DS 203 : Programming for Data Science
Tutorial and Assignment Sheet – 8
Basic Machine Learning Practice

Submission guidelines:

- Prepare an ipython notebook and name it <roll no.>.ipynb and submit it on Moodle before 11:59pm on Nov 5, 2020.

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.
    b. Divide students into unspecified personality groups based student based on their performance in various courses and extra-curricular activities.
    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.
    d. Predict the best type of job for a student based on his/her performance in various courses and extra-curricular activities.

2. Check out scikit-learn documentation and PRML book by C Bishop, and complete the following table about hyper-parameters and parameters of various ML frameworks:

Problem	Framework	Target output variable type (e.g., one-hot, integer, floating point, or none)	Parameters	Hyper-parameters and their typical value range	Scikit-learn commands for defining, training, and testing
Classification	SVM-C with Gaussian kernel				
Regression	SVM-R with Gaussian kernel				
Classification	NN with one hidden layer				
Regression	NN with one hidden layer				
Classification	Random forest				
Regression	Random forest				
Clustering	k-means				
Clustering	DBSCAN				
Dimension reduction	PCA				
Dimension reduction	Kernel PCA				

3. 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.
    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.
    c. Decide which variables might be relevant for predicting the target variable.
    d. Decide which variables are usable.
    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.
    f. If there are too many variables, then consider dimension reduction techniques such as PCA.
    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.
    h. Divide the data into training, validation, and testing subsets, roughly in 70:15:15 ratio.
    i. List hyper-parameters for the ML frameworks selected, and form hyper-parameter grids.
    j. Train the ML frameworks with the hyperparameters set as per the grid search, and test their performance on validation subsets.
    k. Select the ML framework and hyperparameter combinations with the best validation performance, and test them on the test data.
    l. Comment whether the test results indicate if the model is usable or not.

4. 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

5. 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. Plot a graph of dimensions retained versus normalized RMSE.