

Data Science et Machine Learning

TP sur l'Analyse Discriminante (AD)

A réaliser dans l'environnement Python avec les bibliothèques Scikit-Learn, Pandas et Matplotlib

Ecrivez un programme pour la reconnaissance des chiffres écrits à la main. Nous utilisons 2 mille chiffres extraits de la base de Facebook. Ils sont en figures de taille $28 \times 28 = 784$ pixels. Chaque pixel correspond à une variable. La matrice X fait 2000 lignes et 784 colonnes. Une colonne supplémentaire précise la classe de chaque individu.

1. Utilisez la bibliothèque Pandas pour extraire les données à partir du fichier chiffres.xlsx.

```
df = pd.read_excel('chiffres.xlsx')
XY = df.values
```
2. Réalisez une AFD descriptive en construisant la projection des individus et des moyennes des 10 classes sur le premier plan discriminant. Il n'est pas nécessaire de centrer et réduire les données. Utilisez seulement les premiers 1800 individus pour l'apprentissage et laissez 200 individus pour la validation.

```
lda = LinearDiscriminantAnalysis()
lda.fit(X_train, Y_train)
print(lda.classes_)
```
3. Réalisez un classement avec la méthode Predict et évaluer le taux de réussite sur l'ensemble d'apprentissage et l'ensemble de validation.

```
classement = lda.predict(X_train)
```
4. Construisez le dessin du nuage projeté avec Matplotlib.

```
Coul_code = ['#F0C0FA', 'c', '#FAF785', 'y', '#F1B909', '#FA859C', '#85B7FA', '#D0E2AF', '#0BF4F7', '#85FA85']
Coul_classe = dict()
m=10
for k in range(m):
    Coul_classe[lda.classes_[k]] = Coul_code[k]
coul_ind = []
for i in range(t):
    coul_ind.append(Coul_classe[...])
fig, ax = plt.subplots()
ax.scatter(F[:,0], F[:,1], color = ...)
```
5. Appliquez votre programme pour la reconnaissance du chiffre 6 dont la figure est fournie avec ce TP.

```
A = img.imread('six.png')
print(A.shape)
n = A.shape[0]
m = A.shape[1]
V = np.zeros(n*m)
for i in range(n):
    for j in range(m):
        V[m*i+j] = int(A[i][j][0]*255)
V = V.reshape(...,...)
```