K plus proches voisins*

INFORMATIQUE COMMUNE - TP n° 3.5 - Olivier Reynet

À la fin de ce chapitre, je sais :

- ☞ importer des données stockées depuis un fichier de type csv
- ☞ coder l'algorithme knn pour classifier des données étiquetées
- ☞ visualiser la performance de l'algorithme de classification

A Préparation des données pour la classification

On dispose de trois jeux de données étiquetées sous la forme de fichiers au format csv :

1. `iris.csv` pour la classification de variétés d'iris : iris.csv. Les étiquettes (variety) 0,1 et 2 correspondent aux variétés Setosa, Versicolor et Virginica.
2. `diabetes.csv` pour la prédiction du diabète : l'étiquette (Outcome) 1 signifie que le patient souffre de diabète, 0 qu'il n'en souffre pas.
3. `bdiag.csv` pour la prédiction de tumeurs cancéreuses : l'étiquette (diagnosis) 0 signifie que la tumeur est bénigne, 1 maligne.

Ⓡ La classe de chaque échantillon est toujours la **dernière colonne du fichier**.

A1. Coder une fonction de prototype `import_csv(filename)` où `filename` est une chaîne de caractères décrivant le nom d'un fichier. Cette fonction renvoie deux objets :

1. la liste `E` des paramètres de chaque échantillon (ligne) sous la forme d'une liste de liste de flottants,
2. la liste `C` des classes de chaque échantillon, sous la forme d'une liste d'entiers.

Vérifier que cette fonction est opérationnelle sur les deux jeux de données.

On souhaite mélanger les données pour que les classes n'apparaissent pas regroupées. Pour cela, on peut utiliser la fonction suivante qui réunit, modifie l'ordre de deux listes simultanément puis renvoie les deux listes (conservant ainsi la cohérence des données) :

```
def mixid(E,C):
    zipped = list(zip(E, C))
    random.shuffle(zipped)
    E, C = zip(*zipped)
    E, C = list(E), list(C)
    return E,C
```

*from scratch!

Afin de mesurer a posteriori l'efficacité du jeu de données, on crée deux jeux de données à partir des données initiales : un jeu d'entrainement et un jeu de test.

A2. Écrire une fonction de signature `split_train_test(E,C,ratio)` qui renvoie les listes :

1. `Xtrain` le jeu d'échantillons d'entrainement,
2. `Ytrain` les classes des échantillons d'entraînement,
3. `Xtests` le jeu d'échantillons de test,
4. `Ytests` les classes des échantillons de test.

Le paramètre `ratio` est un nombre compris entre 0 et 1 qui précise le rapport entre la taille des données d'entrainement et la taille du jeu de données. Par exemple, un paramètre `ratio` à 0.7 sépare le jeu de données en 70% pour l'entrainement et 30% pour les tests. On pourra utiliser le tranchage de liste (slicing).

B KNN pour la classification

Algorithme 1 k plus proches voisins (KNN)

Fonction KNN(k, E, x, δ)
  $n \leftarrow |E|$
  $\Delta \leftarrow \emptyset$ ▷ Distances à calculer
  pour chaque échantillon étiqueté $e \in E$ **répéter**
    Ajouter $\delta(x, e)$ à Δ
  Sélectionner les k voisins les plus proches de x en utilisant Δ
  Compter le nombre d'occurrences de chaque classe des k voisins de x
  renvoyer la classe c la plus représentée parmi les k plus proches voisins

B1. Écrire une fonction de signature `de(A : list[float],B : list[float])-> float` qui renvoie la distance euclidienne entre deux vecteurs de dimension n.

B2. Écrire une fonction de signature `k_plus_proches(k : int, E : list[list[float]], X : list[float])-> list[int]:` qui renvoie la liste des k plus proches voisins de X dans E. On utilisera la distance euclidienne. On peut utiliser un tri des distances et un tri des indices des échantillons conforme au tri des distances, ou bien trier directement une liste de tuples `(dist, indice_ech)` d'après le premier élément du tuple. Plusieurs solutions sont possibles :

- tri par sélection partiel : on s'arrête aux k plus petits ($O(nk)$),
- ★ utiliser un tas (--→ HORS PROGRAMME),
- modifier le tri rapide afin qu'il sélectionne les k permiers (Quick Select, $O(n)$ en moyenne).

B3. Écrire une fonction de signature `classe_majoritaire(K : list[int], Ytrain : list[int], N : int)-> int` qui renvoie la classe majoritaire d'un échantillon X d'après la liste de ses k plus proches voisins `K`. `N` est le nombre de classes, `Ytrain` les classes des échantillons du jeu d'entrainement.

B4. Écrire une fonction de signature `knn_test(filename, N, ratio)` où `N` est le nombre de classes, `ratio` la proportion de données d'entrainement. Cette fonction renvoie la matrice de confusion associée aux prédictions de KNN sur le jeu de tests.

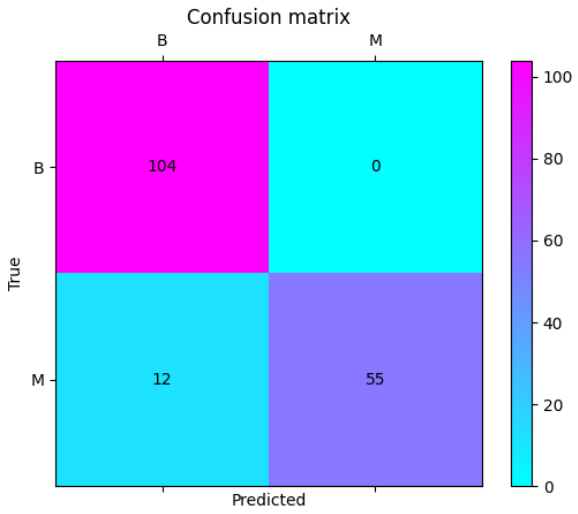

FIGURE 1 – Matrice de confusion pour la détection de tumeurs cancéreuses

B5. Tester l'algorithme sur les différents jeux de données et tracer les matrices de confusion associées à l'aide du code suivant :

```
from matplotlib import pyplot as plt

def draw_confusion_matrix(cm, classes_labels):
    fig = plt.figure()
    ax = fig.add_subplot(111)
    cax = ax.matshow(cm, cmap='cool')
    plt.title('Confusion matrix')
    fig.colorbar(cax)
    ax.xaxis.set_ticks([i  for i in range(len(classes_labels))])
    ax.set_xticklabels(classes_labels)
    ax.yaxis.set_ticks([i  for i in range(len(classes_labels))])
    ax.set_yticklabels(classes_labels)
    plt.xlabel('Predicted')
    plt.ylabel('True')
    for i in range(len(cm)):
        for j in range(len(cm[0])):
            ax.text(j, i, '{:d}'.format(cm[i][j]), ha='center', va='center')
    plt.show()
```

Vous devez obtenir des résultats proches de ceux inscrits sur la figure 1.

C KNN pour la régression

On dispose d'un jeu de données `Short_Student_Performance.csv` permettant de connaître un indice de performance (nombre flottant) d'un étudiant en fonction de paramètres tels que le nombre d'heures de travail ou le nombre d'heures de sommeil. L'objectif de cette section est de prédire la performance d'un étudiant, c'est-à-dire d'effectuer une régression.

C1. Modifier le code d'importation du fichier pour prendre en compte l'indice de performance correctement (car c'est un flottant).

Pour effectuer une régression, on ne calcule plus la classe majoritaire. À la place, on cherche à calculer la moyenne des indices de performance des k plus proches voisins.

C2. Modifier le code pour effectuer une régression sur l'indice de performance.

C3. Analyser les résultats en faisant varier la valeur de k et en calculant la racine carrée de l'erreur au carré (RMSE). Quelle valeur de k faudrait-il choisir ?