UMAP - Découverte

Master 280 - Théo Lopès-Quintas

Introduction

Le dataset MNIST (*Mixed National institute of Standards and Technology*) est une base de données de chiffres écrits à la main créé en 1994 par les équipes de Yann Le Cun chez Bell Labs. Il consiste en 70 000 images de chiffres écrits à la main et représente un des tests les plus classique de l'apprentissage supervisé. Dans le papier original un SVM est utilisé pour classifier ces images.

Les images sont au format 28 par 28 pixels, donc une dimension de la base de 784. Nous souhaitons travailler dans une dimension inférieure sans pour autant perdre beaucoup en terme de performance.

Sujet

Nous souhaitons tester l'algorithme de réduction de dimension UMAP sur le dataset MNIST. Pour cela, nous décidons de nous comparer à un algorithme KNN.

1. **Préparation**
   (a) Construire une fonction qui prend en paramètre un modèle (type sklearn), une matrice d'information X et une réponse y, qui calcule le score moyen et la variance associée à une cross-validation stratifiée les f1-score (moyenné *macro* avec sklearn)
   (b) Construire une fonction qui prend en paramètre une matrice embeddings de deux colonnes, afficher le scatter plot de cette matrice où les axes représente les colonnes. La fonction prend en paramètre également un vecteur y et la possibilité de définir un titre.
2. Donner des baselines avec un KNN avec plusieurs nombres de voisins en utilisant les données MNIST.
3. Visualiser le résultat d'une réduction de dimension avec UMAP en prenant comme hyperparamètre :
   - k : 5
   - min_dist : 0.1
4. Utiliser la projection précédente pour mesurer la performance de l'algorithme KNN sur ces données.
5. Reprendre les deux points précédents en utilisant cette fois UMAP en mode supervisé, *mutatis mutandis*, commentez.
6. Challenger les valeurs choisies pour les paramètres de UMAP et du KNN final afin d'obtenir une meilleur classification.