

TP 3 - Classification de chiffres 8*8

Olivier Goudet

February 24, 2025

Exercice

Questions

1. Charger le dataset "digits_train.csv" disponible sur Moodle dans un dataframe. Les chiffres (8,8) sont encodés en ligne avec 64 valeurs et les labels sont donnés sur la dernière colonne.
2. Quelle est la dimension d pour ce problème ? Combien y a-t-il de classes différentes à prédire dans ce problème ?
3. Afficher le premier chiffre de la base après un redimensionnement. Pour l'affichage, on pourra utiliser la fonction `imshow` du module `matplotlib.pyplot` avec l'option `cmap=gray_r` pour afficher des nuances de gris.
4. Séparer ces données en 2 ensembles : apprentissage (60%) et validation (40%).
5. Implémenter l'algorithme des k plus proches voisins pour ce problème.
6. Régler la valeur du paramètre k sur l'ensemble de validation.
7. Une fois la meilleure valeur de k choisie, faites des prédictions sur l'ensemble de test pour chaque chiffre qui se trouve dans le fichier "digits_entries_test.csv".
8. Mesurez la performance de votre modèle en comparant vos prédictions avec les vraies classes qui se trouvent dans le fichier "digits_target_test.csv".
9. Pour chaque chiffre du dataset de test, afficher vos prédictions ainsi que l'image du chiffre correspondant (cf. question 2 plus haut). Qu'est-ce que vous observez ?
10. Répéter le même processus d'entraînement puis de test en appliquant l'algorithme de régression multivariée (voir script donné en cours).
11. Comparer les résultats obtenus par les deux méthodes sur l'ensemble de test.