Classification II

Learning Outcomes

- Upon successful completion of this lab, you will have demonstrated the abilities to:
  - Building a k-NN model for different data sets
  - Determine/compare the efficiency of models
  - Perform hyperparameter setting and model selection

Instructions:

1. Read the tutorial **here** **(http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial6/tutorial6.html)**
2. Download the following data sets from the UCI Machine Learning Repository:
   - **https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)** **(https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))**
   - **https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+(Version+1)** **(https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+(Version+1))**

Do the following steps on both data sets:

Part I (inference efficiency):

Build a k-NN model and compare its efficiency with another model:

1. Perform preprocessing (normalization) if it is necessary
2. Build k-NN classifier for k = 5:
   A. Use 90% of data set for the **train** and 10 % for the **test**, and perform evaluation 5 times, the final results are the average of trails performance
   B. You should report the final average *F-measure,* and *average test time (*the time that model spends to predict labels for the test dataset instances). Use bar charts.
3. Repeat (2) for building a decision tree classifier (use default parameters).
4. Compare results of part (2) and (3) using appropriate charts

Part II (model selection):

Perform model selection for the k-NN and decision tree:

1. Perform preprocessing (normalization) if it is necessary
2. Build *k-NN* classifier for different **k** (1, 2, 3, 4, 5) and select the best **k:**
   A. Use 90% of data set for **train** and 10 % for the **test**, and 10% of the train for **validation**
   B. Build the k-NN model using the **train** data set and select the best **k** based on *F-measure* on the **validation** set

3. Build the decision tree model using the **train** data set and select the best tree:
    A. Change the tree depth (3, 4...10) and calculate *F-measure* on the **validation** set
    B. Compare results of part (2) and (3) using the appropriate charts

Report:

1. Your report should have a cover letter including the group member names
2. Organize all your *diagrams* and *interpretations* in your lab report *(PDF format)*
3. Include your code and report in a folder (you can zip the folder) and submit it

Resources:

1. **https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)**
2. **https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html)**
3. **https://matplotlib.org/3.1.1/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py (https://matplotlib.org/3.1.1/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py)**