétés influencent la performance des algorithmes choisis? Cette dernière est-elle robuste? Qu'est-ce qui pourrait y nuir?
- Qu'avez-vous appris sur vos données? Selon vos observations, êtes-vous en mesure d'identifier une application dans le monde "réel"? De quelle manière est-ce que la base de données représente une application dans le monde "réel"? De quelle manière cela ne serait pas le cas?
- Que pouvez-vous offrir comme suggestion à quelqu'un qui ferait face à ce problème dans le monde "réel"? (Algorithme proposé, hyper-paramètre, cueillette de données, etc.)
- Faites une analyse qualitative de la performance de vos algorithmes. Montrez des examples où il échoue, d'autres où il est mélangé et réussi bien. Est-ce que les examples "difficiles" et "faciles" d'une base de données sont les mêmes pour tous les algorithmes choisis?
- Suites à vos observations, recommandez des pistes à explorer davantage (dans un projet subséquent bien entendu!).