

## DS203 Programming for Data Science

### Assignment 7 – Basic Machine Learning

#### Instructions:

1. Use python and libraries such as numpy, matplotlib, and pandas.
2. Submit your assignment, preferably a single .ipynb file, by October 16, 2022, 11:59pm on Moodle at: <https://moodle.iitb.ac.in/mod/assign/view.php?id=95560>
3. In your assignment, show not only the final operations that worked, but also the results of one or two other things that you tried that did not work.
4. Include copious comments about your thoughts in choosing the operations for a particular problem, and your observations about what worked and what did not work, including possible reasons.
5. Comment every line of code to demonstrate understanding of how the code works. Write as much code as you can on your own, but you can use functions from the aforementioned libraries.
6. Cite sources referred at the place where they were used, and include a reference section at the end.

#### Problems:

1. Categorize the following problems into classification, regression, clustering or dimension reduction, while also noting down the input and output variables:
  - a. Summarize the grades of all 50 courses that a student using 5 numbers assuming that the performance on some of the courses might be correlated. [.5]
  - b. Divide students into unspecified personality groups based student based on their performance in various courses and extra-curricular activities. [.5]
  - c. Predict the salary of a student after two years of graduating based on his/her performance in various courses, extra-curricular activities, and their first job type. [.5]
  - d. Predict the best type of job for a student based on his/her performance in various courses and extra-curricular activities. [.5]
2. Predict the rating of clothing items based on other variables for the data at URL <https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish> or <https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand> using the following general guidelines.
  - a. Decide whether the variable to be predicted is discrete or continuous. Also, decide if this is a supervised or an unsupervised problem. For the former, find the target variable. [1]
  - b. Decide on a measure of performance, e.g. accuracy, area under ROC curve, F1-score, sensitivity, specificity, variance explained, Davies-Bouldin criteria, reconstruction error, root mean square error (RMSE), RMSE normalized by standard deviation of the target variable, mean absolute error etc. [1]
  - c. Decide which variables might be relevant for predicting the target variable. [1]
  - d. Decide which variables are usable. [1]
  - e. Convert categorical variables into one-hot bit dummy variables, and standardize (or normalize) the continuous variables. Also, figure out a way to deal with a few missing variables. [1]
  - f. If there are too many variables and they are correlated, then consider eliminating some of the variables whose information is adequately covered by other variables. [1]
  - g. Consider ML frameworks that work with the type of problem and the input variables, and try to order them based on whether they are likely to succeed based on the number of variables and the number of samples. For example, some ML frameworks fare better with fewer samples and higher dimensions (e.g. LASSO regression), while others are scalable with more samples (e.g. neural networks, RF, and kernelized SVM). Pick two-three ML frameworks. [1]

- h. Divide the data into training, validation, and testing subsets, e.g., in 70:15:15 ratio. [1]
  - i. List hyper-parameters for the ML frameworks selected, and form hyper-parameter grids. [1]
  - j. Normalize the data, and use the normalization parameters of the training data for validation and testing as well, instead of recomputing the parameters. [1]
  - k. Train the ML frameworks with the hyperparameters set as per the grid search, and test their performance on validation subsets. [1]
  - l. Select the ML framework and hyperparameter combinations with the best validation performance, and test them on the test data. [1]
  - m. Comment whether the test results indicate if the model is usable or not. [1]
3. Repeat the exercise for predicting gestures based on muscle activity for the following dataset: <https://www.kaggle.com/kyr7plus/emg-4> or <https://archive.ics.uci.edu/ml/datasets/Credit+Approval>. This time **replace step (h) with five-fold cross-validation**. All steps in this part carry [0.5] points, except part (h)
4. Compress the 64 input dimensions in the same dataset <https://www.kaggle.com/kyr7plus/emg-4> to an appropriate number of dimensions using PCA such that the RMSE reconstruction error is less 1% of the standard deviation of the L2 norm of the 64 variable input. [3] Plot a graph of dimensions retained versus normalized RMSE. [2]