Analyse de données
Projet (à rendre pour le 18 décembre 2020)

Charlotte Baey (charlotte.baey@univ-lille.fr)

On s'intéresse dans ce projet à la classification de différentes espèces animales. Il y a deux parties, portant sur deux jeux de données distincts. Dans la première partie, on cherche à construire un modèle permettant de prédire la classe d'un animal en fonction d'un certain nombre de caractéristiques, et dans la deuxième partie on cherche à proposer une classification d'un ensemble d'animaux en plusieurs catégories.

Consignes

Le travail est à effectuer **en binôme**. Un rapport **au format PDF** doit être déposé sur la page Moodle du cours, accompagné du script correspondant. Merci d'indiquer vos noms dans le nom des fichiers, au format : `NOM1_NOM2_rapport.pdf` et `NOM1_NOM2_code.R`. Le rapport devra contenir la présentation du problème, ainsi que les réponses aux questions et les justifications nécessaires, sorties graphiques si besoin, ... il n'est pas nécessaire d'y inclure le code. Le fichier de code ne doit retourner aucune erreur.

La qualité de la rédaction sera évaluée tout autant que le contenu. Ainsi, chaque figure, chaque tableau, doit être numéroté et légendé, et doit être mentionné dans le texte.

1 Prédiction de la classe d'un animal

On s'intéresse dans cette partie à la prédiction de la classe d'un animal en fonction de plusieurs de ses caractéristiques. Les données se trouvent dans la base `animaux_I.csv`.

Sur le jeu de données considéré, la classe d'un animal peut être : mammifère, oiseau, reptile, poisson, amphibien, insecte ou invertébré (en anglais dans la base : Mammal, Bird, Reptile, Fish, Amphibian, Bug, Invertebrate). Les autres variables sont toutes booléennes (1 : vrai, 0 : faux) sauf mention contraire :

- `animal_name` le nom de l'animal,
- `hair` la présence de poils,
- `feathers` la présence de plumes,
- `eggs` indique si l'animal pond des œufs
- `milk` indique si l'animal produit du lait
- `airborne` la possibilité de voler
- `aquatic` indique si c'est un animal aquatique
- `predator` indique si l'animal est un prédateur
- `toothed` la présence de dents
- `backbone` la présence d'une colonne vertébrale

— `breathes` la possibilité de respirer
— `venomous` indique si l'animal est venimeux
— `fins` la présence de palmes
— `legs` le nombre de pattes
— `tail` la présence d'une queue
— `domestic` indique s'il s'agit d'un animal domestique
— `catsize` indique si l'animal a la taille d'un chat

Proposer un modèle permettant de prédire la classe d'un animal à l'aide des différentes variables de la table.
Expliquer la démarche et les différentes étapes suivies. Quelles sont les caractéristiques qui influencent le plus l'affectation d'un animal dans chacune des 7 classes de la base ?

2 Classification de photos d'animaux

Ici, on s'intéresse à la classification d'animaux en plusieurs sous-catégories à l'aide de photographies. Les données sont stockées sous forme de fichiers `.jpg` (un fichier par animal).

Vous devrez réaliser les étapes suivantes :

1. lire chaque fichier `.jpg` et l'importer sous R en le codant comme un vecteur. Pour pouvoir traiter tous les fichiers de façon homogène, on ne gardera que les 3 premières dimensions de l'image, correspondant aux couches de couleurs RGB (Rouge, Vert, Bleu). Certaines images contiennent une quatrième couche, qui correspond à la transparence de l'image, et à laquelle on ne s'intéresse pas dans ce projet. On doit obtenir une base de données contenant autant de lignes que de photos, et autant de colonnes que de pixels dans les trois canaux R, G et B (donc 3 fois le nombre de pixels de la photo).
2. Proposer une classification en sous-catégories, en justifiant le nombre de sous-catégories retenu. On pourra si besoin avoir recours aux étapes suivantes :
   — proposer une méthode de réduction de dimension (ACP) **en choisissant de manière pertinente** le nombre d'axes, *en rapport avec l'objectif final du projet*
   — comparer différentes approches et évaluer leurs performances de façon appropriée
   — proposer une interprétation des classes obtenues