Compte rendu Machine Learning

OUTAYEB Katia – ABED Amir Chawki

1- Description du jeu de données :

Le jeu de données est composé de caractéristiques extraites de 7 vidéos avec 3 personnes gesticulant, visant à étudier la segmentation par phase de geste. Expérience réalisée à l'université de Sao Paulo avec un capteur de mouvement 'Kinect', ce capteur rend en sortie les images des vidéos avec un fichier des positions 3D de : la main gauche, main droite, poignet gauche, poignet droit, de la colonne vertébrale et la tête et il représente un fichier brut.

On dispose également d'un deuxième fichier traité (à partir du premier fichier), qui contient la vitesse et l'accélération des mains et des poignets et qui a en tout 32 features et un target 'Phase' qui prend les valeurs ('D', 'P', 'S', 'H', 'R').

a)- Les données du projet :

Dans notre projet on s'intéresse à la classification selon la classe 'S'et la classe 'D'et on utilisera le deuxième fichier. Nous avons donc au total 5691 données sur lesquelles nous allons travailler (Classe 'D' : 2741, Classe 'S' : 2950).

b)- Propriétés des données :

- Le jeu de données est normalisé : toutes les données sont dans l'intervalle [0,1]
- 0 valeurs manquantes.
- La corrélation des données est comme suit :

Attributs	Corrélation
X1 - X7 (Main gauche X - Poignet gauche X)	0.858064
X2 - X8 Main gauche X - Poignet gauche X)	0.935514
X3 - X9 (Main gauche Z - Poignet gauche Z)	0.852701
X4 - X10 (Main droite X - Poignet droite X)	0.885638
X5 - X11 (Main droite Y - Poignet droite Y)	0.947806
X6 - X12 (Main droite Z - Poignet droite Z)	0.870646
X13 - X19 (Main gauche X - Poignet gauche X)	0.688900
X14 - X20 (Main gauche Y - Poignet gauche Y)	0.805323
X15 - X21 (Main gauche Z - Poignet gauche Z)	0.691074
X16 - X22 (Main droite X - Poignet droite X)	0.758995
X17 - X23 (Main droite Y - Poignet droite Y)	0.831011
X18 - X24 (Main droite Z - Poignet droite Z)	0.662363
X25 - X27 (Main gauche - Poignet gauche)	0.921779
X26 - X28 (Main droite - Poignet droite)	0.934080
X30 - X32 (Main droite - Poignet droite)	0.841442
X31 - X29 (Poignet gauche - Main gauche)	0.817196

- Les colonnes de X1 à X12 représentent la vitesse vectorielle, de X13 à X24 nous avons l'accélération vectorielle , et de X25 à X32 c'est la vitesse scalaire.
- Lorsque la corrélation entre deux colonnes est supérieure à 0.90 nous avons choisis d'éliminer une des deux colonnes. Ce qui revient à éliminer les colonnes {X8, X11, X27, X28}

2- Nom des méthodes et protocole de comparaison des méthodes

Nous allons diviser le DataSet de sorte à deviner la valeur de la dernière colonne qui est 'Phase'.

- **Feautres** : les colonnes X1, X2, ….,X32
- **Target** : 'Phase'

Méthodes à utiliser :

KNN : KNN est simple à utiliser et peut donner des résultats très précis et significatifs. Il a également tendance à être coûteux en calcul, de sorte qu'il peut ne pas être un meilleur choix pour des jeux de données plus volumineux mais dans notre cas nous avons un petit DataSet avec classification binaire donc KNN peut être un bon choix.

SVM : Nous avons choisi d'utiliser SVM, car il répond parfaitement à notre problème qui est un problème de classification, en effet SVM a pour but de trouver une frontière parfaite afin de séparer les données, ce qui cadre parfaitement avec notre but qui est de déterminer la classe d'une donnée suivant ses différentes colonnes.

K-Fold Cross Validation : Les méthodes seront entrainées sur l'ensemble du jeu de données en utilisant une validation croisée.

Pourquoi cette méthode : Car les données sont divisées en k sous-ensembles. La méthode est répétée k fois, de sorte que chaque fois, l'un des k sous-ensembles est utilisé en tant qu'ensemble de test et que les k-1 autres sous-ensembles sont rassemblés pour former un ensemble d'apprentissage. L'avantage avec cette méthode c'est que toutes les données seront utilisées comme ensemble d'apprentissage et comme ensemble de test, et dans notre cas ça nous arrange car nous avons un petit DataSet.

GridSearch : GridSearch est une approche de tuning des paramètres qui va construire et évaluer méthodiquement un modèle pour chaque combinaison de paramètres d'algorithme spécifiés. Ce qui va permettre de trouver les paramètres optimaux.

Comparaison des méthodes :

Pour comparer les 2 méthodes nous allons nous utiliser les métriques extraites des statistiques issues à la fin d'exécution de chaque à savoir :

- L'Exactitude des classifications (Accuracy): le pourcentage de prédiction que le modèle a correctement deviné.
- Confusion Matrix