Projet d'apprentissage statistique

Fabrice Rossi

1 Objectif général

L'objectif du projet est de mettre en œuvre des méthodes d'apprentissage statistique dans un cadre prédictif. Les données fournies présentent une sélection des problèmes qu'on peut rencontrer en pratique : il s'agit de données mixtes, avec des valeurs manquantes et de taille suffisante pour qu'une exploration manuelle ne soit pas facile. D'autres problèmes ont été introduits. Il faudra donc mettre en œuvre diverses techniques pour s'adapter à ces données, à la fois au niveau des modèles employés mais aussi pour prétraiter les données, par exemple.

2 Consignes générales

2.1 Aspects logiciels

Le projet sera réalisé de préférence en R (à défaut en python), en utilisant tous les *packages* jugés utiles pour l'application. En plus des *packages* qui fournissent des méthodes d'apprentissage, on pourra ainsi utiliser *caret* (pour la validation croisée et les méthodes associées) et l'un des *packages* qui permettent d'estimer les valeurs manquantes dans des données (*mi*, *mice*, *missForest*, etc.).

2.2 Données

Les données sont fournies sous la forme de deux fichiers csv. Le fichier *learn* contient environ 5000 observations d'environ 20 variables explicatives et de la variable numérique à prédire (colonne Y). Le fichier *test* contient aussi environ 5000 observations mais sans cette variable à prédire et est à utiliser pour la prédiction finale (cf les résultats attendus).

2.3 Pré-traitements

Les données contiennent des `NA` qui sont autant de valeurs manquantes. Comme le fichier d'évaluation (*test*) contient aussi des valeurs manquantes, il est impératif de proposer une chaîne de traitement capable de réaliser des prévisions même quand des données sont manquantes. La solution la plus simple est de s'appuyer sur un procédé d'imputation (donc sous forme de pré-traitement), mais d'autres solutions sont possibles.

Tous les pré-traitements réalisés doivent l'être par des programmes R. Aucun pré-traitement manuel (par exemple avec un tableur) n'est autorisé.

2.4 Méthodes

Le projet devra mettre en œuvre au moins deux méthodes prédictives différentes comme par exemple les *support vector machines* et les *random forests*. L'analyse exploratoire préalable devra s'appuyer sur au moins une méthode de classification non supervisée.

3 Travail à effectuer

L'objectif final du projet est de construire deux modèles prédictifs et de fournir pour chaque modèle ses prévisions sur les données de test et une évaluation de ses performances attendues. On commencera par une analyse exploratoire des données, suivie par une modélisation prédictive. Lors de celle-ci, on tentera de sélectionner les variables les plus pertinentes.

3.1 Analyse exploratoire

Avant de construire des méthodes prédictives, les données devront être étudiées par une analyse exploratoire classique. Parmi les angles d'approche possibles, citons :

- présence d'éventuelles données aberrantes ou extrêmes ;
- dépendance éventuelle entre la présence de données manquantes et les valeurs des données observées ;
- présence d'une éventuelle structure de classes (*clusters*) dans les données ;
- lien à priori (corrélation par exemple) entre les variables explicatives et les variables à prédire (cf ci-dessous) ;
- etc.

3.2 Analyse prédictive

Les tâches de prédiction sont les suivantes :

1. le premier modèle cherche à prédire la valeur exacte de Y, en utilisant une perte quadratique. Il s'agit donc d'un problème de régression ;
2. le second modèle cherche à prédire une valeur discrète Z dérivée de Y de la façon suivante : si $Y < 0$, alors $Z = 0$, si $Y \in [0, 2]$, alors $Z = 1$ et enfin si $Y > 2$, alors $Z = 2$. Il s'agit donc d'un problème de discrimination pour lequel on utilisera la fonction de perte classique $l_0(u, v) = \mathbf{1}_{u \neq v}$. Cette variable n'est pas fournie dans le fichier *learn* et doit donc être construite dans le programme de traitement.

Attention, l'objectif d'obtenir deux *modèles* ne correspond pas à l'obligation d'étudier deux *méthodes* différentes. On peut très bien arriver à la conclusion que les deux modèles sont de même nature (par exemple deux fois un arbre de décision), l'important étant de tester au moins deux méthodes pour chaque situation.

3.3 Sélection de variables

Un autre objectif est de déterminer quelles sont les variables importantes pour ces deux modèles. Une discussion sur les variables est donc attendue dans le rapport à rendre, cf ci-dessous. Si les données le justifient, il est opportun de construire des modèles n'utilisant qu'une partie des variables d'origine.

3.4 Sélection des paramètres et évaluation des modèles

Les paramètres des modèles seront sélectionnés par une procédure de ré-échantillonnage adaptée. De même, on choisira le modèle le plus adapté à chaque tâche par une méthode adaptée. On évaluera aussi les futures performances du modèle retenu d'une façon robuste.

4 Résultats attendus

4.1 Contenu du rapport

Le rapport doit contenir les éléments suivants :

1. analyse des données manquantes (volume, modèle d'absence de données, etc.) ;
2. justification du choix des techniques utilisées pour contourner l'absence de certaines observations ;
3. analyse exploratoire minimale des données (statistiques univariées, dépendances, etc.) ;
4. justification des modèles prédictifs choisis dans les deux problèmes (prévision de Y et de Z) ;
5. description précise de la chaîne de traitement : prétraitements éventuels, ajustement des modèles, choix du modèle, évaluation de ses performances attendues (le mémoire doit impérativement contenir un tableau indiquant la qualité numérique attendue pour les prévisions sur le fichier *test*) ;
6. analyse de l'importance des variables : cela peut être fait avant l'ajustement des modèles, pendant celui-ci ou après le choix du modèle final. Dans tous les cas, le rapport doit discuter de l'opportunité de construire des modèles sur une partie seulement des variables. Si c'est le cas, les prévisions finales et les performances attendues doivent concerner les modèles n'utilisant que les variables pertinentes.

4.2 Remise du travail

Les étudiants doivent rendre leur travail sous forme de trois fichier :

1. un rapport de quelques pages, le fichier `rapport.pdf` (format pdf uniquement), détaillant le choix des modèles, les procédures et méthodes employées, et les résultats obtenus. Il est très vivement conseillé d'utiliser le système *knitr* pour écrire ce rapport (avec markdown dans Rstudio). Le rapport devra notamment faire apparaître explicitement les performances attendues pour les deux modèles retenus. Aucun code ne devra apparaître dans le rapport ;
2. un code R (ou à défaut python), le fichier `traitements.R`, réalisant l'intégralité des traitements demandés, avec chargement des données dans le dossier d'exécution (chemins absolus interdits) et sauvegarde des résultats dans ce dossier. Aucune intervention humaine ne doit être demandée dans ce code ;
3. un fichier `prevision.Rds` contenant une `data.frame` avec exclusivement deux colonnes. La colonne `Y` contiendra les valeurs estimées de la variable Y pour les données de test. La colonne `Z` contiendra les valeurs estimées de la variable Z pour les données test, au format `factor`. Les prévisions seront données dans l'ordre du fichier *test*.