IE 48B Time Series Analytics
Homework 3, due December 13th, 2021

Instructions: Please solve the following exercises using R (http://www.r-project.org/) or Python (https://www.python.org/). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

- A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
  - https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
  - https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps-917519326e41

  Things are little easier if you install Anaconda (https://www.anaconda.com/). Please export your work to an html file. Please provide your *. ipynb file in your repository and a link to this file in your html report will help us a lot.

- A Markdown html document. This can be created using RMarkdown for R and Python Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself. As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Task: Comparison of NN classifiers with alternative representations

The aim of this task is to compare alternative distance measures for classification. Recall that distance calculations can be done using raw time series as well as the represented versions of them. You are expected to implement a pipeline that can evaluate alternative distance measures on different representation settings.

1- Select five time series classification datasets from www.timeseriesclassification.com (your selections should have at least 100 training time series). Direct link to the datasets is http://www.timeseriesclassification.com/Downloads/Archives/Univariate2018_arff.zip. Note that you may want to restrict yourselves with the datasets that do not have large number of training and test instances (due to the computational requirements of certain distance-based approaches).

2- You are expected to evaluate two alternative representations (in addition to raw time series observations) and three alternative distance measures (in addition to Euclidean distance) in nearest neighbor (NN) classification setting. You also need to determine the only parameter of NN, namely *k*. For the setting of *k*, you can evaluate the set {*k=1, 3, 5*}. This strategy requires you to evaluate 3 representations x 4 distance measure x 3 *k* settings = 36 parameter combinations. Note that this calculation discards the parameters of distance measures (i.e. window type of DTW if it is considered as an alternative). You are expected to use the default parameters of selected distance measures. Similarly, representation approaches require certain parameters (i.e. interval size for piecewise aggregate approximation). For the representations, evaluate two alternative parameter settings of your choice. Finally, this will make 36 x 2 = 72 parameter combinations.

Please mention your choices and the default parameters of your selections explicitly. Evaluate your proposals based on 5 repeats of stratified 10-fold cross-validation and report the accuracy of each alternative. The aim of this evaluation is to set the parameters for testing purposes (i.e. parameter tuning). Compare the accuracy of your alternatives in terms of mean and standard deviation of the accuracy over the test folds for each alternative parameter combination.

Use the best parameter combination in order to perform classification on the test time series and report your test accuracies. You can present your results as a table where you have dataset name, cross-validation mean accuracy, test accuracy and the parameters of the best classification setting.