



Machine Learning Projet

M2 Informatique BI&A

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Ce projet sera effectué par groupe de deux étudiants maximum et porte sur l'apprentissage dans un contexte de classification binaire.

Le travail sera à rendre par mail à guillaume.metzler@univ-lyon2.fr avant le **14 Janvier 2024**. Ce travail se composera d'un dossier retraçant vos démarches et résultats entrepris pour traiter le sujet, de plus amples explications vous sont données ci-dessous. Vous êtes libres d'utiliser le langage de votre choix pour effectuer ce travail, R ou Python. Choisissez celui avec lequel vous êtes le plus à l'aise.

1 A propos des données et préparation

Vous pourrez trouver plusieurs jeux de données à l'adresse suivante

[Téléchargement des jeux de données.](#)

Vous pourrez également trouver un code Python permettant de mettre en forme les différents jeux de données afin que ces derniers soient prêts à être employés.

[Préparation des données](#)

On pourra effectuer les différentes étapes :

- regarder le code précédent afin d'identifier les différentes étapes

- détecter d'éventuelles valeurs manquantes et faire de l'imputation
- identifier les caractéristiques des jeux de données (taille, dimension, type de problème, ratio des classes)

Vous trouverez à la fin de ce document quelques indications pour la rédaction de votre rapport.

2 Travail à effectuer

Votre travail se décomposera en trois parties. Pour les différentes phases de votre travail, vous devrez

- présenter les algorithmes/méthodes employé(e)s
- le protocole expérimental afin que la présentation permette la reproduction des résultats obtenus
- concevoir des variantes personnelles des algorithmes standards (*i.e.* combinaison avec des méthodes d'échantillonnage, cost-sensitive, . . .)
- évaluer le temps d'apprentissage des algorithmes

On prendra par exemple soin de préciser quelles sont les mesures de performances employées ou encore les loss optimisées dans le cas où cette dernière est personnalisée.

2.1 Approches non paramétrique

Dans cette première partie, vous concentrerez sur des approches basées sur l'algorithme des k - plus proches voisins uniquement.

Vous devrez rappeler le fonctionnement de l'algorithme et la comparer au moins avec trois variantes que vous proposerez en fonction des jeux de données considérées et des difficultés qu'ils présentent.

2.2 Approches paramétrique linéaires

Dans cette deuxième partie, on va chercher à comparer les algorithmes qui apprennent des classifieurs linéaires

- SVM linéaire (ou noyau linéaire, éventuellement gaussien, mais ce n'est pas très *fair* en pratique)
- Régression logistique

A nouveau, on proposera au moins deux variantes pour ces deux algorithmes qui permettent d'améliorer les performances de base de ces algorithmes.

2.3 Approches non linéaires

Dans cette dernière partie, on va chercher à tester les approches non linéaires ou par boosting

- Arbres de décisions/forêts aléatoires
- Adaboost
- Gradient Boosting

3 Quelques suggestions

Je vous donne ci-dessous une liste non exhaustive des méthodes que vous pourriez utiliser pour votre travail.

- **Pre-Process sur les données** : utilisation d'algorithmes d'over-sampling (random - SMOTE - Adasyn - ...) ou encore des approches d'under-sampling (random - Tomek Link - Edited Nearest Neighbour - NearMiss - ...) vous pouvez aussi utiliser des méthodes dites cost-sensitive qui vont accorder plus de poids aux exemples d'une classe donnée, voire même des poids spécifiques à chaque exemple.
- **Post-traitement** : combinez les résultats issus de différents modèles (bagging) afin de créer un modèle potentiellement plus puissant.

N'hésitez pas à regarder sur internet quelques exemples d'utilisations des algorithmes sus-mentionnés et votre objectif sera de les adapter au contexte des données (consulter des sites comme Kaggle - MachineLearningMastery ou encore Medium qui seront pour vous une bonne source d'inspiration).

Ce qui suit est une suggestion de rédaction pour le rapport de votre projet.

Introduction

Vous présenterez rapidement le contexte du projet les objectifs et vous pourrez également présenter les caractéristiques des jeux de données considérées.

Méthodologie

Pour chaque partie précédemment citée, vous procéderez de la façon suivante

Vous commencerez par présenter les notations que vous allez employer tout au long de la rédaction de votre rapport.

Vous présenterez ensuite les outils que vous allez utiliser dans la partie expérimentale. On commencera par parler de ce que l'on souhaite maximiser, *i.e.* la mesure de performance en la définissant avant de s'attaquer à la présentation des algorithmes (de façon très courte)

Par exemple, si vous faites une contribution basée sur du boosting et que vous combinez avec des méthodes à noyaux, il faudra rappeler ce qu'est le boosting, ce que sont les méthodes à noyaux (ce sont les approches de bases) et ensuite vous expliquerez comment vous combinez les deux pour résoudre le problème confié.

Il ne faut pas hésiter à présenter le processus de façon *abstraite*, c'est-à-dire avec des notations mathématiques et ne pas être uniquement verbeux.

Un pseudo-code est également appréciable pour synthétiser l'approche proposée.

Dans le cas où vous proposez plusieurs approches à des fins de comparaisons, il faudra prendre soin de présenter les différentes approches et de justifier pourquoi vous intéressez à ces approches là.

Remarque : il n'est pas nécessaire de présenter tous les algorithmes employés, mais uniquement ceux qui servent à l'élaboration d'une version "exotique".

Expériences

Vous dresserez ensuite votre protocole expérimentale qui présentera la ou les méthodes sélectionnées pour répondre à la tâche demandée. Celui-ci comprend en général trois parties

Protocole expérimental

Pour chaque partie, vous procéderez de la façon suivante

On commencera par présenter les données que l'on utilise, en général sous la forme d'un tableau qui comprend : *le nom du jeu de données, le taille de l'échantillon, la dimension du jeu de données* et toute autre information que vous pourriez juger pertinente.

Vous présentez rapidement les expériences que vous allez faire, *i.e.* les différents algorithmes testés, le range des hyper-paramètres employés ainsi que la façon dont sont optimisées ces hyper-paramètres (cross-validation en k -folds, simple validation ou est-ce que vous faites le choix de les fixer). Quels sont vos ensembles d'entraînement/validation/test ?

Les informations que vous fournissez dans cette section doivent permettre au lecteur de pouvoir reproduire les résultats que vous allez présenter dans la sous-section suivante.

Résultats

Ici vous allez présenter et analyser les résultats obtenus à l'aide de graphiques et/ou tableaux. Outre les performances, on pourra aussi s'intéresser au critère de rapidité d'un algorithme.

L'analyse doit aussi permettre de mettre en exergue les avantages/inconvénients des méthodes proposées. Cela peut passer par l'utilisation d'autres mesures de performances/critères pour évaluer/comparer vos algorithmes.

Pour ce qui est des résultats, on pourra présenter les résultats sous la forme d'un tableau similaire à celui présenté en Table 1.

Conclusion

Il s'agit de conclure quant à votre étude. Reprendre le travail proposé et les principales conclusions. Il est également important de proposer des perspectives à votre travail en fonction des résultats obtenus et de l'approche proposée. Quelle(s) méthode(s) non explorées ici auraient pu être utilisées pour améliorer les résultats et pourquoi cela vous semble pertinent étant données les expériences effectuées.

TABLE 1 – Mean test F1 over 10 iterations

Dataset	knn	svm_linear	svm_poly	svm_gauss
00.62% abalone20	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
01.39% abalone17	13.7 ± 6.3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
02.36% yeast6	44.0 ± 14.6	0.0 ± 0.0	38.9 ± 18.0	0.0 ± 0.0
03.31% wine4	5.0 ± 5.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
06.67% libras	81.1 ± 12.7	71.0 ± 10.0	88.3 ± 7.9	75.6 ± 11.0
10.23% pageblocks	84.9 ± 2.2	75.7 ± 2.9	82.3 ± 2.5	78.6 ± 2.5
10.98% yeast3	66.8 ± 5.9	65.6 ± 10.8	76.4 ± 5.2	71.3 ± 8.2
13.60% abalone8	22.7 ± 1.3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
14.29% segmentation	85.1 ± 3.3	45.8 ± 3.7	64.1 ± 4.9	50.0 ± 4.6
22.73% hayes	14.8 ± 10.2	0.0 ± 0.0	39.4 ± 19.9	0.0 ± 0.0
23.52% vehicle	88.1 ± 3.4	60.3 ± 5.3	92.6 ± 1.7	56.7 ± 6.5
30.00% german	41.3 ± 6.4	5.3 ± 7.6	56.9 ± 6.8	16.8 ± 6.6
32.71% glass	70.2 ± 4.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
33.15% wine	86.5 ± 4.3	89.6 ± 5.9	90.1 ± 5.8	86.2 ± 6.1
34.90% pima	41.3 ± 6.6	5.0 ± 5.4	44.7 ± 6.1	17.5 ± 4.8
35.90% iono	82.5 ± 4.1	88.9 ± 4.0	92.0 ± 3.4	93.3 ± 3.4
37.50% autmpg	82.4 ± 5.5	0.0 ± 0.0	83.0 ± 3.4	0.0 ± 0.0
46.08% balance	95.3 ± 1.9	94.5 ± 1.2	99.9 ± 0.3	97.3 ± 0.8
Mean	55.9 ± 5.4	33.4 ± 3.2	52.7 ± 4.8	35.7 ± 3.0
Average Rank	2.00	3.72	1.72	2.56