Université de Lille **Data Sciences**
M2 MFCA obouaziz@univ-lille.fr

TP. 2

Modèle de Mélange Gaussien

Exercice 1

Dans cet exercice on va travailler un jeu de données (diabetes.csv) qui comporte 8 variables explicatives et une variable d'intérêt : `Outcome` qui prend deux valeurs 1 si le patient souffre de diabète et 0 sinon. C'est cette variable que l'on va chercher à prédire.

1. Charger la base de données, regarder quelques statistiques descriptives sur les différentes variables. On pourra travailler avec des dataframe et la librairie `pandas`.
2. Creée une base Test et une base apprentissage (il y a plusieurs façon de faire, manuellement et grâce à la fonction `train_test_split` de `sklearn`). Pourquoi a-t-on besoin de scinder ainsi en deux la base de données ?
3. Afin de faciliter la visualisation graphique on commence par se concentrer sur 2 variables, `Glucose` et `Age`. Tracer le nuage de points associé en colorant chaque point en fonction de sa classe d'appartenance.
4. Regarder l'aide des fonctions :
   `LinearDiscriminantAnalysis` et `QuadraticDiscriminantAnalysis`,
   elles permettant de faire respectivement de l'analyse gaussienne homoscédastique et hétéroscedastique. Les appliquer aux données.
5. Tracer sur un même graphique le nuage de point et la frontière de décision dans ces deux cas.
6. Calculer l'erreur de classification des ces deux modèles (on pourra utiliser la fonction `confusion_matrix`).
7. Reprendre la question précédente en utilisant cette fois l'ensemble des variables explicatives.

Exercice 2

Dans cet exercice on va reprendre un jeu de données d'images de chiffres et se concentrer sur les classes $\{1, 7, 8\}$. Chaque chiffre est représenté par une matrice de taille 28×28, chaque élément de cette matrice correspondant à un niveau de gris. L'objectif de cet exercice est de décrire, puis de prédire, l'appartenance d'une image à chacune des trois classes. les données contiennent :

— une matrice x contenant 3000 lignes et 784 colonnes, correspondant à l'échantillon d'apprentissage,

— une matrice xt contenant 1500 lignes et 784 colonnes, correspondant à l'échantillon test,

— un vecteur y contenant les étiquettes de l'échantillon d'apprentissage et

— un vecteur yt contenant les étiquettes de l'échantillon test.

1. Importer les données et tracer quelques images. Faire une ACP et tracer le nuage de points dans le premier plan discriminant. Commenter.
2. Réaliser une Analyse discriminante sur la base d'apprentissage.
3. Proposer une règle de décision pour affecter un nouveau point à l'un des trois groupes, et appliquer cette règle à l'échantillon test. Quel est le taux de mal classés ? Que remarque t-on ? Pourquoi ?