TP mixte UP2 : Apprentissage statistique

Indications générales :

1. Le TP se fait impérativement en groupe de 2 à 4 personnes.
2. Le travail doit être démarré durant la séance de TP, à terminer chez soi pour être remis sur Campus avant le 16/11/2023 à 23h55.
3. Un compte rendu en format PDF doit être soumis par chaque groupe avant le 16/11/2023 à 23h55.
4. Dans le compte rendu vous présentez le code utilisé pour résoudre chaque partie ainsi que les résultats obtenus et l'interprétation détaillée des résultats le cas échéant.
5. L'évaluation est principalement sur votre capacité d'analyser, de critiquer et d'interpréter les résultats. Ainsi, il est essentiel d'expliquer clairement vos conclusions.
6. Les codes sont donnés en R-Studio. Mais si vous êtes plus à l'aise avec un autre langage, n'hésitez pas à l'utiliser.

Problème : Arbre de décision/Foret d'isolement et réduction de dimension :

À partir du répertoire en ligne « Échantillons de données » de Campus, nous considérons l'échantillon de données intitulé « Income_Inequality.csv ». Ces données, collectées par la Banque mondiale, représentent le niveau d'inégalité des revenus de différents pays sur une période s'étendant de 2010 à 2019. Les pays marqués d'un « H » sont ceux qui présentent un niveau élevé d'inégalité des revenus au cours de l'année respective, tandis que les pays marqués d'un « L » sont ceux avec une faible inégalité des revenus. L'échantillon contient un total de 870 observations réparties sur 20 variables. La variable « Income_Inequality » est celle que nous cherchons à expliquer, et les 19 autres variables explicatives sont supposées continues. Elles représentent des mesures économiques (telles que le PIB par habitant), énergétiques (comme la consommation d'électricité), financières (comme le capital en actions), de gouvernance, sociales et d'autres qui peuvent influencer la classification des pays en fonction de leur niveau d'inégalité des revenus. Nous sommes donc confrontés à un problème de classification binaire, où l'objectif est de prédire si un pays a un niveau élevé (« H ») ou faible

(« L ») d'inégalité des revenus en utilisant ces variables explicatives. Cet exercice évalue votre capacité à analyser et à construire un modèle de classification dans le contexte de données économiques et sociales réelles.

1. Analyse statistique : Faire l'étude statistique des variables quantitatives de l'échantillon : moyenne – variance écart type – p-box ...
2. Construisez un arbre de décision de haute performance qui a pour but de bien classer les observations de la variable "Income_Inequality" entre "H" ou "L". Respecter les consignes suivantes :
   (a) Diviser votre base de données en 70% pour l'apprentissage et 30% pour le test en utilisant le paramètre d'initialisation suivant "set.seed(1234)".
   (b) Optimisez votre modèle en passant par des techniques vues dans le cours (par exemple, l'optimisation des hyperparamètres par validation croisée, élagage, etc.).
   (c) Vous serez évalué sur les résultats de votre modèle optimal **appliqué aux données test** :
      i. Explorer le résultat de la classification :
         A. Detailed Accuracy By Class (Precision, Recall,...)
         B. Confusion Matrix
         C. Etc.
      ii. Veuillez inclure dans le compte rendu la courbe ROC et l'AUC de votre modèle optimal, ainsi que le code sous R correspondant.
      iii. Donnez une conclusion/interprétation globale par rapport aux résultats obtenus.
3. Maintenant, la tâche consiste à implémenter une forêt d'isolement pour calculer le score d'anomalie pour chaque observation :
   (a) Ne divisez pas les données en ensembles d'entraînement et de test.
   (b) Construisez une forêt d'isolement en utilisant l'ensemble complet de données et calculez les scores d'anomalie pour les différentes observations (en conservant les hyperparamètres tels quels, comme définis par défaut en R ou Python).
   (c) Fournissez les 10 observations ayant les scores d'anomalie les plus élevés et les 10 observations ayant les scores les plus bas. Comparez, analysez et interprétez les résultats.
   (d) Répétez la partie (2) après avoir retiré du jeu de données les 50 observations ayant les scores d'anomalie les plus élevés, puis comparez les résultats.
4. Mettre en œuvre une ACP sur $\mathbb{R}^p$:
   (a) Justifier si vous standardisez les datas ou non.
   (b) En utilisant les données train de la partie 2(a) :

i. A partir de votre rendu du TP1 ACP, vous disposez d'un script qui fait l'ACP, donc utilisez-le sur les données train (70% de la partie 2(a)).

ii. Tester la réduction par ACP : tester plusieurs règles pour sélection la part d'inertie expliquée et fournir la dimension $k < p$ de réduction retenue.

iii. Visualisez le nuage des n points dans les premiers plans retenus en mettant une couleur associée à chacune des classes.

(c) Donner la qualité de la projection des individus test (les 30% de la partie 2(a)) sur k composantes : quels sont les 10 individus les plus mal projetés dans le nuage ? quels sont leurs caractéristiques statistiques ? sont-ils les 10 individus les plus mal projetés sur les p composantes de l'ACP ? Ont-ils un rôle particulier dans le nuage ? Comparer vos résultats obtenus par les forets d'isolements.

5. Mettre en œuvre une AFD sur $\mathbb{R}^p$:

(a) En utilisant les données train de la partie 2(a) :

i. A partir de votre code sous R ou Python : fournir les matrices V matrice B et la matrice W et la matrice de variance-covariance totale Σ.

ii. Vous disposez d'un script qui vous permet de 1) calculer la diagonalisation de la matrice $\Sigma^{-1}B$ ou plus exactement obtenir les valeurs propres et les vecteurs associés U , 2) de faire la projection des individus dans ce nouvel espace. Pour finaliser l'AFD vous pouvez utiliser un de packages sous R disponibles.

iii. Tester la réduction par AFD : tester plusieurs règles pour sélectionner la part d'inertie expliquée et fournir la dimension $k < p$ de réduction retenue.

iv. Visualisez le nuage des n points dans les premiers plans retenus en mettant une couleur associée à chacune des classes.

(b) En utilisant les données test de la partie 2(a) : donner la qualité de la projection des individus test sur les k composantes : quels sont les 10 individus les plus mal projetés dans le nuage ?Comparer vos résultats obtenus par l'ACP.

6. Mise en œuvre de l'AFD prédictive :

(a) En utilisant le modèle entraîné de la partie 5(a), mettre en œuvre la fonction discriminante prédictive sur les données test et comparer avec la matrice de confusion de l'arbre de décision.

(b) Puis mettre en œuvre un arbre de décision et une AFD sur les données réduites après ACP (partie 4) et comparer les matrices de confusion sur les données tests. Comparer vos résultats obtenus par l'arbre de décision et l'AFD des parties (2) et (5).