CSCI 5408, Summer 2018

Assignment 6: Analyzing data patterns and different classification methods

(Issue: July 18, Due: July 31, 11:59PM)

- TA: Dijana Kosmajac (dijana.kosmajac@dal.ca)
- Tutorial: July 18, 11:05-12:25, Room: 127
- Help Hours: Fri, 11:05-12:25, Room 127; Wed, 18:00-20:00, Room: 127

---

1. **Objectives:**
    1) To learn to use scikit - machine learning tool for data analysis.
    2) To revisit using visualization to explore the data and use those insights in further analysis.
    3) To learn how to identify patterns through simulation of data

2. **Tasks:**
    1) Install Python SciKit (http://scikit-learn.org/stable/) for machine learning and Python Matplot library (https://matplotlib.org/) for visualization.
    2) Retrieve the language dataset from DSL 2014 workshop (https://github.com/Simdiva/DSL-Task). The dataset is targeted for problem of language identification. For the details take a look at the (http://corporavm.uni-koeln.de/vardial/sharedtask.html).
    3) Use PCA (Principal Component Analysis) to plot the training dataset clusters [scatterplot – where every sample represents a dot, and color (or shape) of the dot represents class].
    4) Use PCA or Chi Square statistics for selecting K best features. Find what number of features yields best accuracy.
    5) Use one of the vectorizers provided by sklearn to transform the raw data.
    6) Train the model for: Linear SVM, Logistic Regression, Decision Tree and Naïve Bayes.
    7) Use Pipeline to combine all the steps into one multi-stage model (pipeline per classifier type).
    8) Calculate accuracy for all models and report which model worked best.
    9) Plot confusion matrix for each model and discuss on the plots.
    10) Write a report including the following sections:
        a. Task Description: Present the problem scenario (i.e., the application and the requirements), and the DB (provide references to any external data used within the application).
        b. Feature Extraction and Selection: Provide complete description on the steps performed for extracting data and what methods were used to select useful features.

c. Classification Algorithm: Provide a detailed report on algorithms. Which performed best? Did feature selection impact the performance?
d. Output: Write a comparative analysis between output of classification data relying on accuracy measures and plots.
e. Code Submission: Please provide link to the GitHub repository where all code snippets are uploaded. You can use either public repository or private repository. If using a private repository, please don't create a new one and re-use existing ones that you have created for earlier assignments.
f. Note: You are allowed to take help and guidance from any code source available online (research papers or GitHub repos). However, you are not allowed to copy-paste code into your program. Please provide references for all external sources/libraries including data source and GitHub repositories that you used for performing this assignment.

3. Submit your Ass6 report electronically:

1) Please use Bright Space to submit your assignment.
2) In addition to the report please also provide GitHub repo link where programs are uploaded. Please also include README file in GitHub repository. All teams are required to submit their code through GitHub repositories and no exceptions are granted in this regard.
3) Make sure to cite the dataset you used in your report.
4) Please also provide any additional scripts you used in this assignment (all uploaded to GitHub in a single repository).

*** Plagiarism and Intellectual Honesty:** (http://plagiarism.dal.ca)

Dalhousie University defines "plagiarism as the presentation of the work of another author in such a way as to give one's reader reason to think it to be one's own." Plagiarism is considered a serious academic offense which may lead to loss of credit, suspension or expulsion from the University, or even the revocation of a degree.