

**Rapport de projet de Python**

Analyse Discriminante Linéaire

*Sous la direction de M. Ricco Rakotomalala*

*Réalisé par : Aleksandra KRUCHININA, Aymeric DELEFOSSE et Mamadou DIALLO*

L3 MIASHS – IDS 2020-2021

Sommaire

[Introduction 4](#_Toc61298837)

[Module : calculations.py 4](#_Toc61298838)

[verification\_NA 4](#_Toc61298839)

[recodification\_var\_expl 4](#_Toc61298840)

[freq\_relat 5](#_Toc61298841)

[means\_class 5](#_Toc61298842)

[cov\_matrix 5](#_Toc61298843)

[pooled\_cov\_matrix 5](#_Toc61298844)

[wilks 6](#_Toc61298845)

[wilks\_log 6](#_Toc61298846)

[p\_value 7](#_Toc61298847)

[Module : reporting.py 8](#_Toc61298848)

[Classe HTML (Objet) 8](#_Toc61298849)

[Méthodes : 8](#_Toc61298850)

[create\_html\_head 8](#_Toc61298851)

[stepdisc\_html\_output 8](#_Toc61298852)

[discrim\_html\_output 9](#_Toc61298853)

[discrim\_html\_output\_datapane 9](#_Toc61298854)

[Classe PDF(FPDF) (Objet) 9](#_Toc61298855)

[footer 9](#_Toc61298856)

[discrim\_pdf\_output 9](#_Toc61298857)

[Module : discriminant\_analysis.py 11](#_Toc61298858)

[Classe LinearDiscriminantAnalysis (Objet) 11](#_Toc61298859)

[Attributs 11](#_Toc61298860)

[Méthodes 12](#_Toc61298861)

[\_\_init\_\_ 12](#_Toc61298862)

[\_stats\_dataset 12](#_Toc61298863)

[\_stats\_pooled\_cov\_matrix 12](#_Toc61298864)

[\_stats\_classes 12](#_Toc61298865)

[\_stats\_wilks 12](#_Toc61298866)

[fit 13](#_Toc61298867)

[predict 14](#_Toc61298868)

[confusion\_matrix 14](#_Toc61298869)

[accuracy\_score 15](#_Toc61298870)

[wilks\_decay 15](#_Toc61298871)

[stepdisc 15](#_Toc61298872)

# Introduction

Pour réaliser notre application, nous avons choisi de créer une classe LDA, LinearDiscriminantAnalysis. Cette solution permet de créer des instances, de mieux structurer et d’organiser le code. De plus, cela permet d’utiliser les variables calculées dans certaines fonctions (qui n’étaient pas utilisées dans leurs retours) à partir d’autres fonctions. Par exemple, l’attribut **coef\_** qui est construit dans la fonction **fit()** et qui est utilisé ultérieurement dans la fonction **predict().**

Le code est divisé en trois modules : **calculations**, **reporting** et **discriminant\_analysis**.

# Module : calculations.py

C’est le module qui regroupe des fonctions de vérification et de modification de données ainsi que des fonctions permettant d'effectuer les calculs nécessaires pour la pratique de l'analyse discriminante linéaire (lambda de Wilks, matrice de covariance, etc.).

## verification\_NA

On a repris l’approche de la procédure DISCRIM de SAS : les valeurs nulles ne sont pas prises en compte pour l’analyse, donc les observations précédant ces valeurs sont annulées.

**Fonction verification\_NA(inputData : matrice de réels, targetValues  : vecteur de réels) :**

**Début**

Définir la taille de l’échantillon et le nombre de variables explicatives

Concaténation de inputData et targetValues

Suppression des observations/lignes avec les valeurs nulles

Calcul du nombre de lignes supprimées

Redéfinition de X et y sans lignes supprimées

Affichage du nombre de lignes supprimées

Renvoyer nouveaux X et y

**Fin**

## recodification\_var\_expl

La fonction *recodification\_var\_expl()* de notre module **calculations**  permet de convertir des variables qualitatives en variables quantitatives. L’Analyse de Discriminante Linéaire (ADL), ne cadrant tout simplement pas avec les variables qualitatives, nous oblige de recoder ces variables avant l’application de cette méthode statistique sur nos données. Plusieurs approches existent afin de recoder des variables qualitatives. Parmi elles, nous trouvons la méthode DISQUAL ainsi que la méthode des *dummy variables* (celle que nous utiliserons), pour ne citer que ces deux-là.  La méthode des *dummy variables* (Ricco Rakotomalala, *Pratique de l’Analyse Discriminante Linéaire*, section 3.3, p.76) octroie M-1 variables à chaque variables qualitatives possédant M modalités (notion modalité de référence). À ceux-ci, nous attribuons soit 0, soit 1, représentant la présence ou non de catégories.

À savoir :

* Cette fonction n’est utile qu’en cas de présence de **variables qualitatives** dans les **variables explicatives** des données ;
* Si nous sommes en présence que de variables quantitatives, il faut directement implémenter la fonction **fit()** de la classe LDA.

**Fonction recodification\_var\_expl(inputData : matrice de réels) :**

**Début**

Création d’un dictionnaire vide et lui

Attribuer des valeurs soit vraie ou fausse selon que la variable est positive ou pas

Créer une liste de valeurs booléen

Convertir le type de la liste en type object de python si sa valeur égale à faux

Recoder les variables de type object en tenant compte de la modalité de référence

Jointure des données initiales avec les données recodées

Suppression de variables non numériques de nos données finales

Retourner les données recodées

**Fin**

## freq\_relat

La fonction calcule les fréquences relatives.

**Fonction freq\_relat(targetValues : matrice de réels, n : integer):**

**Début**

Calcul du nombre d'effectifs par classes

Calcul des fréquences relatives par classes

Renvoyer ces deux vecteurs

**Fin**

## means\_class

La fonction calcule des moyennes conditionnelles selon le groupe d'appartenance.

**Fonction means\_class(inputData : matrice de réels, targetValues : vecteur de réels):**

**Début**

Calcul du nombre d'effectifs par classe

Initiation d'une matrice remplie de 0

Remplir cette matrice avec les effectifs par classe

Division par le nombre d'effectifs par classe

Renvoyer les moyennes conditionnelles

**Fin**

## cov\_matrix

La fonction calcule la matrice de covariance totale (V) et biaisée (Vb) entre les variables explicatives directement à partir du jeu de données.

Cette fonction a recours à la fonction [**cov()**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.cov.html) de la librairie **pandas**.

**Fonction cov\_matrix(dataset : pandas DataFrame) :**

**Début**

Définir la taille de l’échantillon

Calcul la matrice de covariance totale

Calcul la matrice de covariance biaisée

Renvoyer les deux matrices

**Fin**

## pooled\_cov\_matrix

Cette fonction permet de calculer la matrice de covariance intra-classe totale (W) et biaisée (Wb).

D’après Ricco Rakotomalala, dans son livre sur la *Pratique de l’Analyse Discriminante Linéaire*, « en pratique, la matrice de covariance commune est estimée à l’aide de la matrice de variance covariance intra-classe » (section 1.2, p.25). Le calcul pour obtenir la matrice de covariance intra-classe est donc :

Où :

* N : taille de l’échantillon ;
* K : nombre de classes ;
* Vk : matrice de covariance pour la classe k.

**Fonction pooled\_cov\_matrix(dataset : pandas.DataFrame, className : string) :**

**Début**

Définir la taille de l’échantillon

Définir le nombre de classes

Initiation de la matrice

Parcourir les modalités des valeurs uniques de la variable cible :

Définir le nombre de valeurs par classe

Calcul de la matrice de la covariance des moyennes conditionnelles

Calcul des sommes de ces matrices selon la formule

Fin de la boucle Pour

Calcul de la matrice de covariance intra-classe totale

Calcul de la matrice de covariance intra-classe biaisée

Renvoyer les matrices de covariance intra-classe totale et biaisée

**Fin**

## wilks

La fonction calcule les valeurs du lambda de Wilks où :

Avec :

Ces deux matrices sont obtenues respectivement grâce aux fonctions **pooled\_cov\_matrix** et **cov\_matrix**, vues précédemment.

Cette fonction a recours à la fonction [**det()**](https://numpy.org/doc/stable/reference/generated/numpy.linalg.det.html) de la librairie de calcul matriciel [**numpy.linalg**](https://numpy.org/doc/stable/reference/routines.linalg.html).

**Fonction wilks(Vb : matrice de réels, Wb : matrice de réels) :**

**Début**

Calcul du déterminant de la matrice de covariance totale biaisée

Calcul du déterminant de la matrice de covariance intra-classes biaisée

Calcul les valeurs du lambda de Wilks

**Fin**

## wilks\_log

La fonction calcule les valeurs du lambda de Wilks à partir des logarithmes naturels des déterminants, où :

Cela permet de contourner les limitations des librairies utilisées dans les calculs scientifiques. En effet, si en entrée, les matrices sont trop grandes, le calcul standard avec la fonction **wilks()** ne sera pas en mesure d’aboutir.

Cette fonction a recours à la fonction [**slogdet()**](https://numpy.org/doc/stable/reference/generated/numpy.linalg.slogdet.html) de la librairie de calcul matriciel [**numpy.linalg**](https://numpy.org/doc/stable/reference/routines.linalg.html).

**Fonction wilks\_log(Vb : matrice de réels, Wb : matrice de réels) :**

**Début**

Calcul du logarithme naturel du déterminant de la matrice de covariance totale biaisée

Calcul du logarithme naturel du déterminant de la matrice de covariance intra-classes biaisée

Calcul du Lambda de Wilks par le rapport entre les logarithmes naturels des déterminants

**Fin**

## p\_value

La fonction calcule la p-valeur d’un test unilatéral de Fisher.

La fonction a recours à la librairie [**scipy.stats**](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f.html).

**Fonction p\_value(F : réel, ddl1 : entier, ddl2 : entier) :**

**Début**

Si la statistique de Fisher < 1 alors :

Renvoyer p\_value selon formule1

Fin Si

Renvoyer p\_value selon formule2

**Fin**

# Module : reporting.py

Ce module regroupe les différentes classes pour la création de reporting automatique, en format HTML et PDF, dont les sorties ressemblent à celles des PROC DISCRIM et STEPDISC de SAS.

## Classe HTML (Objet)

## Méthodes :

### create\_html\_head

Permet de créer le début d’un fichier HTML. La feuille de style est directement issue d’une feuille de style toute faite très populaire : [Bootstrap](https://getbootstrap.com).

Voilà ce que le fichier donne lorsqu’on lance simplement cette fonction. Notons qu’elle prend en paramètre une chaîne de caractères « proc », qui permet de changer le titre de la page avec le nom de la fonction souhaitée (STEPDISC ou DISCRIM).

<!DOCTYPE html>

<html lang="fr" dir="ltr">

  <head>

    <title>R&#233;sultats : %s</title>

    <meta charset="utf-8" />

    <style></style>

    <link

      rel="stylesheet"

      href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.2/css/bootstrap.min.css"

      integrity="sha384-JcKb8q3iqJ61gNV9KGb8thSsNjpSL0n8PARn9HuZOnIxN0hoP+VmmDGMN5t9UJ0Z"

      crossorigin="anonymous"

    />

  </head>

  <body>

    <div class="container text-center">

      <h2>Proc&#233;dure %s</h2>

### stepdisc\_html\_output

Ecriture des étapes de la sélection de variables (proc stepdisc) dans un fichier **.html**.

Les **pandas.DataFrame** sont transformés en tables HTML grâce à la fonction [**to\_html()**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_html.html).

**Fonction stepdisc\_html\_output(ProcStepdisc : objet LinearDiscriminantAnalysis,**

**fileName : chaîne)**

**Début**

Si (les 5 derniers caractères de fileName) sont différents de ".html " :

Ajouter ".html" à la fin du nom du fichier

Fin Si

Ouverture du fichier

Création du début du fichier grâce à la fonction "create\_html\_head("STEPDISC")"

Ecriture des étapes de la procédure de sélection de variables

Fermeture du fichier

**Fin**

### discrim\_html\_output

Ecriture des étapes de la fonction **fit()** (proc discrim) un fichier **.html**, style basé sur **stepdisc\_html\_output**.

Les **pandas.DataFrame** sont transformés en tables HTML grâce à la fonction [**to\_html()**](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_html.html).

**Fonction discrim\_html\_output(ProcStepdisc : objet LinearDiscriminantAnalysis,**

**fileName : chaîne)**

**Début**

Si (les 5 derniers caractères de fileName) sont différents de ".html " :

Ajouter ".html" à la fin du nom du fichier

Fin Si

Appel des différents attributs de l’instance de la classe LinearDiscriminantAnalysis

contenant l’information qu’on a besoin de récupérer pour restituer les sorties

Ouverture du fichier

Création du début du fichier grâce à la fonction "create\_html\_head("STEPDISC")"

Ecriture des étapes de la procédure de sélection de variables

Fermeture du fichier

**Fin**

### discrim\_html\_output\_datapane

Idem que discrim\_html\_output mais le style est différent, car basé sur la librairie [**datapane**](https://datapane.com).

**Fonction discrim\_html\_output\_datapane(ProcStepdisc : objet LinearDiscriminantAnalysis,**

**fileName : chaîne)**

**Début**

Si (les 5 derniers caractères de fileName) sont différents de ".html " :

Ajouter ".html" à la fin du nom du fichier

Fin Si

Appel des différents attributs de l’instance de la classe LinearDiscriminantAnalysis

contenant l’information qu’on a besoin de récupérer pour restituer les sorties

Ouverture et écriture des différents attributs de la classe dans le fichier en

format .html grâce à datapane

Fermeture du fichier

**Fin**

## Classe PDF(FPDF) (Objet)

Cette classe hérite la classe FPDF du package fpdf. Inspiré par le tutoriel : https://pyfpdf.readthedocs.io/en/latest/Tutorial/index.html

## Méthodes :

### footer

Footer facilite l'affichage du bas de page lors d’une création d’un fichier pdf.

### discrim\_pdf\_output

Cette méthode créée un fichier pdf, dont l’utilisateur indique le nom. Il contient des sorties ressemblant aux sorties de la procédure DISCRIM de SAS : la description générale de l’échantillon (Taille d'échantillon totale, le nombre des variables et des classes, les DDL), l’information au niveau classe (les fréquences et proportions des classes), l’informations sur la matrice de covariance combinée, les statistiques de lambda de Wilks, les coefficients et l’intercept construits par le modèle.

La matrice de confusion et le taux de précision ne sont pas présentes dans cette méthode car on peut les invoquer par des méthodes séparées - confusion\_matrix() et accuracy\_score().

**Fonction discrim\_pdf\_output(ProcDiscrim : objet LinearDiscriminantAnalysis,**

**fileName : string) :**

**Début**

Si les 4 derniers caractères du nom du fichier ne contiennent pas ".pdf" :

Concaténer le nom du fichier avec ".pdf"

Fin Si

Appel des différents attributs de l’instance de la classe LinearDiscriminantAnalysis

contenant l’information qu’on a besoin de récupérer les sorties

Création d’une instance de la classe PDF

Définition des caractères des polices pour le titre

Ajout du titre du paragraphe

Ligne rouge

Définition des caractères des polices pour le texte

Parcours en récupérant l’index et l’élément de l’information général des attributs de ProcDiscrim :

Ajout des cellules de pdf contenant les informations selon l’attribut du modèle

choisi

Ligne rouge

Fin Boucle

Idem pour les statistiques des classes, l’information sur la matrice de covariance,

sur la fonction de classement, sur le lambda de Wilks

Sauvegarde du fichier

**Fin**

# Module : discriminant\_analysis.py

## Classe LinearDiscriminantAnalysis (Objet)

Ce module est le module principal et le seul dont l’utilisateur aura besoin pour pratiquer l’Analyse Discriminante Linéaire. Après avoir chargé les données (grâce à pandas) et avoir repéré le nom de la variable catégorielle, l’utilisateur devra créer un objet issu de cette classe en spécifiant les deux paramètres.

**LDA = LinearDiscrimnantAnalysis(jeu\_de\_données, nom\_de\_la\_classe)**

À partir de là, l’utilisateur pourra directement appeler les différentes méthodes pour réaliser les calculs qu’ils souhaitent et les différents attributs s’il veut récupérer les paramètres precis.

### Attributs

* **dataset** : matrice *# le jeu de données*
* **classEtiquette** : string *# le nom de la variable cible*
* **classNames** : liste de chaînes *# les noms des valeurs prises pour la variable cible*
* **varNames** : liste de chaînes *# les noms des variables explicatives*
* **n** : entier *# la taille de l'échantillon*
* **p** : entier *# le nombre de variables explicatives*
* **K** : entier *# le nombre de classes*
* **V** : matrice de réels *# les matrices de covariance totale*
* **Vb** : matrice de réels *# les matrices de covariance biaisée*
* **W** : matrice de réels *# la* *matrice* *de covariance intra-classe*
* **Wb** : matrice de réels *# la matrice de covariance intra-classe biaisée*
* **infoDataset** : pandas.DataFrame *# l’informations de bases sur le jeu de données*
* **infoCovMatrix** : pandas.DataFrame *# les statistiques de la matrice de covariance intra-classe*
* **infoClasses** : pandas.DataFrame *# les effectifs et fréquences relatives des classes*
* **infoWilksStats** : pandas.DataFrame *# les statistiques du Lambda de Wilks*
* **intercept**\_ : vecteur de réels *# l’intercept calculé par le modèle*
* **coef\_** : matrice de réels *# les coefficients du modèle*
* **infoFuncClassement** : pandas.DataFrame *# les valeurs de la fonction de classement*
* **confusionMatrix** : matrice de réels *# la matrice de confusion*
* **confusionMatrixGraph** : graphique *# graphique de la matrice de confusion*
* **accuracy** : réel *# le taux de précision*
* **infoWilksDecay :** liste de réels *# liste des valeurs du lambda de wilks en fonction du nombre de variables sélectionnées*
* **figWilksValue :** graphique *# courbe de la décroissance du lambda de Wilks*
* **\_htmlStringOutput :** string *# pour la sortie HTML de la sortie STEPDISC*
* **infoStepResults :** pandas.DataFrame *# lorsque l’approche de sélection de variables, affiche les différentes valeurs pour la dernière étape*
* **stepdiscSummary :** pandas.DataFrame *# résumé de l’approche de sélection de variables*

### Méthodes

Note : L’utilisation d’un **underscore** devant les méthodes à « weak internal usage » permet à ces méthodes de ne pas être convoquées lors de l’importation de la librairie   
(import \* from LinearDiscriminantAnalysis).

#### \_\_init\_\_

Création de l’objet à l’aide du constructeur python et affectation des différents attributs de base pour le bon fonctionnement des méthodes.

Prend en paramètres le jeu de données et le nom de la variable catégorielle. Il est également possible d’indiquer le nom des variables explicatives.

A la création de l’objet, on retrouvera les attributs suivants : **dataset**, **classEtiquette**, **classNames**, **varNames**, **n**, **p**, **K**, **V**, **Vb**, **W**, **Wb**.

#### \_stats\_dataset

Sert à créer l’attribut infoDataset qui représentent une matrice contenant l’information suivante : "Taille d'échantillon totale", "Variables", "Classes", "Total DDL", "DDL dans les classes", "DDL entre les classes".

#### \_stats\_pooled\_cov\_matrix

Cette méthode calcule des statistiques de la matrice de covariance intra-classes

**Fonction \_stats\_pooled\_cov\_matrix(self.W) :**

**Début**

Calcul du rang de la matrice de cov. intra-classes

Calcul du logarithme naturel du déterminant de cette matrice

Restitution de l’attribut infoCovMatrix qui contient "Rang de la mat. de cov. intra-classes" et "Log. naturel du det. de la mat. de cov. intra-classes"

**Fin**

#### \_stats\_classes

Cette méthode calcule des statistiques des classes.

**Fonction \_stats\_classes () :**

**Début**

Restitution de la variable cible

Calcul des effectifs et fréquences relatives des classes en utilisant la fonction du module calculations

Restitution de l’attribut infoClasses qui contient "Effectifs" et "Fréquences"

**Fin Fonction**

#### \_stats\_wilks

Cette méthode calcule les statistiques du Lambda de Wilks.

**Fonction \_stats\_wilks () :**

**Début**

Calcul du lambda de Wilks

Calcul du ddl du numérateur

Calcul du ddl du dénominateur

Calcul de la F-statistique

Calcul de la p\_value

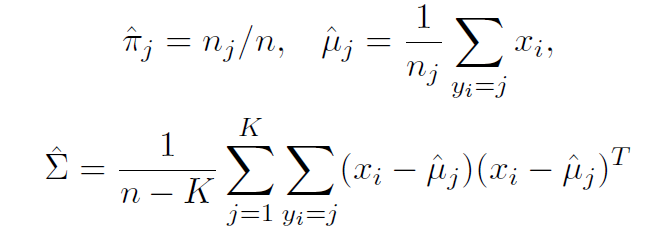
Restitution de l’altribut infoWilksStats qui contient "Valeur", "F-Valeur", "DDL num.", "DDL den.", "p-value"

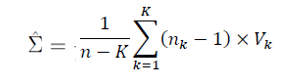
**Fin Fonction**

#### fit

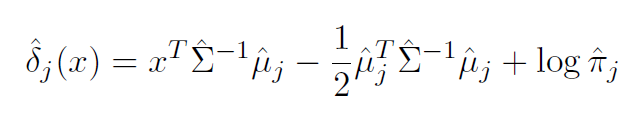
Les valeurs nulles ne sont pas prises en compte et sont supprimées. Le message affiche le nombre d’observations supprimées. La methode renvoie les coefficients et l’intercept du modèle.

D’abord on calcule les moyennes conditionelles, les frequences relatives et la matrice de variance co-variance intra-classes[[1]](#footnote-1) :

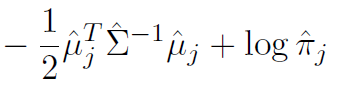




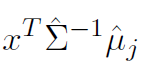
Après on estime la fonction de classement par la formule :



Où l’intercept :



Et les coefficients :



**Fonction fit () :**

**Début**

Sélection des valeurs des variables explicatives et cible

Vérification et suppression des valeurs nulles de dataset

Transformation des données en numpy

Calcul des effectifs et des fréquences de classes

Calcul des moyennes conditionnelles

Si le déterminant de la matrice de var-covar intra-classes (W) ≠ 0 :

Calcul de la matrice inverse de W

Calcul des coefficients et de l’intercept

Récupération des valeurs de la fonction de classement

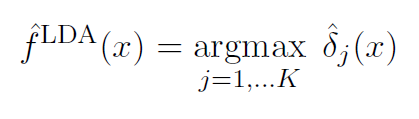
Sinon :

Afficher un message d’erreur que la matrice n’est pas inversible

Fin Si

**Fin**

#### predict

Pour calculer les prédictions il faut prendre le maximum parmi les scores d’appartenance à chaque classe. Ainsi on a finalement la formule pour calculer les prédictions : 

La fonction retourne le vecteur des classes prédites.

**Fonction predict(inputData : matrice des valeurs à prédire) :**

**Début**

Création d’une liste vide pour la remplir avec les prédictions

Parcours la matrice inputData

Calcul des probabilités d’appartenir aux classes

Stockage de ces valeurs dans une liste

Prédictions des classes en choisissant le maximum parmi les probabilités

**Fin**

#### confusion\_matrix

Avoir les prédictions c’est très bien, mais comment on peut savoir ou plutôt mesurer si ces prédictions sont correctes ou pas ? La matrice de confusion et le taux de précision y aident. La première « permet d’obtenir une évaluation non biaisée des performances du modèle en déploiement »[[2]](#footnote-2). Elle prend le vecteur des vraies valeurs de la variable cible qui n’étaient pas utilisées lors de l’apprentissage, et le vecteur des valeurs cibles prédites. Ainsi on peut estimer si la fonction est capable de bien prédire.

« Chaque ligne correspond à une classe réelle, chaque colonne correspond à une classe estimée. La cellule ligne L, colonne C contient le nombre d'éléments de la classe réelle L qui ont été estimés comme appartenant à la classe C »[[3]](#footnote-3).

Cette matrice permet de mesurer plusieurs d’autres indices. Mais comme cela n’est pas l’objectif de notre projet, on a retenu que la matrice de confusion et le taux de précision.

La fonction renvoie la matrice numérique.

En analyse des données c’est toujours utile de construire des graphiques ou d’avoir les données en format plus lisible. Le paramètre de la méthode graphShow est mis à True par défaut, cela nous affiche la matrice en mode plus représentatif (grâce à la librairie seaborn). Si on ne veut pas l’afficher, il faut mettre ce paramètre à False.

**Fonction confusion\_matrix(y\_true : vecteur des vrais valeurs de la cible, y\_pred : vecteur des valeurs de la cible à prédire, graphShow : booléen, True par défaut)**

**Début**

Conversion des noms des classes en numérique

Création d’une matrice zéro de taille K x K

Parcours des indices de 0 à K-1 (deux boucles imbriquées) :

Calcul des valeurs de la matrice de confusion

Si le paramètre graphShow est True :

Créer et afficher la matrice de confusion en mode graphique

Renvoyer la matrice de confusion

Fin Si

**Fin**

#### accuracy\_score

La fonction prend en entrée les vraies valeurs de la variable cible et les prédictions et renvoie la proportion des prédictions correctes.

Accuracy = (True Positive + True Negative) / (Positives + Negatives)

**Fonction accuracy\_score(y\_true : vecteur des vrais valeurs de la cible, y\_pred : vecteur des valeurs de la cible à prédire):**

**Début**

Conversion des noms des classes en numérique

Calcul du taux de précision

Renvoyer le taux de précision

**Fin**

#### wilks\_decay

Permet de calculer la valeur du lambda de Wilks en fonction du nombre de variables explicatives sélectionnées et d’afficher la courbe du décroissance du lambda de Wilks. Par exemple, s’il y a 40 variables explicatives, les valeurs du lambda seront pour {1, 2, …, 40}.

**Fonction wilks\_decay(graphShow : booléen, True par défaut) :**

**Var**

On crée une liste vide qui contiendra les valeurs du lambda de Wilks

On crée une liste vide qui contiendra les noms de variables (varSelection)

Pour chaque nom dans les noms des variables explicatives :

On ajoute le nom de cette variable dans la liste (varSelection)

Calcul du Lambda de Wilks en sélectionnant les lignes et colonnes de Vb et Wb

par rapport à la liste créée (varSelection) et ajout de la valeur dans une

liste (wilksValues)

Fin Pour

Si graphShow est mis à True :

Créer et affiche le courbe de la décroissance du lambda de Wilks

Fin Si

**Fin**

#### stepdisc

Cette méthode permet de réaliser une méthode de sélection de variable. Elle est basée sur la section 4 livre sur la *Pratique de l’Analyse Discriminante* *Linéare* de Ricco Rakotomalala (pp. 87-108) et sur la PROC STEPDISC de SAS.

Rappel des calculs utilisés dans la fonction :

Approche ascendante :

Approche descendante :

Elle permet à l’utilisateur d’aborder une approche ascendante, ou descendante.

**Fonction stepdisc(slentry : réel, method : chaîne)**

**Début**

**Si le paramètre de la méthode est forward :**

On initialise le lambda de Wilks à 1 (car )

**Pour q allant de 1 à p :**

**Pour chaque nom dans les noms des variables explicatives :**

On ajoute le nom de la variable dans la liste wilksVarSelect

On calcule le lambda de Wilks sur cette liste de variables

(en faisant une sélection des lignes et colonnes de Vb et Wb)

On calcule les degrés de libertés (ddl1 := K-1,ddl2 := n-K-q)

Calcul de la statistique de Fisher

Calcul du R²(partiel)

Calcul de la p-valeur du test de Fisher

Ajout de ces informations dans une liste

**Fin Pour**

Récupération des valeurs calculées pour chaque variable dans la boucle

dans un DataFrame (information sur l’étape)

Récupération du nom de la variable qui maximise F (qui minimise le

Lambda de Wilks)

Ecriture du détail de l’étape dans une chaîne de caractères pour faire

la sortie html plus tard

**Si la p-valeur de la variable maximisant F est supérieure à la**

**condition d’arrêt  :**

Ecriture du détail de l’étape dans une chaîne de caractères

pour faire la sortie html plus tard

Arrêt et sortie de la boucle

**Sinon :**

Ajout de la variable dans la liste des variables en entrée

On enlève le nom de la variable des noms des variables

explicatives

On stocke les valeurs de la valeur choisie dans un DataFrame

On initialise le lambda de Wilks à la valeur du lambda de

cette variable ()

Ecriture du détail de l’étape dans une chaîne de caractères

pour faire la sortie html plus tard

**Fin Si**

Création d’un dataframe regroupant les informations de chaque variable

choisie par la sélection ascendante

Ecriture du détail de l’étape dans une chaîne de caractères

pour faire la sortie html plus tard

**Fin Pour**

**Sinon si method = “backward”:**

On initialise le lambda de Wilks pour q = p à l’aide de la fonction wilks\_log

**Pour q allant de p à 1 :**

**Pour chaque nom dans les noms des variables explicatives :**

On ajoute le nom de la variable dans la liste wilksVarSelect

On calcule le lambda de Wilks sur cette liste de variables

(en faisant une sélection des lignes et colonnes de Vb et Wb)

On calcule les degrés de libertés (ddl1 := K-1,ddl2 := n-K-q)

Calcul de la statistique de Fisher

Calcul du R²(partiel)

Calcul de la p-valeur du test de Fisher

Ajout de ces informations dans une liste

**Fin Pour**

Récupération des valeurs calculées pour chaque variable dans la boucle

dans un DataFrame (information sur l’étape)

Récupération du nom de la variable qui minimise F

Ecriture du détail de l’étape dans une chaîne de caractères pour faire

la sortie html plus tard

**Si la p-valeur de la variable maximisant F est inférieure à la**

**condition d’arrêt  :**

Ecriture du détail de l’étape dans une chaîne de caractères

pour faire la sortie html plus tard

Arrêt et sortie de la boucle

**Sinon :**

Ajout de la variable dans la liste des variables retirées

On enlève le nom de la variable des noms des variables

explicatives

On stocke les valeurs de la valeur éliminée dans un DataFrame

On initialise le lambda de Wilks à la valeur du lambda de

cette variable ()

Ecriture du détail de l’étape dans une chaîne de caractères

pour faire la sortie html plus tard

**Fin Si**

Ecriture du résumé des variables choisies ou éliminées dans un DataFrame

Ecriture du détail de l’étape dans une chaîne de caractères

pour faire la sortie html plus tard

**Fin Pour**

**Fin Si**

**Fin**

1. https://www.stat.cmu.edu/~ryantibs/datamining/lectures/21-clas2.pdf [↑](#footnote-ref-1)
2. http://eric.univ-lyon2.fr/~ricco/cours/cours/Pratique\_Analyse\_Discriminante\_Lineaire.pdf [↑](#footnote-ref-2)
3. https://fr.wikipedia.org/wiki/Matrice\_de\_confusion [↑](#footnote-ref-3)