Q1: Implementing PASAD on SWaT and TE Datasets

1. **Dataset Analysis and Plotting (30 marks):**
    - Implement PASAD on the SWaT and TE datasets with the following parameter settings:
        - For the TE dataset, use initial measurements (ie. N=1600 to 2400 out of 4800 measurements) for training PASAD, and the remaining measurements for testing. For the SWaT dataset, select recent measurements from the normal dataset file for training and the entire attack dataset for testing. Ensure that the training and testing portions come from the same timeframe for each sensor within a dataset.
        - The choice of parameters should be justified to ensure superior performance. Ensure that the lag parameter "L" is consistent across all sensors within the dataset and that the reduced dimensionality parameter "r" (r>1) is derived from the significance of the vectors in the eigenspace.
    - Generate your preferred plots, one from the SWaT dataset and one from the TE dataset. Each plot should consist of two subplots:
        - The first subplot should display the sensor measurements.
        - The second subplot should show the corresponding departure scores.
        - Highlight the following in the plots: training measurements, attack measurements, and classifier threshold.
2. **Performance Comparison (10 marks):**
    - One of the key contributions of PASAD is projecting the lag vector onto the signal subspace using the linear map U^T instead of the projection matrix UU^T. This improvement significantly reduces the computation cost in online deployment. Implement both variants and report the following:
        - Measure the average run time (on both datasets) taken by each variant (ie. $P=U^T$ and $P=UU^T$) to generate a departure score for a single measurement.
        - Did you notice any change in the detection capability? Empirically justify your answer with the help of departure score plots, one from TE and one from the SWaT dataset.
3. **Attack Scenario Analysis in TE Dataset (10 marks):;**
    - The TE dataset contains five distinct attack scenarios, SA1, SA2, SA3, DA1, and DA2. Implement PASAD in each scenario to evaluate the total alarm count with a zero false alarm rate. Present the results as a bar plot, with attack scenarios on the x-axis and total alarm counts on the y-axis. To ensure no false alarms, set the classifier threshold at the maximum departure score observed during normal measurements while computing the alarm count.

Q2: Exploring the Use of Centroid Instead of Mean of the Cluster:

In PASAD, the departure score is calculated as the departure of a projected test point from the mean of the cluster. Consider replacing the mean with the centroid.

1. **Justification (10 marks):** Discuss whether this modification improves or degrades the method, and provide reasonings for your answer.
2. **Implementation and Comparison (10 marks):** Implement the modified departure score on the five attack scenarios of the TE dataset. Generate a bar plot comparing the alarm counts from the modified departure score with those from Q1.3. Ensure that the implementation setup, including the training data portion and parameters, is consistent with the previous implementation.

Q3: Exploring the Use of Mahalanobis Distance

1. **Justification (10 marks):** In PASAD, the departure score is calculated using Euclidean distance. What would be the impact of replacing it with Mahalanobis distance? Discuss the potential advantages and disadvantages of this modification.
2. **Implementation and Comparison (15 marks):** This variant uses Mahalanobis distance and the centroid of the normal cluster for calculating the departure score. Generate a bar plot comparing the results with those from the previous two methods (Euclidean distance with mean and Euclidean distance with centroid).
    - You may use the normal cluster generated from the training data's signal subspace to determine the covariance matrix for the Mahalanobis distance calculation.
3. **Runtime Analysis (5 marks):**
    - Measure the average runtime of this modified method to generate a single departure score and compare it with the other variants.

——---—---------- Instructions —----------

Submission includes the following:

- Well-structured implementation code with README in a zip file. Exclude the dataset.
- A report in PDF format answering the questions, including results and plots.

Dataset:

SWaT: "https://drive.google.com/drive/folders/1zn0DMCdSXA9b_CiaaoDzvkbwX5O4ikLn?usp=drive_link"

TE: "https://github.com/mikeliturbe/pasad/tree/master/data" (Use CSV files in Scenarios DA1, DA2, SA1, SA2, and SA3)

Programming Language: Python