CME466 (2022-23) **Deliverable # 4: Machine learning problems** **Total marks: 20**

NOTE: For more information on the deliverables, please follow the lecture materials and in-class discussions. If you have further questions, please consult with the instructor(s).

Please complete this assignment in a Jupyter notebook.

1. Problem: Classifier Performance Evaluation and Parameter Tuning

For this problem, the dataset to be used is the **Iris** dataset.

a. For the module sklearn.metrics, discuss what other metrics should be applicable here, and compare your classifiers in terms of these metrics.

b. For the kNN, plot the accuracy metric as a function of the n_neighbors parameter. What is the optimal value? Does your answer differ depending on the validation strategy used to assess the performance? Explain your answer.

c. Design an SVM classifier for this dataset, and comment on the results.

d. Investigate the computational times for the various classifiers, in terms of both training and classification execution times. You should find the magic function %timeit useful.

2. Problem: SVM with non-linear kernels

For this problem, recall the synthetic dataset generated in the example notebook, using make_circles(100, factor=.1, noise=.1, random_state=0).

a. Design a suitable SVM classifier for this dataset. Justify your parameter choice and kernel used.

b. Investigate the effect of the amount of training used on the classifier design. For this purpose, you can consider plotting the testing performance as a function of the amount of training used. Comment on your findings.

3. Problem: Regression Estimator

For this problem, the dataset to be used is the **diabetes** dataset.

a. Design a suitable regressor for this dataset. You may consider alternatives among any built-in regressors (supported by scikit-learn). Justify your final selected design, including parameter selection and performance metric used.

b. Investigate the effect of the amount of training used on the regressor design. For this purpose, you can consider plotting the MSE testing performance as a function of the amount of training used. Comment on your findings.

4. Problem: Classifier Design and Imbalanced Datasets

For this problem, use the **digits** dataset.

a. Explain the meaning of an *imbalanced* dataset, and why it can be a problem in ML. Comment on the given dataset, with respect to this issue.

b. Explain the PROs/CONs of the accuracy score vs. the F1-score.

c. Design a suitable classifier for this dataset. You may consider alternatives among any built-in classifiers (supported by scikit-learn). Justify your final selected design, including parameter selection and performance metric used.

d. Explain the principles of K-fold cross validation. For the classifier selected in part (c), evaluate the performance using this method, and comment on your results.

Deliverables:

- A report containing:
  - answers to the above questions
  - your python codes (ipynb files)