Apprentissage statistique

Projet

Promo 2024

Auteurs :
Lucien Perdrix
Juliette Limes

Table des matières

Introduction

Le jeu de données **Music_2023.txt** est extrait d'un challenge de reconnaissance de genre de musique, et ne contient que des morceaux de jazz et de musique classique. Il s'agit de mettre en compétition différentes méthodes pour différencier ces deux genres. Le jeu de données est plus précisément décrit en annexe.

1 Partie I : Régression logistique

Dans cette première partie, nous allons réaliser une régression logistique pour la classification.

1.1 Question 1

On effectue les analyses univariée et bivariée.

On trouve des proportions de 47% et de 53% respectivement pour le Jazz et la musique classique.

Il peut être judicieux d'appliquer une transformation log aux variables PAR_SC_V et PAR_ASC_V afin de normaliser les valeurs. On peut de plus supprimer les variables numérotées de 148 à 167 car ce sont les mêmes que les paramètres 128 à 147, d'après l'annexe.

On remarque trois couples de variables très corrélées (corrélation supérieure à 0.99) : 36-37, 71-72 et 160-164. On supprime donc une variable de chaque couple (arbitrairement 37, 72, 160). De plus, les variables PAR_ASE_M, PAR_ASE_MV, PAR_SFM_M et PAR_SFM_MV représentent des moyennes d'autres variables, on peut donc également les éliminer.

On définit ensuite le modèle logistique. On suppose que le modèle est binomial car en effet, on souhaite expliquer une variable binaire. Cependant, on remet en cause l'hypothèse d'independance au vue de la segmentation des données (elles sont liées entre elles intuitivement).

1.2 Question 2

On définit l'échantillon d'apprentissage de la façon suivante :

```
set.seed(103)
train=sample(c(TRUE,FALSE),n,rep=TRUE,prob=c(2/3,1/3))
```

1.3 Question 3

On estime les modèles :

- Mod0 formé des variables PAR_TC, PAR_SC, PAR_SC_V, PAR_ASE_M, PAR_ASE_MV, PAR_SFM_M, PAR_SFM_MV.
- ModT contenant toutes les variables que vous aurez retenues à la question 1.
- Mod1 formé par toutes les variables signicatives au niveau 5% dans ModT.
- Mod2 formé par toutes les variables signicatives au niveau 20% dans ModT.
- ModAIC obtenu par sélection de variables stepwise (stepAIC) sur critère AIC à partir d'un modèle initial que vous définirez. Indiquer dans le fichier R la définition précise du modèle Y $\sim$ x....

1.4 Question 4

On trace les courbes ROC (fonctions prediction et performance du package ROCR) calculées sur l'échantillon d'apprentissage et sur l'échantillon de test pour le modèle ModT, la courbe de la règle parfaite et la courbe de la règle aléatoire.

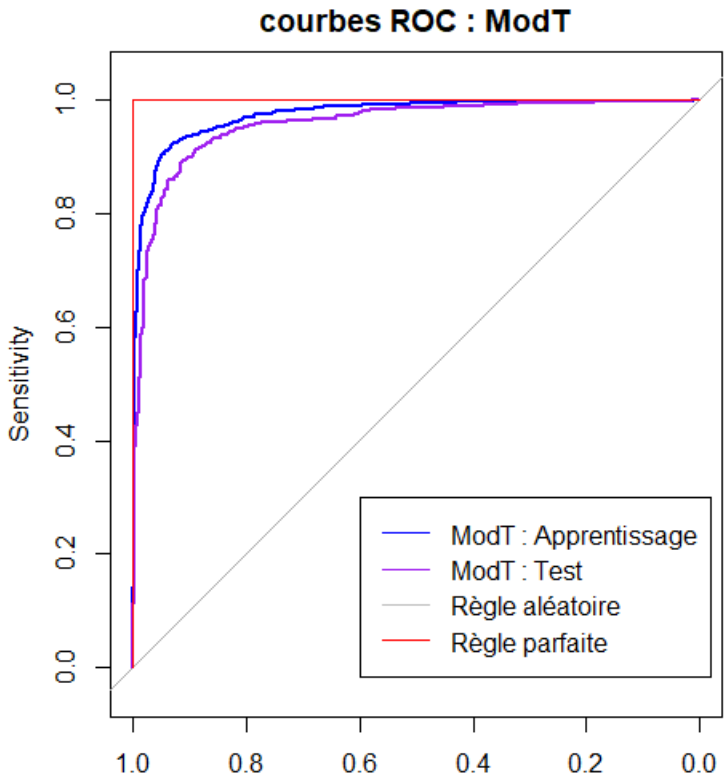

Figure 1 – Courbes ROC

Puis on superpose les courbes ROC de tous les autres modèles calculées sur l'échantillon de test. On peut alors calculer l'aire sous la courbe ROC pour chacun des modèles (performance).

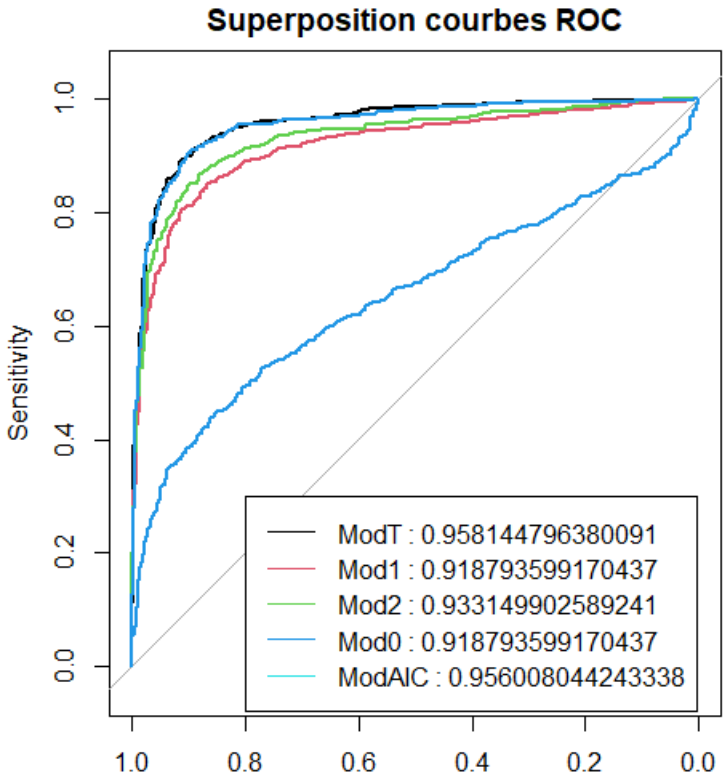

FIGURE 2 – Courbes ROCs (aire sous la courbe en légende)

1.5 Question 5

Pour chaque modèle défini, on calcule l'erreur sur l'échantillon d'apprentissage et sur l'échantillon de test.

Modèles	Type	Erreur
Modèle T	Apprentissage	0.06849315
	Test	0.09154437
Modèle 1	Apprentissage	0.913242
	Test	0.1104123
Modèle 2	Apprentissage	0.08851423
	Test	0.1062194
Modèle AIC	Apprentissage	0.07727432
	Test	0.09923131

On choisit le modèle possédant la plus petite erreur de test.
Il s'agit donc du modèle T.

2 Partie II : Régression ridge

On utilise maintenant la régression ridge.

2.1 Question 1

Comme vu précedemment, les variables sont très corrélées entre elles. De plus, on peut voir la régression ridge comme un problème d'optimisation sous contraintes ℓ_2.
On va donc minimiser la distance entre les solutions possibles.

2.2 Question 2

On utilise la fonction **glmnet** du package **glmnet**, pour un paramètre de régularisation λ variant de 10^{10} à 10^{-2}.

Les deux cas extrêmes s'apparentent à, respectivement, une très forte et une très faible pénalisation.

On obtient le graphique suivant en utilisant la fonction **plot** :

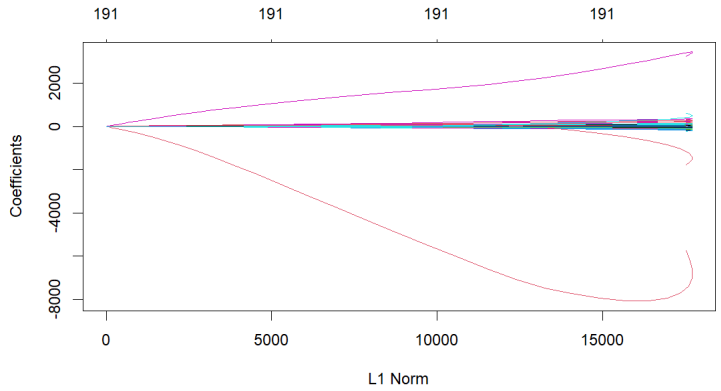

FIGURE 3 – Plot de glmnet

Le plot permet de visualiser l'effet de la régularisation sur les coefficients de régression estimés pour chaque variable d'entrée. On trouve en abscisse la valeur de la régularisation (λ en norme ℓ_1) sur une échelle logarithmique et en ordonnée les valeurs des coefficients de régression.

La figure représente un tracé des chemins de régularisation pour chaque variable : chaque variable est représentée par une ligne qui suit l'évolution de son coefficient en fonction de λ. On remarque que les coefficients varient peu (et sont quasi-nuls) pour la majeure partie des variables.

2.3 Question 3

On définit le germe du générateur à 314, puis on peut estimer le paramètre de régularisation par une validation croisée en 10 segments sur l'échantillon d'apprentissage en utilisant la fonction **cv.glmnet**.

L'algorithme **cv.glmnet** permet d'effectuer une validation croisée de la régression linéaire avec pénalisation ℓ_2 (ridge), en utilisant la méthode de la régression élastique. On détermine le meilleur paramètre de régularisation λ à partir de la plage de valeurs données, en utilisant une approche de validation croisée : la "k-fold cross-validation". Elle divise l'ensemble de données en k sous-ensembles de tailles égales.On construit alors un modèle sur k-1 sous-ensembles et on évalue sur le sous-ensemble restant. Cette procédure est répétée k fois en utilisant des sous-ensembles différents pour l'évaluation, et la performance du modèle est évaluée en moyenne sur les k itérations.

On trouve $\lambda = 10^{-2}$. Cela correspond au cas extrême où il y a une très faible pénalisation.

La valeur moyenne de l'erreur de validation croisée est de 0.249 pour ce λ. La performance de la méthode est de 0.216.

2.4 Question 4

Lorsqu'on utilise la totalité des variables (germe du générateur initialisé à 4658), l'erreur est de 0.0979418, en utilisant la régression ridge.

3 Partie III

On utilise désormais la méthode des plus proches voisins.

3.1 Question

On utilise la méthode des K- plus proches voisins avec K= 1, puis on boucle afin de déterminer le meilleur paramètre K. On trouve K=1 (pour un test jusqu'à K<500).

On peut critiquer l'utilisation de cette méthode car on trouve une erreur de $0,36$ sur l'échantillon. De plus, ayant K=1, on choisit uniquement le plus proche voisin, ce qui crée un fort biais.

Conclusion

D'après nos résultats, le modèle le plus performant en généralisation est le modèle de régression ridge. Il possède une erreur de classification de 0.09652236.

On peut désormer estimer les genres des extraits du fichier Music_2023_test.txt, qui sont enregistrés dans le fichier LIMES-PERDRIX_test.txt

Annexe : description du jeu de données

A database of 60 music performers has been prepared for the competition. The material is divided into six categories : classical music, jazz, blues, pop, rock and heavy metal. For each of the performers 15-20 music pieces have been collected. All music pieces are partitioned into 20 segments and parameterized. The descriptors used in parametrization also those formulated within the MPEG-7 standard, are only listed here since they have already been thoroughly reviewed and explained in many studies.

The feature vector consists of 191 parameters, the first 127 parameters are based on the MPEG-7 standard, the remaining ones are cepstral coefficients descriptors and time-related dedicated parameters :

a) parameter 1 : Temporal Centroid,
b) parameter 2 : Spectral Centroid average value,
c) parameter 3 : Spectral Centroid variance,
d) parameters 4-37 : Audio Spectrum Envelope (ASE) average values in 34 frequency bands
e) parameter 38 : ASE average value (averaged for all frequency bands)
f) parameters 39-72 : ASE variance values in 34 frequency bands
g) parameter 73 : averaged ASE variance parameters
h) parameters 74,75 : Audio Spectrum Centroid - average and variance values
i) parameters 76,77 : Audio Spectrum Spread - average and variance values
j) parameters 78-101 : Spectral Flatness Measure (SFM) average values for 24 frequency bands
k) parameter 102 : SFM average value (averaged for all frequency bands)
l) parameters 103-126 : Spectral Flatness Measure (SFM) variance values for 24 frequency bands m) parameter 127 : averaged SFM variance parameters
n) parameters 128-147 : 20 first mel cepstral coefficients average values
o) parameters 148-167 : the same as 128-147
p) parameters 168-191 : dedicated parameters in time domain based of the analysis of the distribution of the envelope in relation to the rms value.
q) GENRE : Classical, Jazz