Le génie pour l'industrie

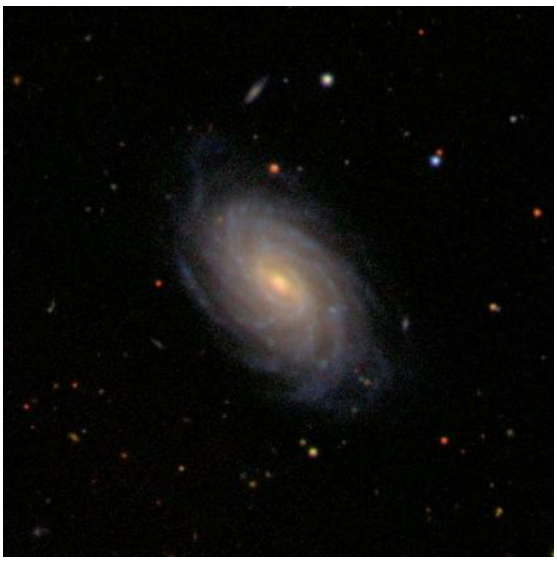
GTI770
Systèmes intelligents et apprentissage machine

TP01 — Définition et extraction de primitives

1. Contexte

Ce premier laboratoire porte sur la définition et l'extraction de primitives sur des fichiers multimédias. Le problème de classification qui vous est présenté est le problème *Galaxy Zoo* (voir l'adresse https://www.galaxyzoo.org/ pour de plus amples informations) dont le but est de classer des images de galaxies dans diverses catégories. En vous basant sur les concepts vus en classe, vous devez définir des primitives que vous jugez pertinentes à extraire sur ces types d'images et effectuer l'extraction de celles-ci sur l'ensemble de données fournies avec cet énoncé.

Veuillez noter que les images qui vous sont fournies ne sont pas nécessairement très faciles à travailler. Plusieurs images comportent du bruit, des artéfacts ou des éléments non pertinents (par exemple, des étoiles lointaines dans l'arrière-plan). Le défi de ce laboratoire repose sur cette difficulté qui est chose courante dans des problèmes d'apprentissage machine moderne. Prenez en considération que certaines images doivent subir certains filtres avant de voir les primitives extraites de celles-ci. Voici, en exemple, deux images de galaxies se retrouvant dans l'ensemble de données :

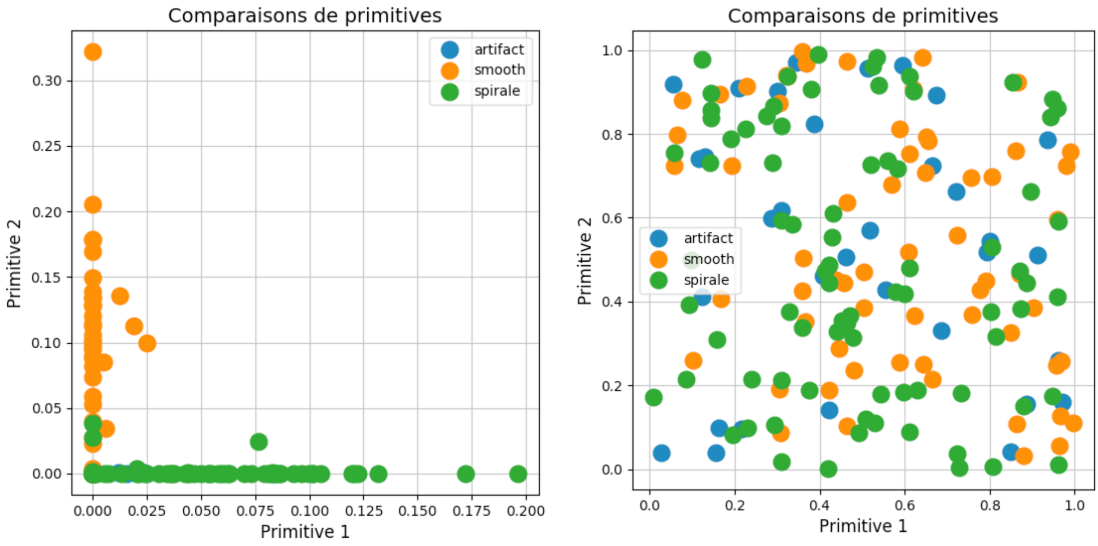


(a)

(b)

Figure 1 : (a) une image d'une galaxie se retrouvant dans la classe « spirale »; (b) une image d'une galaxie se retrouvant dans la classe «smooth».

L'évaluation de ce laboratoire sera basée sur la qualité des primitives proposées, la rédaction de votre rapport ainsi que votre code source produit. Le pouvoir discriminant, c'est-à-dire la capacité des primitives à bien séparer les exemples des classes dans l'espace des primitives, sera également évalué (voir figure 2).



(a) (b)

Figure 2 : (a) primitives x_1 (axe horizontal) et x_2 (axe vertical) très discriminantes; (b) primitives x_1 et x_2 dont le pouvoir discriminantes est faible.

Puisque tous les travaux au courant de cette session se feront avec la technologie *Python3*, vous devez utiliser ce langage de programmation, conjointement avec la librairie de traitement d'images *OpenCV* et la librairie d'apprentissage machine *scikit-learn*, afin de réaliser ce premier laboratoire.

2. Durée du laboratoire

Vous disposez d'un total de trois (3) séances de laboratoire afin d'effectuer ce travail. En fonction de votre groupe, vous disposez des séances suivantes :

Tableau 2.1 : Date des séances et date de remise

Date des séances	GTI770-01	GTI770-02
Séance 1	20 septembre 2018	20 septembre 2018
Séance 2	27 septembre 2018	27 septembre 2018
Séance 3	4 octobre 2018	4 octobre 2018
Remise	10 octobre 2018 - 23h55	10 octobre 2018 - 23h55

3. Objectifs

Les principaux objectifs de ce laboratoire sont :

1. Produire un code source utilisant les technologies *Python3* ainsi que les librairies *OpenCV* et *scikit-learn* permettant d'extraire des primitives;
2. Évaluer le pouvoir discriminant entre les primitives extraites;
3. Programmer un modèle d'apprentissage machine utilisant les arbres de décisions;
4. Se familiariser avec le langage de programmation *Python* et les outils et librairies scientifiques de traitement d'image et d'apprentissage machine.

4. Matériel fourni

Vous disposez du matériel suivant :

- Gabarit du rapport de laboratoire;
- Ensemble de données d'entrainement comportant 31 365 images de galaxies et un fichier CSV contenant l'étiquette associée à chaque image.
- Code source de base permettant de lire les fichiers images dans un dossier et tracer les graphiques;

5. Manipulations

- **Lors de la première séance de travaux pratiques**, vous devez réaliser le TP00 qui consiste en des exercices dans le langage Python. Une fois ce TP initial complété, vous pourrez passer au TP01.
- Téléchargez le code contenant l'ensemble de données à partir de la plateforme *Moodle*.
- Consultez les images et définissez au moins trois (3) types de primitives pouvant être extraites.
- Produisez un code source qui permet d'extraire ces primitives et mettre des valeurs numériques sur celles-ci. Afin de gagner du temps de calcul, pensez à paralléliser votre code avec la librairie Python *multiprocessing* (facultatif).
- Évaluez la pertinence de vos primitives et les performances de classification de celles-ci. Pour ce faire, tracer différents graphiques permettant d'évaluer le pouvoir discriminant des primitives à l'aide de la librairie *matplotlib* de Python.
- À l'aide de la librairie *scikit-learn*, programmez un modèle basé sur un arbre de décision permettant de classifier les galaxies à l'aide des primitives précédemment extraites. Produisez des modèles faisant varier la profondeur de l'arbre de décision (argument *max-depth*). N'oubliez pas de faire calculer un score de précision (*accuracy*) en suivant une des

méthodologies de tests vues en classe (validation croisée, sous-ensembles d'apprentissage/ de tests, etc.)

Tableau 5.1 : Matrice des hyperparamètres - Arbre de décision

Paramètre «*x*»		
max_depth = None	X	
max_depth = 2		X
max_depth = 3		X
max_depth = 4		X
max_depth = 5		X
max_depth = 10		X

- Tracez les frontières de décisions pour chaque paire de primitives que vous disposez avec la méthode *plot_tree_decision_surface()* fournie dans le code du laboratoire.

6. Rapport

Votre rapport devra contenir les réponses aux questions suivantes ainsi que la structure présentée dans un *Jupyter Notebook* en suivant le gabarit du rapport fournit sur la plateforme *Moodle*. Il devra notamment avoir une analyse détaillée de vos primitives ainsi qu'une analyse graphique sur le pouvoir discriminant des primitives. Tout graphique pertinent peut également être ajouté au rapport. N'oubliez pas de **documenter votre code source selon les standards décrits dans le document *Instructions de remise***.

6.1 Questions du rapport

1. Avec les liens fournis en l'annexe de cet énoncé et avec vos trouvailles faites sur Internet par le biais de vos recherches personnelles, faites, à titre d'introduction, une revue de la littérature. Celle-ci doit faire état des recherches ayant été faites en la matière et des possibles pistes que vous pourriez suivre durant ce laboratoire pour effectuer l'extraction de primitives sur les images de galaxies.

2. Expliquer le choix des primitives. Quelle démarche avez-vous suivie afin d'effectuer votre choix de primitives? Sur quelles sources vous êtes-vous basées afin d'établir votre choix de primitives?
3. À l'aide de graphiques générés par votre script, expliquez l'efficacité de deux primitives qui permettent de distinguer le mieux possible les classes du problème.
4. À la suite de votre implémentation de l'arbre de décision, expliquer pour quelles raisons votre arbre de décision donne un tel score de précision. Qu'a fait la variable *max_depth* sur les performances de classification?
5. Quelle autre primitive aurait également pu être ajoutée afin d'améliorer le pouvoir discriminant ou la performance de régression des probabilités?
6. Écrivez une conclusion qui résume le contenu de votre rapport. Dites, en résumé, la manière dont vous avez résolu le problème, quelles ont été vos primitives et les résultats que vous avez obtenus, et glissez un mot sur des améliorations possibles. Dites également, le cas échéant, ce qui a moins bien été durant la réalisation de ce laboratoire.

N'oubliez pas de consulter la grille de correction avant de remettre votre travail afin de vous assurez d'avoir rempli tous les critères.

Annexe 1 : Sources d'information pertinentes

LIENS DIVERS

Galaxy Zoo Decision Trees: https://data.galaxyzoo.org/gz_trees/gz_trees.html

LIVRES

Prateek Joshi. 2015. *OpenCV with Python By Example*. Packt Publishing. 296 p. ISBN-13: 978–1785283932

Aurélien Géron. 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow—Concept, Tools, and Techniques to Build Intelligent Systems.* 566 p. ISBN-13: 978–1491962299

PAPIERS DE RECHERCHES[1]

Fang-Chieh Chou. 2014. *Galaxy Zoo Challenge: Classify Galaxy Morphologies from Images.* University of Stanford. 8p. http://cvgl.stanford.edu/teaching/cs231a_winter1415/prev/projects/C231a_final.pdf

Devendra Singh Dhami. 2015. *Morphological Classification of Galaxies into Spirals and Non-Spirals*. Indiana University. 82p. https://www.researchgate.net/publication/306215181_Morphological_classification_of_galaxies_into_spirals_and_non-spirals

Lior Shamir. 2009. *Automatic morphological classification of galaxy images*. Laboratory of Genetics of Baltimore 24p. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808694/pdf/nihms140565.pdf

Mohamed Abd Elfattah, Nashwa El-Benda, Mohamed A. Abu Elsoud, Aboul Ella Hassanien, M.P. Tolba. 2013. *An Intelligent Approach for Galaxies Images Classification*. Scientific Research Group in Egypt (SRGE). p167-172. http://ieeexplore.ieee.org/document/6920476/

Mohamed Abd El Aziz, I. M. Selim, Shengwu Xiong. 2016. *Automatic Detection of Galaxy Type From Datasets of Galaxies Image Based on Image Retrieval Approach*. 9p. https://www.nature.com/articles/s41598-017-04605-9.pdf

F. Ferrari, R. R. de Carvalho, M. Trevisan. 2015. *Morfometryka—A new way of establishing morphological classification of galaxies.* 15p. http://iopscience.iop.org/article/10.1088/0004-637X/814/1/55/pdf

[1] Certains liens ne fonctionnent seulement qu'avec une connexion VPN et le serveur mandataire (*proxy*) de la bibliothèque de l'École de technologie supérieure. Veuillez vous référez à ce lien pour plus d'informations : https://www.etsmtl.ca/bibliotheque/Infos-generales/Renseignements-utiles/Acces-hors-campus