SY09 P2025
TD/TP 5 — Classification hiérarchique

`numpy`=2.2.3; `seaborn`=0.13.2; `matplotlib`=3.10.1; `pandas`=2.2.3; `sklearn`=1.6.1

Pour ce TP, il sera nécessaire de disposer d'une version récente de `scikit-learn` (au moins 0.22.1).

1 Travaux pratiques

1.1 Visualisation des données `Mutations`

On rappelle que l'AFTD calcule une représentation multidimensionnelle, dans un espace euclidien de dimension $p \leq n$, de données se présentant sous la forme d'un tableau $n \times n$ de dissimilarités δ_{ij} entre n individus ($i, j \in \{1, \ldots, n\}$), dont le tableau de dissimilarités ne donne qu'une description implicite. Cette représentation est exacte lorsque les dissimilarités sont des distances euclidiennes.

Une fois que des variables ont été retenues, la qualité de la représentation peut être évaluée numériquement par un critère similaire au pourcentage d'inertie de l'ACP, ou graphiquement au moyen d'un *diagramme de Shepard* représentant la distance $d_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ entre les représentations de $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ déterminées par l'AFTD en fonction de la dissimilarité initiale δ_{ij}, pour tout couple $(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

1 Charger les données `Mutations`. Vérifier que le tableau chargé est bien carré.

Pour réaliser une AFTD avec `scikit-learn`, il faut charger la classe `MDS` avec l'instruction suivante

```
from sklearn.manifold import MDS
```

Il faut ensuite instancier cette classe en spécifiant la dimension de la représentation voulue avec l'argument `n_components` et préciser que les données sont fournies sous la forme d'un tableau de distance avec l'argument `dissimilarity='precomputed'`.

Il suffit ensuite d'appeler la méthode `fit_transform` en fournissant le tableau de distance en argument. La méthode renvoie les coordonnées de la nouvelle représentation.

2 Calculer une représentation euclidienne des données en $d = 2$ variables par AFTD.

3 Afficher la nouvelle représentation en deux dimensions. On pourra utiliser la fonction `add_labels` du TP03.

4 Afficher le diagramme de Shepard avec la fonction fournie `plot_Shepard`. Que peut-on dire ? Recommencer avec $d \in \{3, 4, 5\}$. Interpréter les résultats.

5 Retrouver le « stress » avec les distances fournies par `plot_Shepard`. On rappelle que la fonction « stress » que cherche à minimiser `scikit-learn` est définie par

$$\text{Stress} = \sum_{i<j} (d_{ij} - \delta_{ij})^2.$$

1.2 Classification ascendante hiérarchique

La bibliothèque `scikit-learn` dispose d'un algorithme de classification ascendante hiérarchique. Pour cela, il faut importer la classe suivante :

```
from sklearn.cluster import AgglomerativeClustering
```

Il faut ensuite instancier cette classe. Les paramètres qui nous intéressent sont
- `linkage` : le critère d'agglomération,
- `metric` : la distance utilisée pour calculer le critère d'agglomération.

Par exemple, pour faire la classification ascendante hiérarchique d'un tableau individus-variables `X` en utilisant la distance euclidienne et le critère d'agglomération maximum, on construit l'objet

```
cls = AgglomerativeClustering(linkage="complete", metric="euclidean")
```

Pour apprendre la classification, il faut utiliser la méthode `fit` en fournissant le jeu de données `X`

```
cls.fit(X)
```

Pour visualiser la hiérarchie indicée obtenue, la fonction `plot_dendrogram` fournie dans le fichier `src/utils.py` pourra être utilisée. Par défaut, `scikit-learn` ne calcule pas les distances permettant de tracer la hiérarchie. Il faut fournir les paramètres `distance_threshold` et `n_clusters` initialisés respectivement à `0` et `None`.

6 Effectuer une classification ascendante hiérarchique des données `Iris`. Commenter les résultats obtenus, en vous appuyant sur votre connaissance de ce jeu de données.

Lorsque les distances entre individus sont déjà calculées, le paramètre `metric` est inutile. On lui attribue la valeur `"precomputed"` et au lieu de fournir le tableau individu-variable à la méthode `fit`, on lui donne directement le tableau de distances.

7 Effectuer la classification hiérarchique ascendante des données de `Mutations` (avec les différents critères d'agrégation disponibles). Commenter et comparer les résultats obtenus, en vous appuyant sur la représentation obtenue par AFTD.

Dans certain cas, il n'est pas nécessaire voire souhaitable de mener la classification ascendante hiérarchique à son terme. Pour cela, on dispose des arguments `n_clusters` et `distance_threshold`. L'argument `n_clusters` signifie qu'il faut arrêter l'agglomération dès qu'on a obtenu `n_clusters` groupements. L'argument `distance_threshold` signifie qu'il faut arrêter l'agglomération dès que le critère d'agglomération descend sous ce seuil.

Dès lors, les attributs suivants sont disponibles depuis l'objet `cls`
- `n_clusters_` : le nombre de groupements obtenu,
- `labels_` : les étiquettes des individus données sous forme de nombre entier indiquant l'appartenance à un groupement.

8 Donner une partition en deux groupes à partir d'une classification ascendante hiérarchique. Visualiser cette partition en utilisant une AFTD en deux dimensions.

1.3 Inertie intra-classe et critère de Ward

Dans cette partie, on cherche à montrer expérimentalement la relation qui existe entre l'inertie intra-classe et le critère d'agglomération de Ward lors d'une classification ascendante hiérarchique. Plus précisément, on a

$$I_W(P') - I_W(P) = \frac{1}{n} D_{\text{Ward}}(A, B),$$

avec P la partition avant fusion, P' la partition après fusion et A et B les deux groupements minimisant le critère d'agglomération de Ward et qui vont être fusionnés.

On va d'abord chercher à calculer l'inertie intra-classe avant chaque fusion lors d'une classification ascendante hiérarchique. On va jouer sur le paramètre `n_clusters` de la classe `AgglomerativeClustering` et récupérer la classification pour calculer son inertie intra-classe.

9 Créer une fonction `inertia` qui prend en argument un sous-jeu de données représentant un groupement ainsi que la taille du jeu de données total et renvoie l'inertie de ce sous-jeu de données.

On pourra utiliser la fonction `np.cov` ainsi que la trace `np.trace`. Attention toutefois au cas où le sous-jeu de données est réduit à un individu.

On contrôlera la justesse de la fonction avec les assertions suivantes :

```
import math
import seaborn as sns

iris = sns.load_dataset("iris")
iris0 = iris.drop(columns="species")
n = iris0.shape[0]
assert math.isclose(inertia(iris0, n), 4.542470666666669)
assert math.isclose(
    inertia(iris0.loc[iris.species == "setosa"], n), 0.10100666666666669
)
assert math.isclose(
    inertia(iris0.loc[iris.species == "versicolor"], n), 0.20410933333333345
)
assert math.isclose(
    inertia(iris0.loc[iris.species == "virginica"], n), 0.29019999999999996
)
```

10 En utilisant la fonction précédente, créer une fonction `intra_class` qui prend en argument un nombre `n_clusters` de groupements et un jeu de données et renvoie l'inertie intra-classe de la CAH avec critère de Ward en `n_clusters` groupements.

On pourra utiliser la méthode `groupby` sur le jeu de données afin de grouper le jeu de données par groupements et ensuite appliquer avec la méthode `apply` la fonction précédente.

Créer ensuite la liste des inerties intra-classe pour un nombre de groupements maximum jusqu'à un nombre de groupement minimum (c'est à dire 1).

11 L'argument `distances_` présent lorsque la classification ascendante hiérarchique est menée à son terme contient tous les critères d'agglomération successifs. Au lieu du critère de Ward c, `scikit-learn` renvoie la quantité $\sqrt{2c}$). Ce recalage est nécessaire si on veut que le critère d'agglomération entre deux feuilles coincide avec la distance entre deux feuilles.

Calculer les critères de Ward successifs.

12 Vérifier que les inerties intra-classe et les critères de Ward sont liés par la relation

$$I_W(P') - I_W(P) = \frac{1}{n} D_{\text{Ward}}(P, P').$$

1.4 Euclidianité et euclidianisation

L'AFTD permet de trouver une représentation euclidienne si elle existe. Au passage, on dispose d'un test pour savoir si une dissimilarité admet une représentation euclidienne. En effet, si D est une matrice de dissimilarité, elle est euclidienne si et seulement si la matrice W_D suivante

$$W_D = -\frac{1}{2} Q_n D_2 Q_n, \tag{1}$$

est semi-définie positive avec Q_n la matrice de centrage et D_2 la matrice D où chaque entrée est mise au carré. Il suffit donc de calculer la valeur propre la plus petite de W_D et de vérifier si elle est positive ou non.

13 Créer deux fonctions, `Wd` et `smallest_eigenvalue` qui renvoient respectivement la matrice W_D et la plus petite valeur propre de W_D.

On pourra utiliser la fonction `eigvalsh` disponible dans le module suivant

```
import scipy.linalg as linalg
```

qui calcule les valeurs propres d'une matrice symétrique.

14 On considère la matrice de distance suivante

$$D = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & a \\ 2 & a & 0 \end{pmatrix}$$

À quelle condition sur a la matrice D est-elle euclidienne ?

15 À l'aide de la fonction `smallest_eigenvalue`, vérifier expérimentalement le résultat précédent.

À partir d'une dissimilarité quelconque D, on définit une autre dissimilarité D^γ comme suit

$$D^\gamma_{ij} = \begin{cases} 0 & \text{si } i = j, \\ D_{ij} + \gamma & \text{sinon,} \end{cases}$$

avec $\gamma > -\min_{i \neq j} D_{ij}$.

16 Montrer expérimentalement que D^γ est une matrice euclidienne à partir d'une certaine valeur de γ.

17 Vérifier expérimentalement que cette valeur est la plus grande valeur propre de la matrice suivante

$$B = \begin{pmatrix} 0 & 2W_D \\ -I & -4W_{D^{1/2}} \end{pmatrix},$$

avec $W_{D^{1/2}}$ la matrice décrite par (1) avec la matrice D au lieu de D_2.

2 Exercices théoriques

2.1 AFTD des données `Mutations`

On reprend le problème de la représentation des données `Mutations`.

18 Dans le diagramme de Shepard, vérifier que l'inertie I_{b_1} du nuage de points par rapport à la première bissectrice b_1 vérifie

$$\text{Stress} = 2mI_{b_1},$$

avec m le nombre de points dans le diagramme de Shepard. Pourquoi ce résultat ?

19 Peut-on choisir la matrice M de telle sorte que $\text{Stress} = mI_{b_1}$?

2.2 Classification ascendante hiérarchique

20 On considère le tableau individus-variables suivant :

$$X = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ 3 & 5 \\ 5 & 2 \\ 7 & 3 \end{pmatrix}$$

Calculer la matrice de distances euclidiennes associée ; faire une CAH avec lien minimum, puis lien maximum.

21 Prouver les formules de récurrence de Lance & Williams, pour les trois critères d'agrégation du lien minimum, du lien maximum, et du lien moyen :

$$\begin{aligned} D_{\min}(A, B \cup C) &= \min\{D_{\min}(A, B), D_{\min}(A, C)\}, \\ D_{\max}(A, B \cup C) &= \max\{D_{\max}(A, B), D_{\max}(A, C)\}, \\ D_{\text{moy}}(A, B \cup C) &= \frac{n_B D_{\text{moy}}(A, B) + n_C D_{\text{moy}}(A, C)}{n_B + n_C}. \end{aligned}$$