

TP 10 – Réseaux de Kohonen

(SOM) – Jeu de données

Wine

NOM : GHODBANI

PRÉNOM : Fares

Groupe : Miage

Réalisé avec Canva,sklearn-som et Visual studio code.

1. Introduction

Le but de ce TD était d'utiliser un réseau de Kohonen (Self-Organizing Map – SOM) pour analyser et visualiser le jeu de données Wine fourni par sklearn.

Chaque échantillon est représenté par 13 caractéristiques et appartient à l'une des 3 classes de vin.

Le SOM permet de projeter les données dans une grille bidimensionnelle, de manière à ce que des points similaires soient proches sur la grille, facilitant ainsi la visualisation et l'analyse des clusters.

2. Méthode

- Grille SOM : 16×16 neurones (choix optimal pour compromis résolution/lisibilité).
- Dimension d'entrée : 13 (nombre de caractéristiques du dataset).
- Nombre d'époques : comparé 1 époque et 100 époques pour observer l'effet de l'entraînement.
- Visualisation :
 - Scatter plot avec jitter pour distinguer les points superposés.
 - Carte des distances (distance map) avec imshow pour visualiser l'organisation des neurones.

3. Résultats

3.1 Scatter plots

- Après 1 époque :

- Les points sont peu organisés.
- Les clusters correspondant aux classes de vin sont peu distincts.
- Cela montre que le réseau a besoin de plus d'entraînement pour structurer les données.
- **Après 100 époques :**
 - Les points se regroupent mieux par couleur (classe).
 - Les clusters sont plus distincts, avec des zones homogènes pour chaque classe.
 - La topologie de la grille commence à représenter correctement les similarités entre échantillons.

(Inclure dans le zip : les figures scatter 1 epoch et scatter 100 epochs.)

3.2 Carte des distances (Distance map)

- La carte des distances montre les différences entre neurones voisins.
- Les zones sombres indiquent des neurones proches (similaires), les zones claires montrent des différences importantes.
- Permet de visualiser les séparations entre clusters et identifier d'éventuels outliers.

(Inclure dans le zip : figure distance map.)

4. Remarques

- L'entraînement du SOM améliore fortement la visualisation des clusters avec un nombre d'époques suffisant.
- La taille de la grille influence la précision et la lisibilité :
 - Trop petite : clusters fusionnent.
 - Trop grande : points dispersés, moins lisible.
- L'ajout de jitter est indispensable pour distinguer les points projetés sur le même neurone.
- La normalisation des données (StandardScaler) peut améliorer la convergence et la séparation des clusters, surtout si certaines variables ont des échelles très différentes.

5. Conclusion

Le SOM est un outil puissant pour visualiser et analyser des données multidimensionnelles.

Après 100 époques, la carte 16×16 a correctement représenté les trois classes de vin avec une séparation claire.

Ce TP montre également l'importance de choisir correctement les paramètres du SOM (grille, nombre d'époques, normalisation) pour obtenir une visualisation significative.