Projet "Méthodes avancées en apprentissage supervisé et non-supervisé"

M2 SISE - Université Lyon 2 - 2021/2022

Responsable : Julien Ah-Pine

1 Objectif du projet

Dans le cadre du cours "Méthodes avancées en apprentissage supervisé et non-supervisé" du Master 2 SISE, il vous est demandé de réaliser un projet afin de compléter vos connaissances et pratique de méthodes vues 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

Il est impératif de respecter les instructions données ci-dessous. Tout manquement pourra donner lieu à une pénalisation.

2.1 Groupes et contenu du dossier

Les projets s'effectuent par groupe de 3. Il est attendu que vous fournissiez :

1. un script R dans lequel se trouvera tout le code (démarquer bien les deux parties -cf ci-dessous-),
2. un fichier comportant les données utilisées,
3. un rapport au format PDF accompagnant le projet.

Le code R doit s'exécuter sans aucun problème et produire l'ensemble des résultats attendus. Il doit également être propre, bien structuré et bien commenté. Similairement, le rapport constitue un élément important de votre rendu. Il doit être bien structuré, bien rédigé et il doit mettre en perspective l'ensemble de votre travail que cela concerne vos implémentations ou vos expériences ou vos conclusions. Vous trouverez dans la suite du sujet plus de détails sur ce qui est attendu.

Dans le cadre de votre rapport, une introduction générale sur l'objectif de ce projet et une conclusion sur les apports/limites de ce travail seront les bienvenues.

2.2 Constitution des groupes et des données

Une partie du projet correspond à l'étude approfondie d'un jeu de données que vous choisirez. Dans cette perspective, vous devrez m'envoyer par mail **au plus tard le 10 décembre 2021 à l'adresse : julien.ah-pine@univ-lyon2.fr**, la constitution de votre trinôme et le jeu de données que vous souhaitez analyser. Celui-ci devra comporter au moins 1500 observations et au moins 12 variables. Dans votre message, vous m'indiquerez la source des données et une brève description du contexte et de la problématique que vous souhaitez étudier.

2.3 Modalités du rendu du dossier

Vous devrez créer une archive .zip contenant les fichiers correspondant aux éléments cités plus haut. Vous nommerez votre archive ainsi que les fichiers qu'elle contient de la manière suivante NOM1_NOM2_NOM3.(zip|R|pdf|RData|csv|txt) où NOMi sont les noms des membres du groupe. Vous devrez soumettre cette archive **au plus tard le 7 janvier 2022 à 23h59 via la boîte de dépôt du cours Moodle accessible à :** `https://moodle.univ-lyon2.fr/course/view.php?id=2626`. Si

vos données sont trop volumineuses vous indiquerez un lien de téléchargement de celles-ci. **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 travail à réaliser

Le projet est constitué de 2 parties :

— Une étude de cas dirigée mettant en oeuvre une approche supervisée le "one-class svm" et une approche non supervisée le "kernel PCA".

— Une étude de cas que vous choisirez afin de mettre en oeuvre et de comparer plusieurs méthodes d'apprentissage supervisé.

3.1 Partie 1

3.1.1 Programmation

h Il s'agit de réaliser le code et les graphiques répondant aux questions posées dans le sujet que vous pourrez récupérer à l'adresse suivante : `https://eric.univ-lyon2.fr/~jahpine/cours/m2_sise-dm/tp4.pdf`.

Dans votre script, vous indiquerez en commentaire, le numéro de la question à laquelle le code correspond.

Idéalement, l'exécution du code doit se faire sans erreur et produire tous les résutats attendus.

3.1.2 Rapport

Le rapport devra contenir :

— Une présentation de l'étude de cas.

— Une présentation succincte des deux méthodes utilisées pour faire la prédiction.

— Une réponse à chaque question posée dans le sujet. Selon le type de questions, la réponse peut consister en :

  — Des lignes de code R avec des commentaires sur les commandes utilisées (ce qu'elle font, ce qu'elles prennent en input et ce qu'elles donnent en output) et/ou sur les variables définies (ce qu'elles représentent).

  — Des explications quant à des questions spécifiques (comme pour la question 16).

  — Des sorties graphiques et des commentaires (question 27 en particulier).

— Dans le prolongement de la question 27, une analyse, comparaison et discussion des résultats obtenus pour chaque type de modèle conduisant à une conclusion sur le modèle spécifique que vous préconisez pour répondre à la problématique.

3.2 Partie 2

3.2.1 Programmation

L'objectif est ici d'étudier un cas présentant un problème de prédiction que ce soit dans le cadre de la régression ou de la catégorisation. Vous devrez pour cela implémenter le code permettant de mettre en oeuvre les points suivants :

— La lecture des données (que vous fournirez avec votre dossier en format .csv ou .xls ou .RData).

— Le pré-traitement des données si nécessaire.

— La validation croisée permettant d'évaluer les performances des méthodes suivantes :

  — Modèle linéaire pénalisé par une fonction de régularisation elasticnet.

  — Réseau de neuronnes avec une couche cachée.

  — SVM.

  — Autre méthode de votre choix sortant des techniques vues en cours (optionel).

  Pour chaque type de méthode, vous testerez plusieurs ensembles de paramètres.

— Des graphiques permettant de comparer les résultats de chaque méthode avec les différents paramétres utilisés.
— Des graphiques permettant de comparer les meilleurs modèles des trois types de méthodes.

3.2.2 Rapport

Le rapport devra contenir :
— Une présentation du jeu de données et de la problématique abordée.
— Une présentation succincte [1] des méthodes utilisées pour faire la prédiction.
— Une présentation succincte des librairies R utilisées [2] pour faire les estimations de chaque méthode.
— Une présentation des différents ensembles de paramètres utilisés pour chaque type de modèle.
— Une présentation de votre protocol expérimental.
— Une présentation, analyse et discussion des résultats obtenus pour chaque méthode vis à vis de chaque ensemble de paramètres utilisés.
— Une présentation, analyse, comparaison et discussion des meilleurs résultats obtenus pour chaque type de modèle conduisant à une conclusion sur le modèle spécifique que vous préconisez pour répondre à la problématique.

1. Sans rappeler tout le cours, faites un résumé des points que vous jugerez importants pour chaque méthode.
2. Pour vous aider, vous pourrez consulter la page suivante : `https://cran.r-project.org/web/views/MachineLearning.html`.