M1- Mathématiques Appliquées Paris-Saclay - ENSTA
Apprentissage statistique (STA203), 2019-2020
Christine Keribin, Rémi Coulaud (Groupe 1), Alann Cheral (Groupe 2),
Simon Grah (Groupe 3)

Mini projet
à envoyer pour le 8 mai 2019 minuit au plus tard

Consignes

Le mini-projet donne lieu à un compte-rendu *rédigé* à effectuer en *binôme*.

- Ne pas oublier de définir un titre, une introduction pour préciser la problématique étudiée et le plan du travail, et une conclusion.
- Commenter les résultats obtenus, inclure les graphiques pertinents dans le corps du texte.
- Les résultats doivent être justifiés. La notation prendra en compte la clarté et le soin de la rédaction.
- A renvoyer sur l'adresse de votre chargé de TD sous forme du `pdf` du compte-rendu (qui peut être manuscrit puis photographié ou scanné) et d'un fichier texte `.R` contenant les commandes. Les fichiers seront nommés avec les noms du binôme: `NOM1-NOM2.pdf` et `NOM1-NOM2.R`.
- Aucun retard ne sera admis.

Introduction

Le jeu de données `Music.txt` est extrait d'un challenge de reconnaissance de genre de musique[1], 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.

Partie I

Pour résoudre ce problème de classification, on commence par effectuer une régression logistique.

1. Effectuer quelques analyses descriptives (univariée et bivariée). Quelle est la proportion de chacun des deux genres de musique ?

   Expliquer pourquoi il peut être judicieux d'appliquer une transformation `log` aux variables `PAR_SC_V` et `PAR_ASC_V`, et de supprimer les variables numérotées de 148 à 167.

   Discuter le cas des variables très corrélées (par exemple avec une corrélation supérieure à .99), et celui des variables `PAR_ASE_M`, `PAR_ASE_MV`, `PAR_SFM_M` et `PAR_SFM_MV`.

   Définir le modèle logistique. Discuter les hypothèses sur ce jeu de données.

2. Définir 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))
   ```

3. Estimer 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`.

[1] http://tunedit.org/challenge/music-retrieval/genres

- ModT contenant toutes les variables que vous aurez retenues à la question 1.
- Mod1 formé par toutes les variables significatives au niveau 5% dans ModT.
- Mod2 formé par toutes les variables significatives 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 (il est possible que vous ayez le temps d'une petite sieste en attendant les résultats).

4. Tracer 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.

   Superposer les courbes ROC de tous les autres modèles calculées sur l'échantillon de test.

   Calculer l'aire sous la courbe ROC pour chacun des modèles (`performance`), et l'afficher dans la légende.

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

   Quel modèle choisissez-vous? Peut-on tester son adéquation?

Partie II

1. Renseignez-vous sur la méthode des K plus proches voisins (ou k-NN, *k-nearest neighbors* en anglais), puis expliquer son principe.
2. Prédire avec $K = 1$ voisin (fonction `knn` du package `class`), puis choisir le paramètre K.
3. Commenter les résultats.

Partie III

On utilise maintenant la régression ridge.

1. Quel peut être l'intérêt de la régression ridge dans cette étude?
2. Utiliser la fonction `glmnet` du package `glmnet`, pour un paramètre de régularisation λ variant de 10^{10} à 10^{-2}. A quoi s'apparentent ces deux cas extrêmes?

   Interpréter le graphique tracé par la fonction `plot` appliquée à l'objet sorti par `glmnet`.

3. Définir le germe du générateur à 314, puis estimer le paramètre de régularisation par une validation croisée en 10 segments sur l'échantillon d'apprentissage en utilisant la fonction `cv.glmnet`. Expliquer l'algorithme, puis commenter son résultat.

   Calculer la performance de cette méthode.

4. Quelle est la performance de la régression ridge lorsqu'on utilise la totalité des variables (germe du générateur initialisé à 4658)?

Conclusion

Quelle méthode préconiser dans cette étude? Pouvez-vous déterminer sa performance de généralisation?

Bonus Vous pouvez aussi faire marcher votre imagination et proposer d'autres méthodes que celles de l'énoncé.

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
```