Mathématiques spéciales — le 7 Janvier 2026

TP n°1 - Programmation de l'algorithme KNN

Dans ce T.P., nous allons programmer l'algorithme des k plus proches voisins pour un étiquetage d'un ensemble O inclus dans $\mathbb{R}^n$ muni de la distance euclidienne.

1. Création d'un ensemble d'entraînement

Q1 : Écrire une fonction `blob(n:int) -> list` qui renvoie une liste de `n` points tirés aléatoirement et uniformément dans $[0,1]^2$.

Q2 : Écrire une fonction `etiquettes(b:list) -> list` qui renvoie une liste d'étiquettes pour une liste de la forme `b=blob(n)` tirées aléatoirement grâce à la loi suivante :

> pour $a=(x,y)\in[0,1]^2$, l'étiquette de a est "rouge" avec probabilité $p(a)=e^{-20\|a-(\frac{1}{2},\frac{1}{2})\|_2^2}$.

2. Programmation de l'algorithme

Q1 : Écrire une fonction `distance(a:tuple,b:tuple) -> float` qui renvoie la distance euclidienne entre les vecteurs a et b de $\mathbb{R}^n$ représentés par les `tuple` `a` et `b` de longueur n.

Q2 : Écrire une fonction `k_min(L:list,k:int) -> list` qui renvoie les **indices** des k plus petites valeurs de la liste `L` qui est une liste de `float` de longueur plus grande que k.

Q3 : Écrire une fonction `knn(a:tuple,E:list,C:list,k:int) -> str` qui renvoie l'étiquette prédite du point `a` par l'algorithme des k plus proches voisins à partir de l'ensemble d'entraînement E représenté ici par le couple (`E`,`C`) où `E` est la liste des coordonnées des éléments de E et `C` la liste des étiquettes (`"red"` ou `"blue"`) des éléments de E.

3. Validation croisée et matrice de confusion

Q1 : Écrire une fonction `vc(E:list,C:list,k:int,np:int) -> float` qui découpe E en `np` morceaux de taille égales puis qui calcule le taux d'erreurs de l'algorithme des k plus proches voisins en prenant pour ensemble de test T chaque tranche du découpage et comme ensemble d'apprentissage le reste. La fonction renverra le taux d'erreurs moyen de l'algorithme.

Q2 : Écrire une fonction `mc(E:list,C:list,k:int,npt:int) -> tuple` qui renvoie la matrice de confusion (sous forme `(VP,FN,FP,VN)`) du découpage en deux sous-ensembles d'apprentissage A et de test T de E où T est contitué de `npt` points tirés aléatoirement et uniformément dans E et A le reste des points.