CSE 5243

Instructor: Jason Van Hulse

Homework 4 (Classification)

Due Date: 3/4/2019 at 11:59 pm (Submit the via Carmen). *There will be a 15% penalty for all homework handed in up to 24 hours late; homework will not be accepted more than 24 hours late]*

Objective: In this lab, you will experiment with multiple off-the-shelf classification algorithms on the two datasets that you used in Homework #1 and #2:

- The **Adult** dataset (https://archive.ics.uci.edu/ml/datasets/Adult).
- The **Wine** dataset (https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv)

You can and should use your insights from the previous homework assignments to help complete this assignment. In other words, you have already done preliminary data analysis, determined which features can be eliminated or need to be transformed, what to do with missing values and/or outliers, etc. You are not bound by those choices, however - feel free to adjust as needed.

For the wine dataset, recall that you will need to create a new binary class variable according to the following rule:

quality <= 5 —> class = “Low”, quality > 5 —> class = “High”

After you do this, you should eliminate the *quality* attribute.

Software: You should use either R or Python for this assignment. You may use any packages that you would like within either of those two platforms.

Adult dataset:

- Build **four** different classifiers. This is *not* four of the same classifiers with different parameters (for example, you shouldn’t build kNN with k = 1, 3, 5, and 7 and claim you have built four different classifiers). For each of the classifiers, you should explain the parameter choices that you have made, with some justification for these choices. For example, if you use kNN, explain your choice of *k* and the *proximity measure*. In addition, you should discuss any data transformations or other preprocessing you may have needed to perform (e.g., for kNN, did you normalize or standardize the variables? How did you treat categorical attributes?) You do not need to perform automated hyperparameter tuning for this exercise; simply make reasonable choices for the modeling parameters given the data you are working with and your knowledge of how these algorithms work.

- Choose an *evaluation approach* (e.g., single partition, cross validation, …) that would allow you to choose among these different models and, once that selection is done, to estimate

the generalization performance. Be thoughtful in your choice given that you will need to pick a different method for the Wine dataset (see below).

- For each of the four models that you built, report the following performance statistics on the *validation data*: confusion matrix (default threshold) with the accuracy, True Positive and False Positive rates, Precision, and F-measure. In addition, plot the ROC curve and compute the area under the ROC curve. From this data, determine which model you choose as the final model, and provide justification.

- For the final chosen model, calculate the *generalization performance* using the same measures listed above (accuracy, True Positive and False Positive rates, Precision, and F-measure. In addition, plot the ROC curve and compute the area under the ROC curve).

Wine dataset: Perform the same steps as outlined above for the Adult dataset, with the following modifications:

- Use at least one different classifier.

- Choose a different evaluation approach (e.g., if you created a single Train/Validation/Test split for the Adult dataset, you may use 10-fold Cross validation for the Wine dataset).

What you need to turn in:

1) Written report, either as a notebook or as a typed document.
2) Code, either in Python or R.

Grading rubric:

1. **Code** (30%): The TA should be able to run your code from beginning to end to *read in the data from the URLs (not the local file system)*, perform data preprocessing, and execute both model training and evaluation (including the output of all of the model performance measures). There should be no errors in your code. If you hand in only a notebook, the TA should still be able to run all of the cells, from beginning to end, to produce the desired output.
2. **Readability of your report** (20%): Is it well organized and does the presentation flow in a logical manner; are there many grammar and spelling mistakes; do the charts/graphs relate to the text, etc…
3. **Model Development & Evaluation** (50%): Did you develop the models that were outlined and discuss their relative performance? Did you justify your choices and discuss different options? Were your choices reasonable and justified by the data? Pay close attention in particular to how you compared the models and how you estimated the generalization error. The report should at a minimum address each of the bullet points outlined above. The first section should cover the Adult dataset, while the second section should cover the Wine classification dataset.

Collaboration: For this assignment, you should work as an individual. You may informally discuss ideas with classmates, but your work should be your own.

How to hand in your work:

Please do this work in either Python or R. All the related files (code and/or report) except for the data will be tarred in a *.zip file or *.tgz file, and submitted via Carmen. Use this naming convention: “Project4_Surnames_DotNumber.zip" or “Project4_Surnames_DotNumber.tgz.” The submitted file should be less than 5MB.