**Note: Deep learning pour l’herbier**

**J Tressou**

**Février 2019**

Les données doivent être préalablement normalisées.

* Dans l’exemple des données MNIST (les chiffres) ou Fashion MNIST (images de vêtement, cf [tuto](https://keras.rstudio.com/articles/tutorial_basic_classification.html)), comme les X sont des images de basse résolution 28x28, on normalise en divisant par 255 car chaque pixel est un entier entre 0 et 255
* Dans l’exemple de régression « Boston housing Prices » ([tuto](https://keras.rstudio.com/articles/tutorial_basic_regression.html)), on a 13 inputs qui ne sont pas de même nature, on normalise en centrant et en réduisant (attention on calcule la moyenne et l’écart-type de chaque input sur les données d’apprentissage et on applique ces moyennes et sd aux données test/validation).

QUESTIONS :

* Pourquoi ne pas normaliser toutes les données d’un coup (par la moyenne et l’écart-type de l’ensemble) ?
* Quel est l’impact de différentes normalisations (moyenne et écart-type vs min-max par exemple) ?

Concernant les données herbier, cela demande donc plusieurs étapes préalables :

1. **Imputation des valeurs manquantes** -> utilisation du package Hmisc, imputation d’une catégorie spéciale valeurs manquantes pour les données qualitatives, et d’une valeur moyenne/médiane pour les variables quantitatives.

QUESTIONS :

* Impact de cette pratique sur le résultat ?
* serait-il intéressant d’ajouter une variable indiquant s’il y a eu imputation ou non sur une variable ou un groupe de variables (car toutes les variables « herbier » manquent pour les plantes qui n’y sont pas, idem quand plusieurs variables sont issues de la jointure avec une autre base …) ?

1. **Cas particulier des données qualitatives** : il faut convertir les variables qualitatives en « dummies » aussi appelé « one-hot encoding » (cf via la fonction to\_dummy du package « sjmisc ») mais cette pratique créée des difficultés car certaines modalités plus rares n’apparaissent pas dans les données d’apprentissage (mais sont bien là dans la base ensuite) et les intégrer initialement ajoute des variables constantes (égales à 0). – une option serait de n’intégrer que des variables avec des modalités pas trop rares (en regroupant des modalités).

QUESTIONS :

* Ne serait-il pas plus malin de partir des données brutes plutôt que des données agrégées dans ce contexte ?