**PROJET D'ANALYSE DE DONNEES 2**

Thaïs Perdriel & Fernanda Tchouacheu

Avril 2019

Université Paris Descartes



# Sommaire

Sommaire 2

Exercice 1 3

Question 1 3

Question 2 et 3 6

Question 4 8

Question 5 8

Question 6 8

Question 7 9

# Exercice 1

## Question 1

Décrivez les 4 datasets. Représentez graphiquement le visage moyen (moyenne de toutes les lignes) de l’échantillon train.

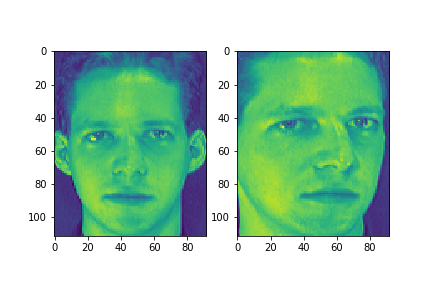
**Réponse :**

Notre jeu de données est séparé en 4 datasets :

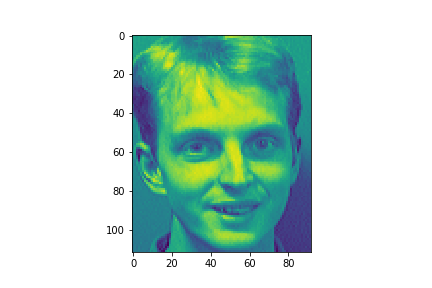
* un jeu de donnée d'apprentissage qui correspond au dataset "train" et ses labels (labels\_train)
* un jeu de donnée de test qui permettra de verifier si les données ont bien été labelisées, dataset « test » et ses labels (labels\_test)

La base de donnée d’entraînement contient 320 photos représentant 40 personnes différentes. Chaque personne est représentée par 8 photos. La base de donnée de test contient 80 photos représentant ces 40 même personnes. Chaque personne est représentée par 2 photos. Les bases de données train et test comportent 10304 colonne. Chaque colonne correspond à un pixel, une photo étant formée de 92\*112 pixels. La valeur prise par chaque pixel est comprise entre 0 et 1.

Voici deux exemples de visages :

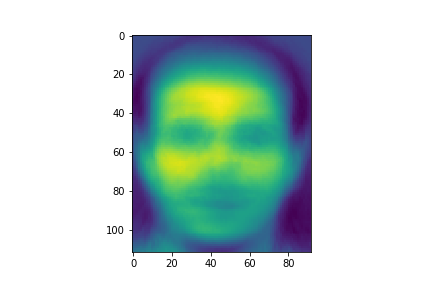


**Figure 1.** Premier visage



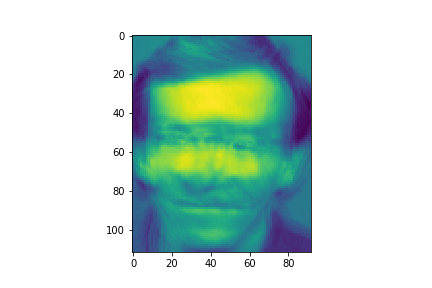
**Figure 2.** Dernier visage

Représentation du visage moyen (moyenne de toute les lignes) du dataset train :



**Figure 3.** Visage moyen du dataset train

Nous pouvons aussi représenter le visage moyen d’un seul individu :

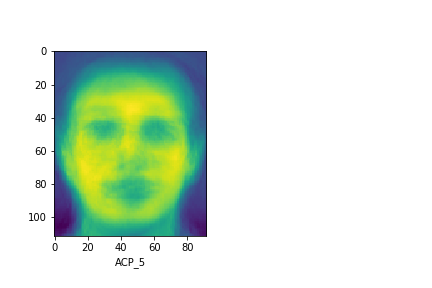


**Figure 4.** Visage moyen de l’individu n°3 du dataset train.

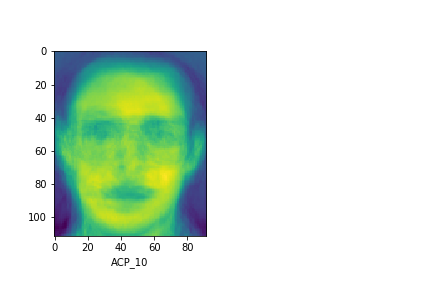
## Question 2 et 3

Appliquez une ACP pour reduire la dimension du dataset pour k=5, 10 et 50 puis choisissez un visage et représentez le graphiquement pour k=5, 10 et 50.

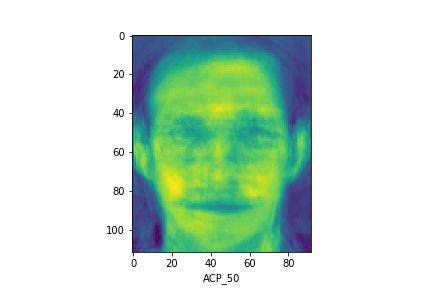
**Réponse :**



**Figure 5.** Représentation graphique du visage pour k=5



**Figure 6.** Représentation graphique du visage pour k=10



**Figure 7.** Représentation graphique du visage pour k=50

## Question 4

L’ACP est plutôt longue lorsque la quantité de colonnes est très grande, proposez une solution.

**Réponse :**

* Choisir les variables actives (en faisant un tableau de corrélations sur les variables)
* Choisir de réduire ou non les variables
* Réaliser l’ACP

## Question 5

En utilisant le dataset réduit par ACP, implémentez un K plus proche voisin, pour pouvoir classifier les images du test.

**Réponse :** Notre algorithme de classification basée sur la technique du K plus proche voisin obtient un score d’adéquation de 90.625 %.

## Question 6

Faîtes une analyse d’erreur sur vos labels prédits et les vrais labels du test.

**Réponse :**

**Tableau 1.** Analyse d’erreur sur les labels prédits et les vrais labels du test pour n\_neighbours = 5

|  |  |  |
| --- | --- | --- |
| **k** | **Score sur le dataset train** | **Score sur le data set test** |
| 5 | 90.625 | 76.25 |
| 10 | 93.125 | 85.0 |
| 30 | 96.875 | 91.25 |
| 50 | 97.5 | 93.75 |
| 150 | 96.875 | 92.5 |
| 300 | 96.25 | 93.75 |

**Tableau 2.** Analyse d’erreur sur les labels prédits et les vrais labels du test pour n\_neighbours = 3

|  |  |  |
| --- | --- | --- |
| **k** | **Score sur le dataset train** | **Score sur le data set test** |
| 5 | 95.0 | 76.25 |
| 10 | 97.5 | 87.5 |
| 30 | 99.0625 | 92.5 |
| 50 | 99.0625 | 92.5 |
| 150 | 99.0625 | 93.75 |
| 300 | 99.0625 | 93.75 |

**Analyse :** Pour n\_neighbours = 5, le meilleur résultat est obtenu pour k=50, pour n\_neigbours = 3, le meilleur résultat est obtenu pour k=150.

## Question 7

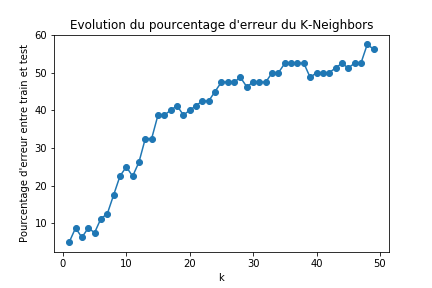
Offrez des recommandations pour améliorer le modèle et appliquez les.

**Réponse :**

**1ère méthode :**

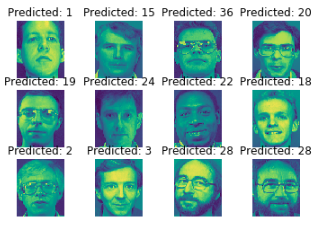
Faire une optimisation du score sur les données test ou une validation croisée afin de déterminer le k optimal. Concrètement nous allons simplement tester le modèle pour tous les k de 1 à 50, mesurer l’erreur test et afficher la performance en fonction de k

Implémentation :



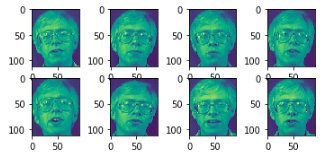
**Figure 8.** Evolution du pourcentage d’erreur du K-Neighbors en fonction du k choisi

Comme on peut le voir, le k-NN le plus performant est celui pour lequel k = 3. On connait donc notre classifieur final optimal : 3-nn. Ce qui veut dire que c'est celui qui classifie le mieux les données.



**Figure 9.** Exemples de résultats de classificiation de l’algorithme K-neighors pour k=3

Voici toutes les représentations du visage 2 dans le data set train :



**Figure 10.** Les 8 représentations du visage n°2

**2ème méthode :**  La pénalisation

Les régressions pénalisées sont des techniques puissantes généralement utilisées pour créer des modèles parcimonieux en présence d'un "grand" nombre de variables.

**Tableau 2.** Erreur absolue à la moyenne obtenue par régression linéaire simple, pénalisation Ridge puis Lasso.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Régression linéaire** | **Pénalisation Ridge** | **Pénalisation Lasso** |
| Erreur absolue à la moyenne | 6.113 | 6.107 | 7.184 |

On choisira la pénalisation Ridge car l'erreur est la plus petite.