Science des données
IFT3700 et IFT6758
Travail 2

Révisé le 20 novembre 2018
(remise électronique avant le **21 décembre 23h59)**

Question 1

Cette question est une compétition entre les équipes et sera évaluée en fonction de la précision du classifieur produit par chaque équipe. Les données sont contenues dans le fichier PATCH.amat et il s'agit de 50000 images (28 x 28) en noir et blanc codées en binaire. Les images sont abstraites et appartiennent à deux catégories distinctes. Le fichier contient une image par ligne et chaque ligne commence par 784=28 * 28 bits associés à la couleur des pixels et suivi d'un bit représentant la classe. Le codage du fichier est lisible, mais nécessite un prétraitement pour être utilisé. L'équipe doit produire un classifieur qui sera mis en production et évalué sur des données fraîches non disponibles à l'équipe (mais qui ont exactement la même distribution). Les équipes avec une solution minimalement raisonnable seront classées en ordre de précision et le rang sera transformé en note variant de 10 à 25 sur 25.

Question 2

Imaginez qu'on vous donne un fichier contenant des données de nature astronomique. Il s'agit de données concernant 6500 milliards d'étoiles. Les données (dépassant la centaine de téraoctets) sont réparties sur **300 serveurs** avec un accès rapide au disque et une bonne capacité de calcul. La connexion entre les serveurs est rapide, mais pas exceptionnelle.

Les données pour chaque étoile comporte, la position dans l'espace (x,y,z), la luminosité apparente de l'étoile et sa catégorie (10 catégories possibles). Aussi, pour chaque étoile une liste de ses caractéristiques physiques représentées par 22 nombres réels est incluse.

Proposez une approche distribuée qui permet de répondre aux questions suivantes et expliquez en détail toute la démarche permettant leur résolution.

- Trouvez les 1000 paires d'étoiles jumelles les plus proches (distance euclidienne de la position).
- Comptez combien d'étoiles il y a dans chaque catégorie.
- Produisez un classifieur qui, étant donné le vecteur de caractéristiques (22 nombres réels), prédit la catégorie de l'étoile.

Question 3

Expliquez en détail comment utiliser un classifieur binaire, capable d'apprendre à effectuer la classification de deux catégories, pour réaliser la classification dans un contexte ou plusieurs catégories doivent être distinguées. Considérez le cas à 3, 25, 12500 catégories et faite le contraste entre les différentes approches étudiées et le nombre de catégories.

Question 4

Faites une analyse détaillée et exhaustive des patrons existants dans le jeu de données adulte. Il est possible que le regroupement de valeurs pour certaines caractéristiques donne des résultat intéressant. Vous devez aussi, s'il y a lieux, discuter de l'aspect éthique concernant l'utilisation des patrons obtenus.

Information complémentaire

- En groupe de 3 ou 4
- Le rapport peut être rendu en français ou en anglais.
- Format PDF et Jupyter Notebook
- Originalité
- La qualité des résultats obtenus
- Qualité de la présentation du rapport
- Les étudiants IFT3700 et IFT6758 seront évalués en tant que groupe séparé et la qualité attendue dans le cas gradué (IFT6758) est significativement supérieure.