IE 582 Statistical Learning for Data Mining
Homework 5 (Optional), due February 18th, 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.

Question 1 (30 points)

Use the UWave data provided in the second homework for this exercise. In the second homework, you were given only the training data. For this homework, you are asked to perform classification task on the test data and evaluate the performance of certain classifiers. Test data is uploaded for each axis (i.e. X, Y and Z). It has the same format as the training data.

The dataset is provided in the following link:
https://drive.google.com/drive/u/1/folders/13553neknux7U8why55KM1WrjgkA9IJKm

a) Suppose we decided to apply a nearest-neighbor (NN) classifier to find the labels of test instances. Propose two distance measures for computing similarity between two time series. For each distance measure alternative, use the training data to identify the optimal value of *k* which minimizes the error of a 10-fold cross-validation.

b) Using the value of *k* (identified for each distance measure) in part (b) and evaluate your final performance on the test data and present your results in a (8-by-8) confusion matrix, showing the counts for actual and predicted labels. In addition, quote the runtime and accuracy for your results.

c) The observations from different axes are weighted equally if we compute the distance over each axis and sum them to obtain a final similarity measure. Is this reasonable? For example, we can compute the distance as below:

$$finalDist = w_1 Dist_x + w_2 Dist_y + w_3 Dist_z$$

where $Dist_x$ is the distance based on the acceleration only on *X* axis, $Dist_y$ is for *Y* axis and so on. Do you think weighting the distances over different axes to obtain a final similarity measure makes sense for classification? Why?

Question 2 (70 points)

In the provided link, you are also given ECG data. You get an R code on Moodle (titled as "NN classification example (Time series data)" under week named as "Jupyter Notebooks"). The script contains a sample NN code (that you can also use this code as an example for the first question).

a) Train a logistic regression model on the training data using fused lasso penalties. Use the learned model to predict the class for test data. Present your results in a (2-by-2) confusion matrix.

Do not forget to learn the parameters of fused lasso through cross-validation on training data. You just need to provide the necessary parameters as arguments to the corresponding function.

b) Comment on the regression coefficients. Is there any interesting information? Try to interpret the model.

c) Suppose you take the difference between consecutive time series observations to transform the time series. Let x_t be the observation at time point *t* and you are asked to create a new time series which is $y_t = x_t - x_{t-1}$ for $t = 2,3, \dots, T$ where *T* is the length of the time series. Perform the same operation in part (a) for this newly created dataset and compare the results.

d) Comment on the regression coefficients for the model trained on the new dataset. Is there any interesting information? Try to interpret the model.

e) Suppose you combine two datasets created for part (a) and (c) column wise and train the model on the combined dataset. Use the learned model on the combined dataset to predict the class for test data. Present your results in a (2-by-2) confusion matrix.

f) Comment on the regression coefficients for the model trained on the new dataset. Is there any interesting information? Try to interpret the model.