3/5/2018

CSE 5243

Instructor: Jason Van Hulse

Homework 4 (Classification)

Due Date: 3/21/2018 5:30 pm. Hand in the report in class and submit the via Carmen.

In this lab, you will experiment with multiple off-the-shelf classification algorithms on the **Adult** dataset (https://archive.ics.uci.edu/ml/datasets/Adult). **This is an individual assignment - you are not permitted to work in a team.**

You should hand in a written report, maximum of 9 pages. You do not need to turn in any code. You can use any software (or combination of software) for this assignment that you would like. Please consider all aspects of the knowledge discovery cycle in this assignment, including preliminary data exploration, data processing/transformation, model building and model evaluation.

Your report should have (at minimum) the following sections:

1) **Preliminary data analysis** - discuss any observed trends with the data, include interesting graphs (histograms, scatterplots), summary statistics, correlations among features, etc. I would suggest including some of the interesting output of this preliminary data analysis to help augment the presentation.
2) **Data Transformations** - What if any transformation or processing of the data is needed (prior to actually building the models) e.g., treatment of outliers and missing values, discretizing, feature subset selection? Even if you decide not to handle these issues, please discuss why.
3) **Model development**- Test each of the following modeling approaches in this section using off the shelf software (you are not expected to code a Decision Tree from scratch). For each approach, discuss the parameters you experimented with and why you ended up choosing what you finally chose.
   - **A.** Decision Tree
   - **B.** Artificial Neural Network
   - **C.** Support Vector Machine
   - **D.** Ensemble learner (such as Adaboost, RandomForest, etc)
   - **E.** One technique of your choosing (e.g., kNN, Logistic Regression)
4) **Model evaluation** - Describe your approach to evaluating and comparing the performance of the different models that you have built in part 3 (e.g., 10-fold cross validation). For each model include performance statistics (such as accuracy, F-measure, ROC Curve, etc). The conclusion should be to select your preferred modeling approach - please justify this choice, and state any pros and cons of this choice.

Grading rubric:

1. **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…
2. **Exploratory data analysis and data preprocessing** (30%): Briefly describe your observations regarding the data. Did you need to perform any types of data preprocessing? Did you clean any data (removing noise, outliers) or need to handle any missing values? Did you perform any type of feature transformation or selection?
3. **Model Development & Evaluation** (50%): Did you develop the 5 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?

If you simply take the data and immediately build models, with no explanation of your EDA or data processing tasks, your score on this assignment will be quite low. Poor quality writing or a disorganized report will also reduce your score.

What you need to turn in:

You should hand in a written report, and submit to Carmen, with a maximum of 9 pages in length. *You do not need to turn in any code or data for this assignment.*