Travaux Pratiques
Apprentissage Automatique I

TP #3 – Arbre de décision

Professeur

Abdelouahab Moussaoui

Partie 1 — Using R

Exercice 1 — with rpart

Questions :

1. Charger la bibliothèque **rpart**.
2. Chargement du jeu de données **tennis**.
3. Afficher la structure des données.
4. Créer un arbre avec un nombre minimal d'exemples nécessaires à la création d'un nœud égal à est 1, 2, 5 puis 20.
5. Visualiser pour valeur l'arbre correspondant.

Exercice 2 — with tree

Questions :

1. Charger les données **iris**.
2. Afficher les premiers enregistrements.
3. Créer puis visualiser l'arbre.
4. Visualiser la qualité de la classification.

Exercice 3 — with C5.0

Questions :

1. Charger le fichier **credit.csv.**
2. Afficher la structure des données.
3. Afficher le nombre d'occurrences des intervalles de **checking_balance** et de **savings_balance.**
4. Afficher les statistiques de **months_loan_duration.**
5. Afficher les statistiques de **amount.**
6. Donner les proportions relatives aux payements et non payement du crédit.
7. Répartir les données credit.csv en 90% pour l'apprentissage et 10% pour les tests.
8. Vérifier les proportions par la commande **prop**.**table**.
9. Installer le package **C5.0**.
10. Créer un modèle d'arbre de décision par C5.0
11. Evaluer le modèle d'apprentissage.
12. Installer la bibliothèque **gmodels.**
13. Utiliser la fonction **CrossTable** pour évaluer le modèle

Partie 2 — Using Python

Exercice 1 — Heart Attack Analysis & Prediction Dataset

Nous avons des données qui classent si les patients ont une maladie cardiaque ou non en fonction de leurs caractéristiques. Nous allons essayer d'utiliser ces données (*"heart.csv"*) pour créer un modèle qui essaie de prédire si un patient a cette maladie ou non. Nous utiliserons un algorithme de décision (classification).

Les données contiennent les informations suivantes :

- **age -** âge en années
- **sex -** (1 = masculin ; 0 = féminin)
- **cp -** type de douleur thoracique
- **trestbps -** tension artérielle au repos (en mm Hg à l'admission à l'hôpital)
- **chol -** sérum cholestoral en mg/dl
- **fbs -** (glycémie à jeun > 120 mg/dl) (1 = vrai ; 0 = faux)
- **restecg -** résultats électrocardiographiques au repos
- **thalach -** fréquence cardiaque maximale atteinte
- **exang -** angine induite par l'exercice (1 = oui ; 0 = non)
- **oldpeak -** Dépression ST induite par l'exercice par rapport au repos
- **slope -** la pente du segment ST d'exercice maximal
- **ca -** nombre de vaisseaux principaux (0-3) colorés par fluoroscopie
- **thal -** 3 = normal ; 6 = défaut corrigé ; 7 = défaut réversible
- **target -** être malade ou non (1=oui, 0=non)

Questions :

1. Charger le dataset *"heart.csv"*.
2. Afficher les données
3. Proposer deux modèles à base d'arbre de décision l'un basé sur le gain d'information et l'autre sur l'indice de Gini.
4. Comparer les deux modèles.
5. Visualiser les résultats graphiquement.

Exercice 2 — IRIS Data Classification using CART Decision Tree

Dans cette exeercice, nous utiliserons l'ensemble de données bien connu Iris Species https://en.wikipedia.org/wiki/Iris_flower_data_set créé en 1936 par le botaniste Ronald Fisher. L'ensemble de données contient des mesures de longueur et de largeur de sépale et de pétale de trois espèces de fleurs d'iris différentes *Iris setosa*, *Iris versicolor*, *Iris virginica*. Tout au long de cette activité, nous utiliserons les mesures fournies dans l'ensemble de données pour classer les différentes espèces de fleurs à l'aide d'un arbre de décision *CART*.

Questions :

1. Charger le dataset *"Iris.csv"*.
2. Afficher les données
3. Proposer deux modèles CART.
4. Visualiser les résultats graphiquement.

Exercice 3 — Credit Card Fraud detection using Decision Tree

Dans cet exercice nous voulons prédire s'il y a fraude ou non en utilisant les le dataset "creditcard.csv".

Questions :

1. Charger les données *"MNIST"*
2. Afficher quelques images.
3. Proposer deux modèles basés sur les arbres de décisions.
4. Visualiser les résultats comparatifs graphiquement.

Partie 3 — Mini-Project Using Python

Exercice 1 — Diabet Prediction with Decision Tree

Le diabète est un groupe de troubles métaboliques caractérisés par une glycémie élevée pendant une période prolongée. Les symptômes de l'hyperglycémie comprennent des mictions fréquentes, une soif accrue et une faim accrue. S'il n'est pas traité, le diabète peut entraîner de nombreuses complications. Les complications aiguës peuvent inclure une acidocétose diabétique, un état hyperglycémique hyperosmolaire ou la mort. Les complications graves à long terme comprennent les maladies cardiovasculaires, les accidents vasculaires cérébraux, les maladies rénales chroniques, les ulcères du pied et les lésions oculaires.

✓ *Ensemble de données et histoire*

Cet ensemble de données provient à l'origine de l'Institut national du diabète et des maladies digestives et rénales. L'objectif de l'ensemble de données est de prédire de manière diagnostique si un patient est diabétique ou non, sur la base de certaines mesures diagnostiques incluses dans l'ensemble de données. Plusieurs contraintes ont été placées sur la sélection de ces instances à partir d'une plus grande base de données. En particulier, tous les patients ici sont des femmes d'au moins 21 ans d'origine indienne Pima.

✓ *Grossesses : Nombre de fois enceintes*

Pregnancies : Nombre de fois enceintes
Glucose : Glucose
BloodPressure : tension artérielle
SkinThickness : Épaisseur du pli cutané du triceps
Insulin : Insuline
BMI : indice de masse corporelle
DiabetesPedigreeFunction : fonction d'arbre généalogique du diabète
Age : Âge (ans)
Outcome : Savoir s'il y a du diabète (c'est notre objectif)

Questions :

1. Charger les données *"diabetes.csv"*
2. Afficher quelques images.
3. Proposer deux modèles basés sur les arbres de décisions.
4. Comparer ces deux modèles avec la régression logistique
5. Visualiser les résultats comparatifs graphiquement.
6. Conclusion

Exercice 2 — Titanic Disaster Prediction with Decision Tree

Le naufrage du Titanic est l'un des naufrages les plus tristement célèbres de l'histoire.

Le 15 avril 1912, lors de son voyage inaugural, le RMS Titanic largement considéré comme "insubmersible" a coulé après être entré en collision avec un iceberg. Malheureusement, il n'y avait pas assez de canots de sauvetage pour tout le monde à bord, ce qui a entraîné la mort de 1502 des 2224 passagers et membres d'équipage.

Bien qu'il y ait eu un élément de chance dans la survie, il semble que certains groupes de personnes étaient plus susceptibles de survivre que d'autres.

Dans ce défi, nous vous demandons de construire un modèle prédictif qui répond à la question : "quels types de personnes étaient les plus susceptibles de survivre ?" en utilisant les données des passagers (c'est-à-dire le nom, l'âge, le sexe, la classe socio-économique, etc.).

Pour le travail sur les données Titanic, nous devons deviner si les individus de l'ensemble de données de test ont survécu ou non. Mais pour notre objectif actuel, découvrons également ce que les données peuvent nous dire sur le naufrage à l'aide d'un arbre de classification. Chargeons les données et obtenons un aperçu.

✓ *Vue d'ensemble de la description de l'ensemble de données*

Les données ont été divisées en deux groupes :

- ✓ Ensemble de formation (*train.csv*)
- ✓ Ensemble de test (*test.csv*)

L'ensemble de formation doit être utilisé pour créer vos modèles d'apprentissage automatique. Pour l'ensemble de formation, nous fournissons le résultat (également connu sous le nom de «vérité terrain») pour chaque passager. Votre modèle sera basé sur des "caractéristiques" telles que le sexe et la classe des passagers. Vous pouvez également utiliser l'ingénierie des fonctionnalités pour créer de nouvelles fonctionnalités.

L'ensemble de test doit être utilisé pour voir dans quelle mesure votre modèle fonctionne sur des données invisibles. Pour l'ensemble de test, nous ne fournissons pas la vérité terrain pour chaque passager. C'est votre travail de prédire ces résultats. Pour chaque passager de l'ensemble de test, utilisez le modèle que vous avez formé pour prédire s'il a survécu ou non au naufrage du Titanic.

Nous incluons également *gender_submission.csv*, un ensemble de prédictions qui supposent que toutes les passagères et seulement des femmes survivent, comme exemple de ce à quoi un fichier de soumission devrait ressembler.

Questions :

1. Charger les données du *titanic.*
2. Afficher les informations concernant les données.
3. Proposer deux modèles basés sur les arbres de décisions.
4. Comparer ces deux modèles avec la régression logistique
5. Visualiser les résultats comparatifs graphiquement.
6. Conclusion

Exercice 3 — MNIST Multi-classes classification with Decision Tree

Ici on veut classer les chiffres décimaux de la base MNIST en utilisant les arbres de décisions.

Questions :

1. Charger les données du *MNIST.*
2. Afficher les informations concernant les données.
3. Proposer deux modèles basés sur les arbres de décisions.
4. Comparer ces deux modèles avec la régression logistique
5. Visualiser les résultats comparatifs graphiquement.
6. Conclusion