CSCE 633
Homework 5: Designing and disseminating ML for a real-world problem

Instructions for homework timeline and submission

a) Form teams of 5 classmates (approximately 5 members per team). Please email us your team names and UIN with a private message on CANVAS.
b) **Work with your teammates in this project.**
c) You will present your work with your team in class on **Tuesday 12/7** during class time (**3.55-5.10pm CT**). Most of the work should be ready by the class presentations, including the data analysis and the e-poster (question (f)). The material for question (g) will be obtained during in-class presentations.
d) The final report is due on **December 13, 2021 @ 11.59pm**. Please create a zip file with two pdf files, including the final report and the e-poster. **One member per team can make the submission.**
e) You can use any publicly available library for this homework.
f) The total for this homework is **15 points**.

The goal is to build machine learning models to estimate one's public speaking anxiety from bio-behavioral data. The data comes from the VerBio dataset, which was collected with the goal to better understand individuals' affective responses while performing public speaking tasks. More details about the dataset, including the experimental setup and type of data, can be found here: `https://hubbs.engr.tamu.edu/resources/verbio-dataset/`.

The data is provided in Google Drive in "Homework 5" folder and includes 55 participants. We have one presentation for each participant, which results in 55 data samples. You are provided with two types of data.

1. The "data.csv" file that contains a set of bio-behavioral features (i.e., skin conductance level, skin conductance response amplitude, skin conductance response frequency, heart rate, wrist acceleration, interbeat interval, speech energy, 12 speech mel-frequency cepstral coefficients (MFCCs), speech zero crossing rate, speech voicing probability, speech fundamental frequency (F0), speech pause frequency), label (i.e., state anxiety), and participants' language information (i.e., native/non-native English speaker). Each row in the csv file corresponds to one data point. The bio-behavioral features were computed based on the entire presentation for each participant.

2. A set of files titled "EDA_PPT_PXXX.xlsx," where XXX is the participant's ID, that include temporal values of participants' electrodermal activity (EDA) signals (i.e., sweat activity from the wrist) collected during the presentation sampled at 4Hz (i.e., 4 values per second).

3. A set of files titled "HR_PPT_PXXX.xlsx," where XXX is the participant's ID, that include temporal values of participants' heart rate (HR) signals collected during the presentation sampled at 1Hz (i.e., 1 value per second).

(a) (2 points) Data pre-processing and exploration. Identify missing data values and replace them with the corresponding feature mean. You can also experiment with any other feature imputation method. Provide visualizations of the features with respect to the state anxiety label (e.g., overlaying histograms, scatter plots), and quantify associations between the features and the label (e.g., via correlation coefficient).

(b) (2 points) Feature selection. Using the bio-behavioral features, explore two different feature selection methods of your choice to identify the features that are the most informative of the state anxiety label. One method should be part of the **Filter** category and the other should be part of the **Wrapper** category. Using a feedforward neural network (FNN), plot the absolute error between the actual and predicted state anxiety values using a 5-fold cross-validation (i.e., average over the 5 folds) against the number of features for both feature selection methods. Compare and contrast between the two (e.g., in terms of performance and computation time).

(c) (2 points) Feature transformation. Use Principal Component Analysis (PCA) to reduce the dimensionality of the bio-behavioral features from D to K. Use a 5-fold cross-validation with a FNN model and experiment with different numbers of resulting dimensions K. You will run an outer loop that examines different K values. Following that, you will run an inner loop, that implements the 5-fold cross-validation for a given K. You will use the training data of each fold to compute the transformation matrix (i.e., by computing the eigenvalues and eigenvectors based on the training data), which will be then used to also transform the test data of each fold. At the end of the inner loop, you will compute the average absolute error across the five folds. Provide a plot of the average absolute error for all the different K values that you have experimented with.

(d) (2 points) Working with time-series. Next you will be using the time series constructed by the EDA and HR measures (i.e., "EDA_PPT_PXXX.xlsx" and "HR_PPT_PXXX.xlsx") to estimate state anxiety, in two ways: (1) Fit the time series using linear and non-linear regression models (i.e., input is time and output is the EDA or HR value). The parameters of the regression models (e.g., the bias and slope of the linear regression) will comprise the features of a FNN; and (2) Use a recurrent neural network (RNN) or long short-term memory (LSTM) network to model the temporal evolution of EDA and HR values in association to the state anxiety label. The output of the RNN/LSTM should be the state anxiety label. Experiment with a 5-fold cross-validation framework for both approaches using the average absolute error computed over all folds. Please report and discuss your findings.

(e) (2 points) Interpreting the model decisions. Based on the results that you obtained in questions (b) and (d), use an interpretable machine learning algorithm of your choice to better understand which part(s) of the data contributed to the algorithms decision. Discuss your findings using a few examples.
Note: You can use existing toolboxes from Github, such as LIME (`https://github.com/marcotcr/lime`) and LIME for time (`https://github.com/emanuel-metzenthin/Lime-For-Time`).

(f) (2 points) Examining individual differences. Examine individual differences with respect to the bio-behavioral features, state anxiety labels, and the models' decisions between native and non-native English speakers (1: native, 2: non-native in "data.csv"). What might be potential sources of these differences (if any)? Consider building group-specific models for estimating state anxiety for each group separately. Please discuss your results in comparison to the models obtained from the previous questions.

(g) (2 points) E-poster. Create an e-poster presentation of your work. The e-poster will give the main gist of your work, including the problem statement, your methodology, and the main results from your experiments. **Add visuals to your poster so that people understand the main concepts.** Do not make your e-poster too crowded, since you want other people to be able to see through the screen projection. You can find here the link to prepare your poster presentation `https://www.youtube.com/watch?v=1RwJbhkCA58&feature=youtu.be`.
Note: Each team will project their e-poster on a screen. Half of the teams will be presenting in the first half of the class, while the other teams will go around the posters and discuss their classmates' solutions. Roles will switch during the other half of the class.

(h) (1 point) Reporting other teams' work. When your team is not presenting, your will go around the posters of the teams that are presenting and report their main findings. In the final report, provide a brief description of the work and main results from **4 other teams**.
Note: You can distribute the work among your team members.