

Projets “Advanced Supervised Learning”

M2 DM - Université Lyon 2 - 2016/2017

Responsable : Julien Ah-Pine

1 Objectif du projet

Dans le cadre du cours “Advanced Supervised Learning” du Master 2 Data-Mining, il vous est demandé de réaliser un projet afin de compléter vos connaissances, pratique de méthodes et concepts vus en cours dans le but de parfaire vos compétences notamment en le langage R.

2 Aspects liés à l’organisation et à la remise du dossier

Les projets s’effectuent par groupe de 3 ou 4. Dans l’esprit, vous êtes une team de data scientists devant répondre à un besoin exprimé par une entreprise. La bonne organisation de votre travail collectif, la répartition efficace des tâches, et le travail en groupe/sous-groupes pour la résolution de points difficiles, constitueront autant d’atouts pour la réussite de votre projet. Il est attendu que vous fournissiez :

1. un script R dans lequel se trouvera toutes les fonctions développées,
2. un ou plusieurs fichiers comportant les données utilisées,
3. un rapport au format PDF accompagnant le projet.

Vous devrez créer une archive .zip contenant ces fichiers. Vous nommerez votre archive de la manière suivante NOM1_NOM2_NOM3_NOM4.zip où NOM_i sont les noms des membres du groupe. **Vous devrez envoyer par email cette archive au plus tard le 03 mars 2017 à l’adresse mail : julien.ah-pine@univ-lyon2.fr. ATTENTION :** vous êtes responsables de votre envoi et donc toute absence de ressource ou tout problème conduisant à l’impossibilité d’accéder correctement à votre travail est de votre responsabilité.

3 Descriptif du rapport accompagnant votre projet

L’objectif du manuscrit est de présenter et de mettre en valeur votre travail. Vous devez le rédiger dans l’esprit d’un rapport scientifique. Vous devrez présenter de façon claire :

- le sujet : les thématiques abordées et l’intérêt de celles-ci selon la nature du travail choisi (cf ci-dessous) ;
- les problématiques scientifiques : mise en perspective des méthodes et/ou des concepts et/ou du type de données que vous allez traiter ;
- la formalisation des problématiques traitées : les méthodes et/ou concepts et/ou type de donnée que vous allez traiter et/ou un état de l’art le cas échéant ;
- les différents choix effectués : librairies R/ou autre utilisées, jeux de données/benchmarks sur lesquels vous allez vous reposer en particulier ;
- les résultats expérimentaux : tables et/ou graphiques des benchmarks montrant notamment votre bonne maîtrise des protocoles d’évaluation en machine learning ;
- les résultats vis à vis des problématiques scientifiques posées : discuter en quoi votre travail permet de répondre aux problématiques posées ;
- une discussion scientifique : les avantages et les limites de la/les méthodes et ce qui est proposé dans la littérature comme extension, les “limites” de votre travail et ce qui serait intéressant de poursuivre ... ;

- une conclusion : en indiquant notamment les apports et les difficultés rencontrées lors de la réalisation de ce projet.

Il est clair également que le manuscrit devra comporter une bibliographie montrant votre investissement lors de la réalisation de ce travail.

L'organisation du manuscrit données ci-dessus n'est pas restrictive et vous pouvez ainsi ajouter toute section que vous jugerez utile pour la mise en valeur de votre travail.

Concernant le code R¹, il est attendu qu'il soit soigné et bien commenté. Vous devez programmer dans l'esprit de la réutilisabilité de votre travail par une tierce personne. Même si cela est encouragé, il n'est pas requis que le code soit optimisé à condition que l'exécution du script se fasse dans des temps raisonnables.

4 Liste de sujets

La nature des sujets diffère selon deux axes : algorithmique et data science.

Concernant le type “algorithmique”, l'objectif principal du travail à effectuer est d'implémenter “from scratch” une ou plusieurs méthodes et d'illustrer sur un ou plusieurs jeux de données simulés ou réels, les apports de la ou des méthodes étudiées. La difficulté principale est donc la maîtrise technique et l'implémentation de l'approche. Il est attendu du rapport que les étudiants présentent la ou les méthodes, le pseudo-code, le code R et qu'ils s'attardent sur les points particulièrement ardues de l'implémentation. Dans l'esprit, le rapport doit permettre à une personne tierce et novice de comprendre aisément les fondements de la méthode et son implémentation.

En ce qui concerne le type “data science”, le but est d'approfondir le traitement d'un type de données particulier ce qui inclut une présentation claire des données et des problèmes associés, les méthodes et outils de représentation numérique (extraction de features), le recensement des méthodes d'apprentissage adapté pour le type de données considéré et une comparaison sur un ou plusieurs jeux de données de ces différentes techniques. La principale difficulté ici est de produire un rapport et un code R complets sur le type de données considéré présentant et mettant en oeuvre les approches de l'état de l'art. Dans l'esprit, le rapport doit permettre à une personne tierce et novice de connaître aisément le traitement du type de données considéré.

Remarque : les références données ci-dessous sont toutes disponibles sur le net.

4.1 Sujet à finalité algorithmique

4.1.1 Sujet “random forest”

Ce projet consiste en l'étude approfondie des arbres de décision et de leur extension par le biais des forêts aléatoires. Les objectifs sont :

- Implémenter le pseudo-code des arbres décisionnels donné en cours. On attend donc deux fonctions `ArbreGeneration` et `DivisionAttribut`. Une seul des deux problèmes, régression ou catégorisation, pourra être considéré.
- Implémenter une fonction `RandomForest` qui utilise les fonctions précédentes et met en oeuvre les ré-échantillonnage sur les individus et sur les variables tels que décrit dans le pseudo-code du cours.
- Présenter dans le rapport un texte synthétique sur les fondements et propriétés des forêts aléatoires. Vous pourrez vous inspirer du cours mais toute autre référence pourra être utilisée pour enrichir le propos.
- Montrer aux travers d'expériences la supériorité des forêts aléatoires sur les arbres de décision permettant ainsi de valider empiriquement les propriétés théoriques discutés en cours sur les méthodes de ré-échantillonnage.

Quelques références à titre indicatif que vous complèterez à votre guise et selon les besoins de votre projet : [Breiman, 2001, Rokach, 2010, Breiman, 1996].

1. Et autres langages le cas échéant.

4.1.2 Sujet “adaboost”

Ce projet consiste en l’étude approfondie des deux algorithmes de base d’adaboost pour la catégorisation. Les objectifs sont :

- Implémenter le pseudo-code des approches adaboost (problèmes binaires) et adaboost.M1 (problèmes multiclassés) tels que décrits dans les articles fondateurs [Freund et al., 1996, Freund and Schapire, 1997] :
- Présenter dans le rapport une synthèse expliquant les fondements et propriétés des méthodes adaboost. Il est attendu des étudiants un “bref” état de l’art sur le sujet. En particulier, les liens/comparaisons avec d’autres méthodes telles que les SVM ou le bagging seront appréciés.
- Montrer aux travers d’expériences les bonnes performances des méthodes adaboost vis à vis d’une ou plusieurs méthodes baseline que vous choisirez.

Quelques références à titre indicatif que vous complèterez à votre guise et selon les besoins de votre projet : [Freund et al., 1999, Rokach, 2010].

4.2 Sujet à finalité data science

4.2.1 Sujet “extraction d’information dans des textes”

Ce projet consiste à traiter la tâche “term annotation” du corpus GENIA : <http://www.geniaproject.org/genia-corpus/term-corpu>. Celui-ci consiste en des résumés d’articles scientifiques issus de pubmed². La tâche consiste à apprendre un classifieur capable de détecter automatiquement des termes appartenant à certaines catégories (des entités nommées) telles que des noms de protéines, de cellules... Les objectifs sont :

- Présenter dans le rapport une synthèse expliquant les fondements et propriétés des différentes méthodes d’apprentissage. Il est attendu des étudiants un bref état de l’art sur ces techniques.
- Pré-traiter les documents du corpus xml de façon à enlever les méta-données et à représenter les textes sous forme numérique selon les besoins des classifieurs.
- Pré-traiter les documents de sorte à extraire les features intéressants pour la tâche de catégorisation et selon la méthode considérée (voir [Zhou et al., 2004] par exemple).
- Traiter la tâche par les modèles probabilistes suivants : Naïve Bayes (NB), Hidden Markov Models (HMM) et Conditional Random Fields (CRF). Pour cela, vous utiliserez des bibliothèques existantes.
- Tester ces trois modèles sur la tâche et faire une analyse critique et comparative des performances des différentes méthodes.

L’ensemble des bibliothèques utilisées doivent être en R.

Quelques références à titre indicatif que vous complèterez à votre guise et selon les besoins de votre projet : [Sun et al., 2012, Jiang, 2012].

4.2.2 Sujet “catégorisation d’images”

Ce projet consiste à traiter la tâche “classification” de la compétition Pascal VOC de 2005 : <http://host.robots.ox.ac.uk/pascal/VOC/voc2005/index.html>. Il s’agit d’images et la tâche consiste à reconnaître si elles contiennent une des classes suivantes : “motorbikes”, “bicycles”, “people”, “cars”. Les objectifs sont :

- Se former sur les méthodes d’extraction de features pour les images. Il est notamment attendu que les histogrammes de couleurs et une méthode basées sur la texture telle que SIFT (ou SURF) soient étudiés et présentés dans le rapport.
- Pré-traiter les images de façon à les représenter sous forme numérique selon les besoins des classifieurs. Existe t-il des outils en R pour effectuer l’extraction de features ? Si non,

2. <https://www.ncbi.nlm.nih.gov/pubmed>

vous pourrez utiliser des outils tierces comme OpenCV³ ou VLFeat⁴.

- Traiter en R la partie apprentissage. Vous comparerez plusieurs méthodes incluant au minimum les SVM et la régression logistique pénalisée (et toute autre méthode que vous souhaitez tester qu'elle ait été vue en cours ou pas -à condition de la présenter dans le rapport-).
- Tester les méthodes choisies sur la tâche et faire une analyse critique et comparative des performances des différentes méthodes.

Quelques références à titre indicatif que vous complèterez à votre guise et selon les besoins de votre projet : [Tuytelaars et al., 2008], http://www.scholarpedia.org/article/Scale_Invariant_Feature_Transform.

May the force be with you !

Références

- [Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2) :123–140.
- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5–32.
- [Freund et al., 1999] Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780) :1612.
- [Freund and Schapire, 1997] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1) :119–139.
- [Freund et al., 1996] Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156.
- [Jiang, 2012] Jiang, J. (2012). Information extraction from text. In *Mining text data*, pages 11–41. Springer.
- [Rokach, 2010] Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2) :1–39.
- [Sun et al., 2012] Sun, Y., Deng, H., and Han, J. (2012). Probabilistic models for text mining. In *Mining text data*, pages 259–295. Springer.
- [Tuytelaars et al., 2008] Tuytelaars, T., Mikolajczyk, K., et al. (2008). Local invariant feature detectors : a survey. *Foundations and trends® in computer graphics and vision*, 3(3) :177–280.
- [Zhou et al., 2004] Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. (2004). Recognizing names in biomedical texts : a machine learning approach. *Bioinformatics*, 20(7) :1178–1190.

3. https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_tutorials.html

4. <http://www.vlfeat.org/>