Projets “Advanced Supervised Learning”

M2 DM - Université Lyon 2 - 2017/2018

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. 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.zip où NOMi sont les noms des membres du groupe. **Vous devrez envoyer par email cette archive au plus tard le 22 décembre 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 rédigez 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 litérature 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 donnée 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 R [1], 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 trois axes : "algorithmique", "données complexes", "contexte particulier".

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 ardus 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 "données complexes", 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 recencement 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 traitenemt du type de données considéré.

Pour ce qui est du type "contexte particulier", il s'agit ici de situations où l'apprentissage nécessite des traitements supplémentaires et spécifiques. En effet, dans des études de cas réels, nous recontrons des situations particulières, qui nécessitent des traitements particuliers en amont ou en aval ou au cours de la phase d'apprentissage. Les problèmes de données mixtes ou de fusion d'information ou de distributions de classes asymmétriques en sont des illustrations. Dans ce cadre, il est attendu des étudiants qu'ils implémentent des techniques permettant de tenir compte de ces contextes particuliers. Dans l'esprit, le rapport doit présenter la problématique, les méthodes classiques et illustrer sur un exemple concret et à l'aide d'une méthode d'apprentissage, l'intérêt des techniques mises en oeuvre dans le cadre de la problématique étudiée.

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 approdondie 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-échantillonage sur les individus et sur les variables tels que décrit dans le pseudo-code du cours.

1. Et autres langages le cas échéant.

— 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-échantillonage.

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

4.1.2 Sujet "adaboost"

Ce projet consiste en l'étude approdondie 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 multiclasses) 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 compléterez à votre guise et selon les besoins de votre projet : [Freund et al., 1999, Rokach, 2010].

4.2 Sujet à finalité "données complexes"

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 pubmeb [2]. 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 librairies existantes.
— Tester ces trois modèles sur la tâche et faire une analyse critique et comparative des performances des différentes méthodes.

L'ensemble des librairies utilisées doivent être en R.

Quelques références à titre indicatif que vous compléterez 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

2. `https://www.ncbi.nlm.nih.gov/pubmed`

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, vous pourrez utiliser des outils tierces comme OpenCV [3] ou VLFeat [4].
- 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 souhaiterez 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 compléterez à votre guise et selon les besoins de votre projet : [Tuytelaars et al., 2008], `http://www.scholarpedia.org/article/Scale_Invariant_Feature_Transform`.

4.3 Sujet à finalité “contexte particulier”

4.3.1 Sujet “apprentissage déséquilibré”

En catégorisation, dans de très nombreux problèmes réels, la distribution des classes est fortement déséquilibrée. Dans le cas binaire cela se traduit par une classe très représentée par rapport à l'autre. On rencontre cette situation pour les problèmes de détection d'anomalies, de fraudes, de maladies rares... Si on ignore cette asymmétrie, toute méthode serait implicitement biaisée et aurait plus de difficultés à détecter la classe minoritaire qui est souvent celle d'intérêt. Les objectifs de ce projet sont les suivants :

- Présenter la problématique et expliquer en quoi celle-ci nécessite un traitement particulier. Dans cette perspective, il faut expliquer également en quoi les mesures d'erreurs vues en cours doivent être adaptées.
- Faire un état de l'art sur les différentes méthodes classiques basées sur l'échantillonnage. Les techniques discutées et implémentées doivent au moins contenir les approches de : sous-échantillonage, sur-échantillonage et échantillonage synthétique (SMOTE, Borderline-SMOTE et ADASYN).
- Implémenter en R, ces différentes méthodes classiques. Tester les méthodes sur l'étude de cas suivante : la détection d'intrusion à partir des données de la KDD Cup de 1999 (`http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`). Vous emploierez pour cela au moins une méthode de catégorisation vue en cours parmi les suivantes : SVM ou régression logistique pénalisée ou arbres de décision.
- Analyser les résultats de classification selon une (ou plusieurs) mesure d'erreur adéquate. Comparer les méthodes d'échantillonage entre elles.

Quelques références à titre indicatif que vous compléterez à votre guise et selon les besoins de votre projet : [Chawla et al., 2002, He et al., 2008, Han et al., 2005].

3. `https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_tutorials.html`
4. `http://www.vlfeat.org/`

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.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote : Synthetic minority over-sampling technique. J. Artif. Int. Res., 16(1) :321–357.

[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.

[Han et al., 2005] Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote : A new over-sampling method in imbalanced data sets learning. In Huang, D.-S., Zhang, X.-P., and Huang, G.-B., editors, Advances in Intelligent Computing, volume 3644 of Lecture Notes in Computer Science, pages 878–887. Springer Berlin Heidelberg.

[He et al., 2008] He, H., Bai, Y., Garcia, E., and Li, S. (2008). Adasyn : Adaptive synthetic sampling approach for imbalanced learning. In Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pages 1322–1328.

[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.