MNIST ou le *Hello Word* de la classification d'images

Mohammed Sedki

Projet à rendre le 28/02/2018

MSP/ES *Apprentissage et agrégation de modèles*

1.Le jeu de données MNIST

Le jeu de données MNIST est hébergé sur le page web de Yann LeCun. Ce jeu de données sert de référence pour les compétitions en apprentissage où la donnée explicative est sous forme d'image. Commençons par une visualisation du jeu de données et plus précisément les 36 premières images. Pour cela nous allons installer le package keras qui installe à son tour tensorflow. Le bloc de code suivant permet cette visualisation

```
#install.packages("keras", dep=TRUE)
require(keras)
mnist   <- dataset_mnist()
x_train <- mnist$train$x
y_train <- mnist$train$y
x_test  <- mnist$test$x
y_test  <- mnist$test$y

# visualize the digits
par(mfcol=c(6,6))
par(mar=c(0, 0, 3, 0), xaxs='i', yaxs='i')
for (idx in 1:36) {
  im <- x_train[idx,,]
  im <- t(apply(im, 2, rev))
  image(1:28, 1:28, im, col=gray((0:255)/255),
        xaxt='n', main=paste(y_train[idx]))
}
```

2.Préparation des données

Afin de pouvoir mettre en place les méthodes d'apprentissage classiques, nous avons besoin d'aplatir les images en vecteurs. Cela revient à transformer une image de dimension 28×28 pixels en un vecteur de dimension 784.

1. Préparer le jeu de données en appliquant un aplatissement[1] ainsi qu'une suppression des

[1] La fonction *array_reshape* permet aplatir des matrices.

pixels nuls pour l'ensemble des images du jeu de données. La fonction nearZeroVar du package caret permet de repérer les variables de variance nulle.

2. Le nombre de variables du jeu de données après aplatissement est très élevé. Proposer une procédure de réduction de dimension indépendante de la méthode d'apprentissage à utiliser. La fonction preProcess du package caret permet un ensemble de transformations de données. Afficher la dimension avant et après suppression et réduction de dimension et faire en sorte à ce que la procédure garde les mêmes dimensions pour les deux jeux de données train et test.

3. Apprentissage par SVM linéaire

3. À l'aide du package caret, entraîner un modèle SVM linéaire avec les paramètres de contrôle suivants :
   a. Pour le paramètre C, on propose de tester la séquence 10**(-3:0).
   b. Utiliser une validation croisée à 5 folds répétée deux fois[2].
   c. Donner le taux de bon classement du jeu de données test.

4. Apprentissage par réseaux de neurones artificiels

Les RNA sont au cœur de l'apprentissage profond (*Deep Learning*). Ils sont polyvalents, puissants et extensibles, ce qui les rend parfaitement adaptés aux tâches d'apprentissage automatique extrêmement complexes, comme la classification de millions d'images (par exemple, Google Images), la reconnaissance vocale (par exemple, Apple Siri), la recommandation de vidéos auprès de centaines de millions d'utilisateurs (par exemple, YouTube) ou l'apprentissage nécessaire pour battre le champion du monde du jeu de go en analysant des millions de parties antérieures puis en jouant contre soi-même (AlphaGo de DeepMind).

Nous proposons ici une introduction aux Réseaux de Neurones Artificiels, en commençant par une description rapide des toute première architecture de RNA. Nous présenterons ensuite les *perceptrons multicouches* (PMC).

4.1. Le perceptron

Le perceptron, inventé en 1959 par F. Rosenblatt, est l'architecture de RNA la plus simple. Il se fonde sur un neurone artificiel appelé unité linéaire à seuil (LTU, *Linear Threshold Unit*). Les entrées correspondent aux variables explicatives et la sortie correspond à la prédiction de la variable à expliquer. Par analogie aux neurones biologiques, la LTU est un modèle qui se caractérise par une fonction dite *d'activation* qu'on notera g et les poids des entrées. La sortie d'un neurone artificiel est donnée par

$$h(x_1, \ldots, x_p) = g\left(w_0 + \sum_{j=1}^{p} w_j x_j\right) = g(w_0 + \boldsymbol{w}'\boldsymbol{x}).$$

[2] Attention le calcul peut prendre entre 15 et 30 minutes

La fonction d'activation applique une transformation d'une combinaison linéaire des entrées pondérées par des poids $w_0, w_1, \ldots, w_p$ où le poids w_0 est appelé le biais du neurone. Le choix de la fonction d'activation g se fait selon la nature de la variable à expliquer y (binaire ou continue). Souvent on fait appel à une fonction d'activation correspondant à l'identité dans le cas d'un problème de régression ou la fonction d'activation softmax quand il s'agit d'un problème de classification[3]. Entraîner une LTU revient à choisir les valeurs des poids $w_0, w_1, \ldots, w_p$ appropriées par rétro propagation élémentaire du gradient[4].

4.2. Le perceptron multicouches

Les performances d'apprentissage d'un perceptron composé d'une seule LTU sont souvent médiocres. Il est souvent possible d'améliorer les performances des perceptrons en empilant plusieurs perceptrons. Le RNA résultant est appelé *perceptron multicouche* (PMC). Un PMC est constitué d'une couche d'entrée (qui ne fait rien, si ce n'est distribuer les entrées aux neurones de la couche suivante), d'une ou plusieurs couches successives de LTU appelées *couches cachées* et d'une dernière couche de LTU appelée *couche de sortie*. Une couche est un ensemble de LTU n'ayant pas de connexion entre elles. Selon les auteurs, la couche d'entrée qui n'introduit aucune modification n'est pas comptabilisée. Une ou plusieurs couches cachées participent au transfert. Dans un perceptron multicouche, une LTU d'une couche cachée est connectée en entrée à chacune des LTU de la couche précédente et en sortie à chaque LTU de la couche suivante. Lorsque un perceptron multicouche possède deux couches cachées ou plus, on parle de réseaux de neurones profond (RNP) au lieu de RNA (en anglais, on parle de *Deep Neural Network (DNN)*).

5. Implémentation d'un RNA avec une seule couche cachée avec keras

Reprenons le jeu de données après aplatissement mais sans réduction de dimension. Appliquons une normalisation en divisant les intensités des pixels par 255 (intensité maximale d'un pixel).

```
require(keras)
mnist <- dataset_mnist()
x_train <- mnist$train$x
y_train <- mnist$train$y
x_test <- mnist$test$x
y_test <- mnist$test$y
y_train <- to_categorical(y_train, 10)
y_test <- to_categorical(y_test, 10)
# reshape
x_train <- array_reshape(x_train, c(nrow(x_train), 784))
x_test <- array_reshape(x_test, c(nrow(x_test), 784))
# rescale
x_train <- x_train / 255
x_test <- x_test / 255
```

4. À l'aide du package keras, entraîner un RNA avec une couche cachée comme suit :

[3] Il est fortement recommandé de consulter le cours de Philippe Besse disponible ici pour plus de détails.
[4] Le cours de Philippe Besse détaille l'algorithme de rétro propagation élémentaire du gradient.

a. Le nombre de neurones (LTU) de la couche cachée est de 784 avec la fonction d'activation relu[5]. De combien de LTU a-t-on besoin à la dernière couche ? préciser sa fonction d'activation.

b. Utiliser la fonction de perte ainsi que la métrique appropriés pour le contexte de la classification. Faire appel à l'algorithme d'optimisation *adam*. Utiliser des mini-lots de données de taille 200 et 10 cycles d'apprentissage. Inclure dans la procédure le jeu de données test comme jeu de données de validation[6]. Préciser la meilleure performance obtenue en terme de taux de bon classement sur les 10 cycles d'apprentissage.

6. Apprentissage par réseaux de neurones convolutifs

Pour certains types de données, en particulier les images, l'utilisation des perceptrons multi-couches impose la transformation des données sous forme de matrice d'intensités de pixels en vecteur. Cette transformation sacrifie l'information spatiale contenue dans les images, telle que les formes. Avant le développement de l'apprentissage par réseaux de neurones profonds, les algorithmes d'apprentissage faisaient appel à l'extraction de variables d'intérêt, appelées caractéristiques[7]. Les réseaux de neurones convolutifs (CNN) introduits par *LeCun 1998* ont révolutionné la classification d'images en contournant l'extraction manuelle des caractéristiques. Les CNN agissent directement sur la matrice d'intensités des pixels.

7. Implémentation d'un réseau de neurones convolutifs

Reprendre le jeu de données précédent sans les transformations *array_reshape*[8].

5. À l'aide du package keras, entraîner un CNN composé des couches suivantes :
    1. Une couche de convolution 2d avec 30 filtres de convolution ainsi qu'un noyaux d'un pas de c(5,5) et une fonction d'activation *relu*.
    2. Une couche de max-pooling 2d avec un noyau de pooling d'un pas de c(2,2).
    3. Une couche de convolution 2d avec 15 filtres de convolution ainsi qu'un noyaux d'un pas de c(3,3) et une fonction d'activation relu.
    4. Une couche de max-pooling 2d avec un noyau de pooling d'un pas de c(2,2).
    5. Une couche *dropout* avec un taux d'extinction de 30 %.
    6. Une couche d'aplatissement flatten pour aplatir les sorties de la couche précédente.
    7. Une couche cachée composée de 128 LTU et une fonction d'activation relu.
    8. Une couche cachée composée de 50 LTU et une fonction d'activation relu.
    9. La dernière couche de sortie composée de 10 LTU et une fonction d'activation softmax.

[5] Il faut absolument consulter l'aide du package keras pour comprendre le principe de fonctionnement et de construction des réseaux de neurones avec keras.

[6] La lecture de ce document: *Optimization algorithms* permet de se familiariser avec le vocabulaire.

[7] Méthode de descripteurs SIFT par exemple.

[8] Conserver la normalisation des intensités.

Reprendre les mêmes paramètres d'entraînement précisés dans le point b. de la section 5. Expliquer intuitivement le rôle de chaque couche du réseau.