RAPPORT : Classification Challenge

Il y’a trois points essentiels dans notre travail sur ce challenge :

* La distance : nous avons, dans un premier temps, choisi d’utiliser la distance dites de « similarité cosine » qui calcule le cosinus entre deux vecteurs, pour en donner la similarité. (Proche de 1, très semblable, proche de -1 radicalement opposé), ce qui nous semblait plus parlants en dimensions 4 qu’une simple distance euclidienne. Cependant, nous nous sommes aperçus qu’il y’avait une grande similarité entre les distances calculées pour un même « data ». Aussi, nous avons décider d’utiliser la distance euclidienne pour augmenter l’écart entre les valeurs (d’où la division par la distance euclidienne), ceci nous permettant de gagner plus de 5% de précision.
* Le choix du k = 5 : cela semblait être la valeur de k pour laquelle nous avons le plus de précisions (testées entre 1 et 10 puis pour 15 et 20)
* Nous avons pris le pari d’accoler les données de « data » et « preTest», n’ayant rien vu allant à l’encontre de cela dans le sujet. Bien que nous ne soyons pas sûr que cela apporte un gain réel en précision (nous avons essayé en accolant seulement 50% de preTest et de tester avec les autres 50% mais cela n’a rien changé a la précision par rapport à data seul).