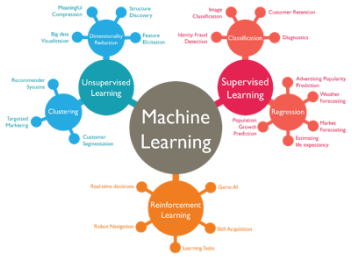

Etude d'un algorithme de Machine Learning : Feuille de route

Une visio est prévue pour échanger sur le sujet et les modalités. Une version initiale est présentée ci-dessous. D'ici là, merci de bien lire cette feuille de route et de commencer à réfléchir aux étapes 1 & 2 et de lister les questions que vous pourrez poser lors de la visio.

Vous remarquerez que nous avons légèrement modifié le sujet initial (c'est aussi plus détaillé, d'où plus d'étapes). L'accent est mis sur votre participation active aux choix (de l'algorithme comme des jeux de données) afin de vous laisser libre de choisir des algorithmes/méthodes/jeu de données qui vous motivent.

A bientôt en visio,

Luc Bouganim et Ludovic Javet

Modalités :

- Visio de 1h30 avec les 4 groupes pour lancer le TER et poser vos questions : merci de compléter ce doodle
- Des échanges par mail dès qu'une étape est finie ou dès qu'il y a un blocage (un correspondant par groupe) mettre systématiquement Ludovic.Javet@inria.fr et Luc.Bouganim@inria.fr en destinataires et merci de mettre le tag [TER] dans le sujet
- Un rapport à rendre par groupe : concaténation des comptes rendus successifs (voir les étapes ci-dessous)
- Une soutenance en fin de projet comme indiqué dans le sujet initial.

Etape 1 : Choisir un algorithme de supervised learning disponible sur Scikit-learn (ci-après nommé ALGO)

Contraintes sur l'algorithme :

- Adapté au semi-supervised learning et plus précisément au self-training (dispose de la méthode predict-proba)
- Applicable à des problèmes de classification **ET** de régression
- Potentiellement distribuable (cf. sujet initial)

Compte-rendu :

- Donner les liens vers les pages Scikit-learn et Wikipédia qui décrivent l'algorithme
- Expliquer votre motivation personnelle pour cet algorithme : 2 ou 3 arguments sur une dizaine de lignes
- Présenter des scénarios pertinents pour ALGO (quels types de données ?) : sur une dizaine de lignes

Etape 2 : Utiliser scikit-learn pour tester ALGO sur les jeux de données fournis avec scikit-learn

Objectifs :

- Choix de 2 jeux de données (datasets) dans Scikit-learn puis découpage en partitions : TRAINING et TEST
- Implémentation de base (paramètres par défaut) sur TRAINING et TEST avec métriques de qualité (à choisir en fonction de ALGO) en classification **ET** en régression (d'où les 2 datasets)

Compte-rendu :

- Donner une courte justification pour les choix des datasets : 4 lignes par dataset
- Implémenter un programme de test (lien vers le code ou zip)
- Discuter des métriques choisies et des résultats obtenus (1 page max)

Etape 3 : Faire varier les paramètres d'ALGO

Objectifs :

- Comprendre les différents paramètres d'ALGO
- Observer les effets (ou les non-effets) de la modification de ces paramètres

Compte-rendu :

- Lien vers le code ou zip
- Description et explication des paramètres : 2-3 lignes par paramètres
- Discuter des variations : pour chaque paramètre, tester, montrer les résultats et donner une courte explication sur une dizaine de lignes

Etape 4 : Faire un deuxième test avec un jeu de données que vous choisirez (nous vous fournirons des pistes)

Objectif : Etudier les différentes pistes et choisir « en conscience » 2 datasets adaptés à ALGO (un en classification et un en régression)

Compte-rendu :

- Indiquer les sources de vos datasets puis expliquer pourquoi vous les avez choisis : 10 lignes max
- Analyse de chaque dataset (1 page max par dataset) :
    - Quels sont les features ?
    - Quelle est la variable à prédire (target) ?
    - Détailler les opérations de nettoyage/préparation des données

Etape 5 : Implémentation sur les datasets de l'étape 4

Objectif : Reprendre les implémentations des étapes 2 et 3 avec de nouvelles données

Compte-rendu :

- Expliquer comment vous avez calibré les paramètres pour obtenir de meilleurs résultats : 1 page max par dataset
- lien vers code ou zip

Etape 6 : Faire fonctionner ALGO en apprentissage semi-supervisé sur les datasets de classification

Objectifs :

- Découper les datasets en LAB (ex. 10%), non-LAB (ex. de 0% à 80%), TEST (ex. 10%)
- Entrainer le modèle en supervisé sur LAB
- Améliorer le modèle en semi-supervisé sur non-LAB
- Tester le modèle sur TEST

Compte-rendu :

- Faire varier la proportion LAB/non-LAB et recalculer les métriques de qualité sur TEST : tracer des graphes avec en abscisse la proportion et en ordonnée les métriques de qualité
- Proposer une tentative d'explication : 1 page max par dataset

Etape 7 : Faire fonctionner ALGO en distribué

⇨ Sera détaillée une fois que l'étape 2 sera terminée