Exercices suggérés et projets guidés

INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

Entre les sessions

De la session 1 à la session 2

- compléter les exercices de la session 1
- télécharger les ensembles de données à partir du site web
- lire Programming Primer (DUDADS)
- installer R/ RStudio (Posit)
- installer les librairies R suivantes : dplyr, tidyverse, ggplot2, arules, arulesViz, rpart, rpart.plot, rattle, party, flexclust, e1071, psych

De la session 2 à la session 3

- compléter les exercices de la session 2

De la session 3 à la session 4

- compléter les exercices de la session 3

Après la session 4

- compléter les exercices de la session 4
- essayer les projets guidés

Projet guidé I

Ce projet utilise l'outil Gapminder

1. Dans la configuration par défaut, nous pouvons identifier quelques règles d'association potentielles. En utilisant des estimations visuelles et approximatives, évaluez la performance des règles suivantes:
   - Revenu > 8000 → Espérance de vie > 70
   - Revenu < 8000 ET Espérance de vie < 70 → Région du monde = Afrique
2. Jouez avec divers graphiques et variables et identifiez/évaluez 5+ règles d'association supplémentaires.
3. Identifiez des groupes de pays "similaires" en 2018 [veillez à valider vos groupes à l'aide de divers graphiques].
4. Dans la configuration par défaut, suivez les trajectoires de la Finlande, de la Suède, de l'Islande, de la Norvège et du Danemark entre 1900 et 2018. Les pays semblent-ils suivre des trajectoires similaires ? Y a-t-il des valeurs aberrantes ou des trajectoires anormales ?
5. Répétez l'étape 4 pour le Brésil, le Paraguay, l'Uruguay, le Venezuela, la Colombie, le Pérou, et l'Équateur.
6. D'après les résultats des étapes 4 et 5, pensez-vous que la trajectoire de l'Argentine ressemblerait davantage à celle des pays nordiques ou à celle des pays d'Amérique du Sud? Ou peut-être ni l'un ni l'autre? Votre réponse est-elle la même pour tous les horizons temporels?

Projet guidé II

Sélectionnez un ensemble de données dans la liste ci-dessous (ou un autre ensemble d'intérêt)

- GlobalCitiesPBI.csv
- 2016collisionsfinal.csv
- HR_2016_Census_simple.xlsx
- custdata.tsv

Pour votre (vos) jeu(x) de données :

1. Effectuez les étapes appropriées pour la compréhension, la préparation, le nettoyage, et l'exploration des données, afin de déterminer si elles sont dignes de confiance et à quoi elles pourraient servir (voir le projet guidé IV [*Les principes fondamentaux de la science des données*] et le projet guidé III [La *visualisation des données et les tableaux de bord*]).
2. Effectuez une analyse de règles d'association des ensembles de données, en déterminant 10 à 20 règles d'association "fortes". Visualisez-les, validez-les, et interprétez leurs résultats.

Projet guidé III

Prenons l'exemple de la base de données sur la prolifération des algues (algae_blooms.csv). Nous essayons de construire un modèle pour prédire la présence ou l'absence d'algues sur la base de diverses propriétés chimiques de l'eau de la rivière. La motivation pour un tel modèle est simple : la surveillance chimique est bon marché et facile à automatiser, alors que l'analyse biologique des échantillons est coûteuse et lente.

Une autre raison est que l'analyse des échantillons en fonction de leur contenu nocif ne permet pas de mieux comprendre les conducteurs d'algues : elle nous indique simplement quels échantillons contiennent des algues.

1. Chargez les données et résumez-les/visualisez-les : vous devez prédire la présence/l'absence des algues a1 et a2.
2. Nettoyez les données et imputez les valeurs manquantes, au besoin.
3. Retirez 20 % des observations et storez-les dans un ensemble de validation.
4. Créez une paire formation/test sur les 80 % d'observations restantes et formez deux arbres de décision pour prédire la présence/absence des algues a1 et a2, respectivement. Évaluez les performances de chaque modèle. Quels sont les modèles les plus performants sur votre paire formation/test ?

Projet guidé III (suite)

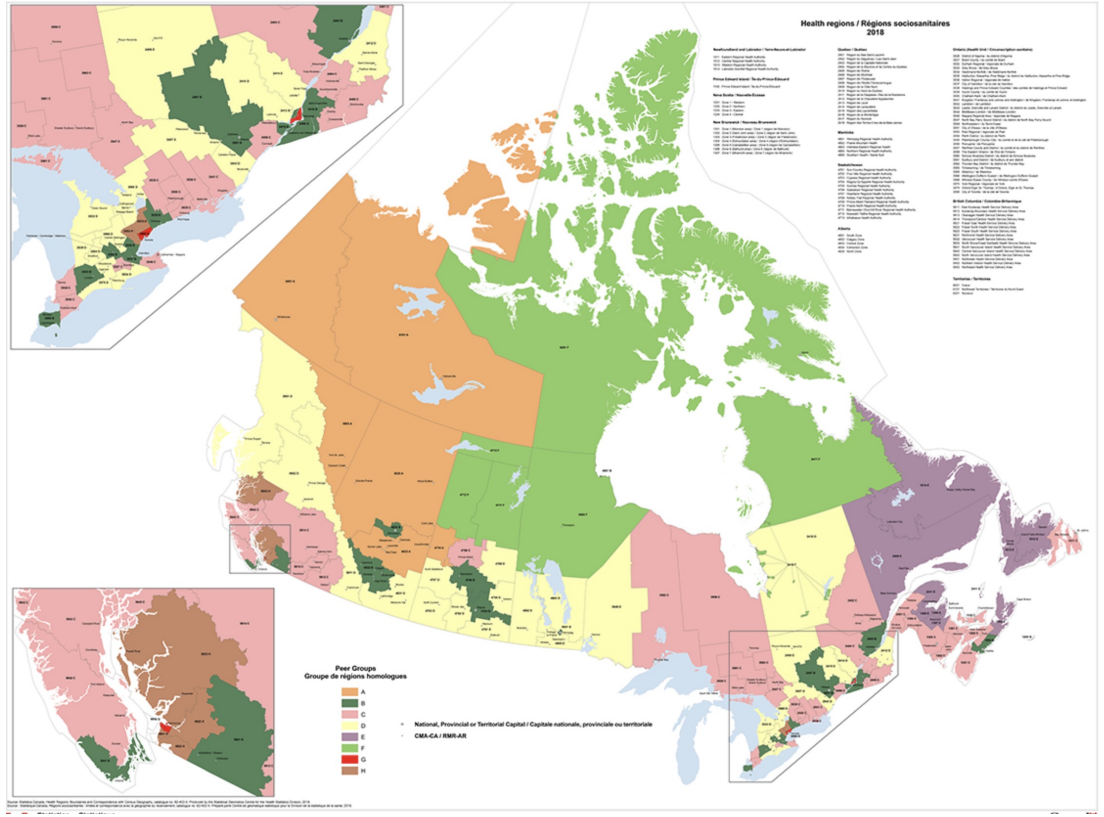
5. Répétez l'étape 4 sur au moins 20 paires formation/test distinctes. Évaluez les performances de chaque modèle et enregistrez-les.
6. Pour chaque algue, choisissez le meilleur des modèles (comment le déterminez-vous ?) et utilisez-le pour faire des prédictions pour les observations de l'ensemble de validation. Évaluez ces prédictions.
7. Au lieu de choisir le meilleur des 20+ modèles, trouvez un moyen de combiner les résultats des 20 modèles et de faire des prédictions pour les lectures de l'ensemble de validation. Évaluez ces prédictions.
8. Parmi les modèles résultant des étapes 6 ou 7, quels sont ceux qui offrent les meilleures performances ? Lequel est le plus facile à interpréter ?
9. Utilisez le même ensemble de validation qu'à l'étape 3. À l'étape 4, utilisez les 80 % de données restantes pour construire un arbre de décision (ne divisez pas d'abord en une paire formation/test). Utilisez ces modèles pour faire des prédictions pour les lectures de l'ensemble de validation. Évaluez ces prédictions. Y a-t-il des signes de sur-ajustement ?
10. Utilisez le même ensemble de validation qu'à l'étape 3. Dans les étapes 4 à 7, utilisez des souches de décision (arbres de décision à un seul point de ramification) au lieu d'arbres de croissance complets. Y a-t-il des signes de sous-ajustement ?
11. Effectuez les étapes d'analyse de 1 à 10 en utilisant d'autres algorithmes de classification. Discuter des résultats.

Projet guidé IV

La population du Canada est divisée physi-calement en régions provinciales et territoriales, dont la plupart sont subdivisées en régions de santé.

Les données du recensement (de 2016) sont disponibles pour ces régions sanitaires. L'ensemble des données équivalentes de 2018 a été regroupé pour produire des groupes de pairs : le résultat est illustré ici.

Les données se trouvent dans HR_2016_Census_simple.xlsx



Projet guidé IV (suite)

1. Chargez les données et résumez/visualisez-les (extraire les lignes avec un géocode à 4 chiffres).
2. Nettoyer les données et mettez-les à l'échelle.
3. Exécutez k −moyennes (avec la distance euclidienne) sur les données mises à l'échelle, en utilisant TOUTES les caractéristiques, pour des valeurs raisonnables de k . Utilisez l'indice de Davies-Bouldin et l'indice Within-SS pour déterminer le nombre optimal de grappes. Ce schéma de regroupement est-il plausible ?
4. Réduisez la dimension de l'ensemble de données de la région sanitaire en effectuant une analyse en composantes principales (ACP) et conservez les composantes principales qui expliquent jusqu'à 80 % de la variabilité des données. Répétez l'étape 3. Les résultats sont-ils significativement différents de ce qu'ils étaient ?
5. Exécutez k-means sur les données originales des régions de santé (question précédente) et sur les données réduites, pour la même gamme de valeurs de k, mais reproduisez le processus 30+ fois par valeur de k. Quelles sont les valeurs optimales de k dans les exécutions agrégées ?
6. Enregistrez les affectations de grappes pour chaque exécution avec les valeurs optimales de k. Deux observations A et B ont une similarité $w(A,B) \in [0,1]$ si A et B se trouvent dans la même grappe dans $w(A,B)\%$ des exécutions. Quelles sont les observations présentant des mesures de similarité élevées ? Des mesures de similarité faibles ?