ESIEA – Systèmes d'information 4ème année
Réseaux de neurones
ali.ziat@lip6.fr

Feuille d'exercices 2

Exercice I
Prétraitement, évaluation

Question (I.1) Définir une fonction `standardize`, qui prend en argument une matrice `mat` et qui sur chaque colonne, normalise toutes les valeurs entre 0 et 1.
Pour ce faire, pour chaque colonne on doit avoir : $Xs = (X - min)/(max - min)$

Question (I.2) Définir la fonction `Split` qui prend en argument une matrice `mat` et un nombre `taux` qui vaut entre 0 et 1 .
Cette fonction sépare la matrice `mat` en deux sous-matrices `train` et `test` de taille respective `taux` et 1 - `taux` et les renvoie. Ainsi, `train` est un sous-ensemble **random** des lignes de `mat` et `test` est constitué des autres lignes.

Question (I.3) Définir une fonction `accuracy` qui prend en argument deux vecteurs qui renvoie leur taux de valeurs égales (taux entre 0 et 1).

Exercice II
KNN

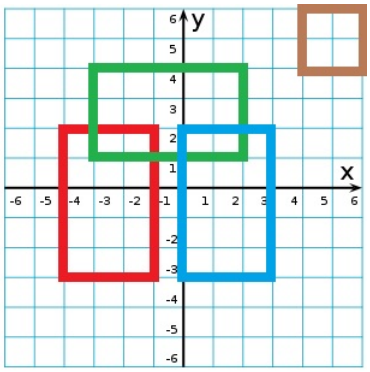
Question (II.1) Écrivez la fonction `knn` qui implémente l'algorithme des K plus proches voisins. Ses arguments sont `vec` qui représente le vecteur de l'exemple que l'on veut prédire, `X` qui est la base d'apprentissage, `k` qui est le nombre de voisin à prendre en compte, et `problem` qui permet de préciser si c'est un problème de régression ou de classification. Le pseudo-code de l'algorithme se définit simplement par :

- Je calcule la distance entre mon exemple à prédire et tous les exemples d'apprentissage
- Je récupère les K-exemples d'apprentissage les plus proches.
- Je prédis la valeur de mon exemple en fonction de la majorité des K plus proches (si problème de classification) ou la moyenne des K plus proches (si problème de régression).

Question (II.2) Testez l'algorithme sur des données de régression et de classification

Exercice III
K-moyenne

Question (III.1) Création d'un jeu de données de test en 2D :
Générez une matrice de points randoms avec 25 points dans chacune des zones délimitées par un rectangle dans l'image du repère qui suit.



Question (III.2) Écrivez la fonction `kmean` qui implémente l'algorithme des K-moyennes. Ses arguments sont `X` qui est la base de données et `k` qui est le nombre de clusters à trouver.
Le pseudo-code de l'algorithme se définit par :

- J'initialise aléatoire K centre de clusters.
- Tant que les centres de clusters continuent à varier :
  - J'assigne le centre de cluster le plus proche pour chaque point de X
  - Je recalcule les centres de clusters comme le centre de tous les points assignés à chaque cluster

Question (III.3) Testez l'algorithme sur les données 2D définies à la question 1.