

Compte rendu du Classification Challenge

Pour réaliser cette classification nous avons décidé d'utiliser la distance euclidienne comme fonction de calcul de distance.

Nous avons utilisé la fonction ApprentissageSupervisé afin de tester toutes nos fonctions qui nous permettent de faire la classification.

Afin de savoir quel k nous allons utiliser, nous avons créé une fonction qui va calculer la précision de notre algorithme pour chaque k allant de 1 à 29.

Pour ce faire, nous avons tout d'abord créé une fonction qui va calculer le pourcentage de précision de la classification en fonction de k .

Pour cela, nous prenons une donnée du fichier, on la stock dans une variable, puis la « retire » du fichier, puis nous voyons si avec notre classification nous retrouvons bien le bon label de départ (stocké dans la variable). Ensuite nous faisons cette opération pour chacune des variables.

Nous avons ensuite créé une fonction qui va tester le pourcentage de précision en fonction du k passé en paramètre. Pour cela, nous faisons un algorithme qui va tester pour les valeurs de k allant de 1 à 30 le pourcentage de précision. Cette fonction a été utilisée avec le fichier preTest et nous a retourné que le meilleur k possible était 4 avec une précision de 85,8%

Enfin pour créer notre fichier final, nous avons donc appliqué notre algorithme de classification avec comme fichier de référence le fichier data, avec un $k = 4$ au fichier finalTest. Cela nous retourne la liste des labels trouvés. Enfin nous créons un fichier qui stocke toutes ces données.

Nous avons aussi pris soin de vérifier que notre fichier avait le bon format avec la fonction que vous nous avez fournis.