Challenge for "Merit Prize" 2024/2025

Deadline 29/11/2024 23:59 via Fenix as PDF

Submit Gxxx.ZIP in Fenix where xxx is your group number. The ZIP should contain two files: Gxxx_report.pdf with your report and Gxxx_notebook.ipynb with your notebook demo according to the suggested templates

NB: This project is pertaining to the challenge "Merit Prize" and has no impact on the course grade. Also, it may rely on material that was not covered in the course, e.g., RBF ("Radial Basis Function") networks, but for which you should have the necessary background.

Consider the breast_cancer dataset

data = datasets.load_breast_cancer()

with binary target variable y='malignant'. Split it 70% for training and 30% for testing.

1) Perform logistic regression and indicate the accuracy.
2) Perform EM clustering on the training data set with different number *k* of clusters. Evaluate the quality of the clusterings using Silhouette. Is the number of clusters correlated with the quality of clustering? Which is the optimal *k*?
3) Map the test set into probability values of the *k*-clusters. If you have a data point represented by a vector of dimension *d*, you will map it into a vector of dimension:

   prob=em_model.predict_proba(X)

4) Perform logistic regression on the mapped data set with the labels of the original test set. Indicate now the accuracy. Is there a relation between the number of clusters, the cluster evaluation and the accuracy of the logistic regression model?
5) Train an RBF network using the clustering with optimal *k* from 2).
6) Discuss your findings on a (up to) 5 page document.