Analyse, classification et indexation des données

(Mini) projet

Présentation

Dans une usine de tri, on veut mettre en place des bras automatisés permettant de trier des vêtements, des chaussures et des accessoires. La première étape du tri consiste à séparer les accessoires des autres objets.

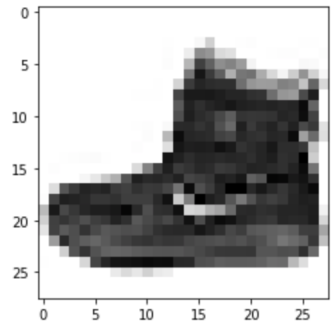
Afin de réaliser une première démonstration (POC ou "proof of concept"), nous allons travailler sur un jeu de données simple : le corpus `fashion_mnist`. Le corpus proposé fait partie des datasets disponibles sur `keras` et peut donc être chargé grâce aux instructions :

```
from keras.datasets import fashion_mnist
(X_train, y_train), (X_test, y_test) = fashion_mnist.load_data()
```

Le corpus contient 60000 images de 28×28 pixels en niveau de gris. Chaque image représente un objet. Il y a 10 catégories d'objets, chacune étant codée par un chiffre :

0 : T-shirt/top
1 : Trouser
2 : Pullover
3 : Dress
4 : Coat
5 : Sandal
6 : Shirt
7 : Sneaker
8 : Bag
9 : Ankle boot

Voici un exemple d'une image du corpus et le code utilisé pour l'obtenir .




```
9
```

```
import matplotlib.pyplot as plt
import matplotlib.cm as cm
%matplotlib inline
plt.imshow(X_train[0], cmap=cm.Greys)
plt.show()
print(y_train[0])
```

On souhaite entraîner un modèle qui permet de trier les images en deux catégories : `vetements` et `accessoires`. Dans la catégorie `vetements`, on retrouvera les types 0, 1, 2, 3, 4, 5, 6, 7, 9 et dans la catégorie `accessoires` le type 8 seulement.

Travail demandé

Le but de ce travail est de tester des algorithmes de classification vus en cours/tds sur le corpus présenté ci-dessus. La rédaction et la méthode suivie seront autant appréciées que les résultats obtenus. Concrètement, l'équipe pédagogique attend de vous de :

1. Modifier le corpus en remplaçant les labels par les deux classes `vetement` et `accessoire`.
2. Entrainer et tester deux classifieurs vus en cours sur cette nouvelle base. Vous devrez justifier votre choix concernant les deux classifieurs retenus ainsi que leur paramètres si nécessaire. Vous devez également expliquer les critères permettant d'évaluer la qualité de la classification.
3. Étudier l'impact de la réduction de dimension par LDA sur la faisabilité, la qualité et la complexité de la classification. Vous proposerez en particulier une visualisation de l'ensemble d'entrainement réduit par LDA.
4. Écrire une conclusion reprenant les différents éléments et justifiant les résultats obtenus.

Modalités pratiques

1. Vous êtes libre d'utiliser le module sklearn.
2. Le projet est à réaliser en binômes
3. Il doit être déposé par un membre du groupe sur moodle
   https ://moodle1.u-bordeaux.fr/course/view.php ?id=9757
   Le nom du fichier doit indiquer les noms du binôme et le groupe. Le format attendu des noms de fichiers est le suivant `GrX-Nom1_Prenom1-Nom2_Prenom2.format`. Attention : des pénalités seront appliquées en cas de non respect de cette convention.
4. La date limite pour déposer vos fichiers est fixée au 02 janvier 2023 à 23h59.