The metastable states of a peptide in water solution (Mandatory choice for PCBS students. Optional for the others)

From the link
`https://github.com/alexdepremia/Unsupervised-Learning-Datasets.git`
download the file high_variance_distances.dat.gz. It corresponds to a time series of the value of distances between pairs of atoms belonging to different residues in the peptide of sequence Arg-Phe-Phe-Glu-Ala. The chosen distances are all those that during the dynamics vary significantly (variance larger than 5 Å^2. This peptide is a fragment of α-synuclein, and studying its conformational dynamics is important to understand Alzheimer disease.

Build a Markov State model from this time series, following, for example, this pipeline:

1. Find the microstates by a cluster analysis performed with k-means or k-medoids.

2. Choose a time lag τ, and estimate the transition matrix between each pair of clusters $\pi_{\alpha,\beta} = P(\beta, \tau|\alpha, 0)$

3. By analyzing the spectrum choose an appropriate number of Markov states. Find the Markov states by inspecting the sign of the eigenvectors.

4. Perform a dimensional reduction with a method of your choice, retaining only two coordinates, and visualize the Markov states in the space of these coordinates.

5. Discuss the stability of the results with respect to the meta-parameters of the approach, in particular, the time lag τ and the number of clusters.

Unsupervised learning on Individual household electric power consumption (optional choice for UniTS and Sissa DS students)

Download the dataset from the link `https://archive-beta.ics.uci.edu/dataset/235/individual+household+electric+power+consumptionAnalizzando`. Read carefully the description of the dataset before performing the analysis. In each entry, the first column is date, the second time and the rest are consumption-related measurements.

1. Data pre-treatment: You can use as data points each time measure or the average value per day. Propose a coherent way for dealing with missing values and, if needed, a way for decimating the data.

2. Perform a cluster analysis with k-means or with a density-based clustering.

3. Perform a dimensional reduction with a method of your choice, retaining only two coordinates, and visualize the clusters in the space of these coordinates.

4. Now recompute the k-means clusters, but with the goal of performing Markov State Modeling (hint: the number of clusters necessary to perform MSM is typically larger).

5. Choose a time lag τ, and estimate the transition probability matrix between each pair of the clusters found in point 4. $\pi_{\alpha,\beta} = P(\beta, \tau|\alpha, 0)$

6. From the spectrum compute the relaxation times of the system. Choose an appropriate number of Markov states. Find the Markov states by inspecting the sign of the eigenvectors.

7. Comment if the MSM clusters have some meaning.

Be ready to discuss if the clusters obtained in point 2, the relaxation times and the Markov States obtained in point 6 can be interpreted in relation with the nature of the dataset.

Unsupervised learning on a dataset of your choice (optional choice for UniTS and Sissa DS students)

Perform a data analysis on a data set on which you have already worked in your previous research projects, or on which you plan to work in the future.

In the analysis implement and use methods of your choice from the following list, up to a total number of credits of at least 5.

- TwoNN intrinsic dimension estimator (1 credit)
- Clustering with k-means, or fuzzy k-means (1 credit)
- hierarchical clustering (1 credit)
- Density-based clustering (2 credits)
- Dimensional reduction with a method of your choice, retaining only two coordinates, and visualizing the clusters in the space of these coordinates (1 credit)
- Markov State modeling, if the data set is a time series (in this case follow the pipeline of the other exercises). (3 credits)

Be ready to provide an interpretation of your results in relation with the nature of the dataset.