Gaspard Palay / Validation Croisée

Gaspard Palay

29/01/2021

1. Critères d'évaluation

1. Comportement du Rmd lors de son exécution
2. Qualité de la rédaction du dossier
3. Didactisme et pertinence du dossier
4. Qualité et lisibilité du Rmarkdown
5. Qualité des explications du modèle mathématique

2. Lien vers le document commenté

En cliquant **ici**, vous trouverez le lien menant au GitHub de Nicolas ALLIX hébergeant le fruit de sa réalisation.

3. Auteurs du document commenté

Le document évalué dans le cadre de ce rendu a été produit par Nicolas ALLIX et Rindra LUTZ étudiants en MSc Data Management à Paris School of Business.

4. Synthèse du document

Le document est une production RMD et PDF présentant une synthèse textuelle de ce qu'est la validation croisée. Il s'organise en plusieurs étapes : - Introduction - méthodes descriptives - Méthodes prédictives - Validation croisée - Résumé de la validation croisée - L'overfitting - La gestion des bases de données non équilibrées

Toutes ses parties s'articules autour d'une définition littéraire des concepts. Elle reprend les définitions et les enjeux.

Pour résumé ce document : Il existe deux modèles d'apprentisage : - L'apprentissage supervisé (méthodes prédictives) - L'apprentissage non supervisé (méthodes descriptives)

Parmis les méthodes prédictives, les principaux algorithmes sont :
- Régressions (linéaire, logistique, etc)
- Arbres de décision
- Réseaux de neuroness
- Analyse discriminante
- SVM - Support Vecteur Machine

Parmis les méthodes dexcriptives les principaux algorithmes sont :
- les analyses factorielles

- les analyses typologiques
- modèles combiantoires
- les modèles à base de règles

La validation croisée sert à mesurer la fiabilité d'un modèle.
Pour cela on divisie nos données en
- Train (population d'apprentissage)
- Test (Population de test de notre modèle)
- Valid (Tester plusieurs modèles sur cet échantillon)

Chaque jeu de données, selon la validation croisée, sont structurées selon trois modèles :
- LOOCV (leave-one-out cross-validation)
- LKOCV (leave-k-out cross-validation)
- k-fold cross-validation

5. Extrait commenté des parties de code

Le document est un RMD ne comportant pas de chunk ou de code R… il est difficile d'en extraire des parties. Il comporte des explications textuelles dont je peux extraire des exemples ci dessous :

Le sur apprentissage : problématique majeure en modélisation

Un modèle trop complexe, intégrant trop d'inputs et « épousant » trop les données d'apprentissage amènera donc une très bonne performance sur l'échantillon d'apprentissage (par construction), mais aura trop appris, notamment les bruits ou cas aberrants.
Il sera alors moins performant sur des données qui n'ont pas servi à la construction du modèle, c'est-à-dire sur les données sur lesquelles on souhaite faire la prédiction.
L'enjeu est donc de trouver le bon niveau de sophistication pour obtenir un bon niveau de performance sur l'échantillon d'apprentissage et sur l'échantillon de test.

Il n'y a pas sur-apprentissage lorsque la performance du modèle en Test est légèrement plus faible que celle en Train. Un écart trop grand est signe de **sur-apprentissage**.

Gestion des bases de données non-équilibrées

Dans les tâches de classification, la répartition des classes dans la base de données peut être déséquilibrée, c'est-à-dire que le nombre d'observations par classe peut ne pas être le même d'une classe à l'autre : si l'on dénote n_i le nombre d'observations de la i-ème classe, alors il existe {i,j} tel que n_i soit différent de n_j. Dans ce cas, pour éviter que la performance de validation (et d'apprentissage) ne soit biaisée par une répartition changeante des classes d'un ensemble de validation (resp. d'apprentissage) à un autre, il est recommandé d'utiliser une validation croisée stratifiée (« stratified cross validation »). La stratification consiste à s'assurer que la répartition des classes soit la même dans tous les ensembles d'apprentissage et de validation utilisés. C'est-à-dire que si la base de données initiale présente, par exemple, 3 observations de la classe 1 pour 7 observations de la classe 2, alors chaque ensemble de validation (resp. d'apprentissage) devra présenter ce ratio de 3 pour 7.

Dans le cas de la validation croisée à k blocs, il s'agit simplement de répartir les classes de la même manière d'un bloc à un autre. Les ensembles de validation et d'apprentissage qui en dériveront hériteront de cette répartition.

6. Evaluation du travail suivant les 5 critères précités

1. Comportement du Rmd à l'exécution

Le RMD s'execute bien mais ne comporte aucune diifciculté puisque ne nécessite pas l'installation de package au préalable. Il n'appelle pas de jeux de données.

2. Qualité de la rédaction du dossier

La rédaction de ce document est de bonne qualité. Il est bien structuré, et les étapes de progression son respectées. L'auteur est très pragmatique. Les termes employés sont bons. Il manquerait certainement des exemples théoriques mathématique ou des essais des différents modèles via des exemple sur des jeux de données.

3. Didactisme et pertinence du dossier

La lecture de ce dossier est aisée et accessible. Il est didactique. Il est pertinent puisqu'il fait partie de notre programme. Le dossier est synthétique et résume les concepts clés de l'apprentissage supervisée ou le datamining.

4. Qualité et lisibilité du RMD

Le fichier est lisible et accessible sur son GitHub. Le fichier est compilé en PDF à partir du RMD. Il est lisible et aéré. Les titres sont très bien mis en forme

5. Qualité des explications du modèle mathématique

L'auteur n'utilise aucune démonstration mathématique de ses propos. Il n'utilise pas Latex.

7. Conclusion

Selon moi, il s'agit globalement travail d'assez bonne qualité.

L'auteur synthétise et définit les différents modèles d'apprentissage. Le document est littéraire et ne comporte pas de partie mathématique. c'est dommage car l'auteur aurait pu appuyer ses propos avec des expressions mathématiques accessibles à tous. Il aurait été aussi possible de montrer les différentes méthodes de validation croisée via un dataset pré néttoyé et configuré. Cependant, ce type de document synthétique reprenant l'existant est important pour la poursuite de nos études et notre apprentissage.

Vous retrouvez ce document sur mon **GitHub**.