**Quantifying natural behavior – the road toward computational neuroethology**

The final project has two parts:

Supervised Learning (created by Ehud Sussman, [ehudsussman@gmail.com](mailto:ehudsussman@gmail.com)) and Unsupervised Learning (created Yizhak Goussha, [gutzcha@gmail.com](mailto:gutzcha@gmail.com) ).

Deadline for submission: **17.11.2024, 23:59**

1:1 meeting to evaluate your understanding of the code: **25-26.11.2024**

Please register for 15 minutes with Ehud and 15 with Yizhak [HERE](https://docs.google.com/spreadsheets/d/1Y5nkKJPnEB8un09BB9kIajSmCyzyCRhB2Y6VPe2L4qI/edit?usp=sharing)

**Part 1: Supervised Learning Assignment**

**General Guidelines for Assignment for Supervised Learning Component**

You must submit your solutions as a Jupyter Notebook for this assignment in both .html and .ipynb formats, as well as .csv file that will be called blind\_test\_prediction.csv. The .html format will be reviewed for readability, while the .ipynb format to verify solutions, the csv file will be check for accuracy. The following instructions outline the key components and guidelines for preparing the notebook:

**Notebook Structure and Content**

1. **Headers and Subheaders**:
   1. Organize the notebook into clearly defined sections using headers and subheaders.
2. **Graphs and Visualizations**:
   1. Include well-labeled graphs with appropriate titles, x and y-axis labels, and scales.
   2. Specify the colors used in each graph to ensure clarity and consistency.
   3. Each graph also includes a subsequent markdown cell to explain the graph's purpose and relevance to the analysis.
3. **Work individually or in pairs**:

We will contact each student to explain the work done individually or in pairs with questions to verify that the student understands the solution submitted in depth. We will only accept assignments submitted jointly by up to two students or submitted individually.

1. **Grading:**

Most of the grading in this section will be on evidence of applied learning in the course. Explain the reasoning behind the methods selected thoughts. Provide visual evidence to confirm the explanations. Achieving a high level of success on performance measures may yield additional evidence of correct thought processes; however, success in performance measures is less significant in grading. Using tools you know or found online that are beyond what was taught in this course is acceptable; however, if you use such tools, please explain them at a level that a freshman majoring in engineering or hard sciences could follow as we want to follow what you are doing as well. Also, we will ask you verbally on components of your solutions so be prepared to demonstrate command of the tools you used.

**The Data**

The supervised learning component of the homework assignment involves analyzing a dataset structured for behavioral analysis of pairs of flies. The dataset consists of 53 Columns: the first 52 Columns represent the features extracted from the videos with the SLEAP package, while the last Column represents the labels. These labels are binary (either 0 or 1), indicating different populations of fly pairs.

**Data Structure**

1. **Feature Columns (1-52)**:
   * The first 52 Columns contain features organized into 13 body parts of the flies, each with an (x, y) pixel coordinate tracked by SLEAP (Social LEAP Estimates Animal Poses).
   * Each body part has separate coordinates for the male and female flies, resulting in 13 body parts \* 2 (x and y axis of coordinate) \* 2 (male and female) = 52 feature Columns.
2. **Label Column (53rd Column)**:
   * The 53rd Column contains the labels, which are either 0 or 1. These labels represent different populations or behavioral states of the fly pairs.
3. **Row Chunks**:
   * The data is in chunks of 150 consecutive Rows, where each chunk corresponds to 150 frames captured at 150 frames per second (fps). This structure provides 1-second snippets of video data, allowing for detailed analysis of fly behavior over short time intervals.
   * Within each chunk of 150 Rows, the label (0 or 1) remains the same since all frames within a chunk are from the same pair of flies.

**Dataset Composition**

1. **Training Set**:
   * The training set consists of 240,000 Rows, divided evenly between the two labels: 120,000 Rows with label 0 and 120,000 with label 1.
   * This set trains your supervised learning models and includes the full feature set with corresponding labels.
2. **Test Set**:
   * The test set contains 120,000 Rows, which are also evenly split between 60,000 Rows with label 0 and 60,000 Rows with label 1.
   * This set evaluates your trained model's performance on data it has not seen before but where the labels are known.
3. **Blind Test Set**:
   * The blind test set also consists of 120,000 Rows, structured identically to the test set, except no labels are provided.
   * This set is for final evaluation, where you will predict the labels based on your trained model. The actual labels are withheld to ensure an unbiased assessment.

**Additional Information**

* **Mutual Exclusivity**: The training set, test set, and blind test set are mutually exclusive regarding the pairs of flies they contain.
* **Repeated Pair Inclusions**: Within each set (training, test, and blind test), the same pairs of flies appear multiple times but from different parts of the video.
* **Label Dataframes**: Two smaller, one-Column DataFrames are provided, containing the labels for the 150-Row chunks for the training and test sets. Note that such a labeled DataFrame is not provided for the blind test set.
* **Link to Data**: <https://drive.google.com/drive/folders/1E9wNyuCGdY-m8DxAe4WcA7G0QFLzwHkk?usp=drive_link>

**Assignment Tasks**

1. **Data Cleaning (4 points)**:
   * Check for outliers or possible missing values in the dataset. (2 pts)
   * Describe the identification process for outliers. (2 pts)
   * Interpolation of outliers is not required.
2. **Create Features (20 points)**:
   * Create at least five features from the raw data; at least one is temporal. (12 pts)
   * Use at least two different window sizes for a temporal feature. (4 pts)
   * Visually display a temporal feature with two different window sizes on the same graph. (4 pts)
3. **Visual Testing of Features by Label (6 points)**:
   * Make graphs to test for differences in features by labeling visually. Ensure the graphs are well-labeled and explain each graph. (2 pts)
   * Which features are likely to be better predictors? Explain (4 pts)
4. **Train and Evaluate Models on Training Set (5 points)**:
   * Perform an 80%/20% train-test split and apply the following machine learning algorithms: Random Forest and one additional machine learning algorithm. (1 pts)
   * Include a confusion matrix and the accuracy measure. (4 pts)
   * Use all rows included (for labels) up to the limitation of a window size selected due to temporal data limitations.
5. **Evaluation Using Full Train and Test Data (5 points)**:
   * Apply the machine learning models and performance metrics from the previous section, but use all training and test sets data. Display the results for comparison. (3 pts)
   * Compare these results to the previous train-test split results and explain possible differences in performance. (2 pts)
6. **Enhanced Preprocessing (5 points)**:
   * Reapply the machine learning models (for training and testing data) with (an) additional processing step(s) to structure the data to fit short labels. Display the results for comparison. (5 pts)
7. **Blind Test Prediction (5 points)**:
   * Apply the enhanced preprocessing steps to the blind\_test\_set with the same additional preprocessing for the Random Forest model. Submit the

**Part 2: Unsupervised Learning Assignment**

**General Guidelines for Assignment for Unsupervised Learning Component**

You must submit your solutions as a Jupyter Notebook for this assignment in both .html and .ipynb formats, as well as a .csv file that will be called electrophysiology\_data\_kmeans.csv.   
The HTML file will be reviewed for readability, the ipynb file will be used to verify solutions, and the CSV file will be checked for accuracy.

The following instructions outline the key components and guidelines for preparing the notebook:

**Notebook Structure and Content:** Unless specifically stated otherwise for example the name of the csv file to submit grading guidelines for notebook structure and content will be the same as in the supervised learning component.

**The Data**

1. The unsupervised learning component of the homework assignment involves analyzing a dataset structured for behavioral analysis of subject rats in dyadic social experiments. The dataset consists of 8 columns representing some of the features extracted from electrophysiological data refined for this assignment by performing data imputation and cleaning. The features are extracted from electrophysiological measurements collected from brain areas important in mediating social behavior.
2. Link to data:   
   [https://drive.google.com/file/d/1c-VLLYdzG213Pyn0LcYuDNL-73SY1ZL6/view?usp=sharing](https://drive.google.com/file/d/1c-VLLYdzG213Pyn0LcYuDNL-73SY1ZL6/view?usp=sharing%20)

**Assignment Tasks**

1. **Dimensionality reduction (25 points)**:
   * Reduce the dimensions of the data to 2 dimensions using t-SNE, UMAP, and PCA
   * Perform the reduction using the following parameters and visualize each reduction:
     1. t\_SNE: perplexity level: [5, 100] (2 plots)
     2. UMAP: n\_neighbors: [2, 10], min\_dist: [0.1, 10] (4 plots)
     3. PCA (default) (1 plot)
   * Visualize the data for each reduction using a scatter plot
   * Explain how the different parameters affect the reduction
2. **Clustering (25 points):**
   * Cluster the data using K-Means, Agglomerative Clustering, and one more method of your choosing
   * For K-means clustering, evaluate the optimal number of clusters using the Elbow method
     1. Fit 10 K-mean models with the number of clusters ranging from 1-10
     2. Extract the inertia of each fitter model
     3. Plot the Elbow plot
   * Add a column to the data frame named "kmeans\_labels" and assign the labels extracted using the K-means method.