CMPT459 Spring 2019
Data Mining
Martin Ester
TA: Ricardo Silva Carvalho

Programming Assignment 4

Total marks: 100
Due date: March 27, 2019

Data

The "Pairs FIFA 19" dataset contains detailed information for soccer players scraped from the website "sofifa.com". Each record is a pair of different players, and the attributes provide basic information about both players and their soccer skills.

The dataset contains 142 attributes, where the first is the pair ID, the following 70 attributes refer to the attributes of the first player, the next 70 to the attributes of the second player, and the last attribute (named "Chemistry") records if the pair of players have "Good" or "Bad" chemistry – in other words, if they play well together or not.

Here are the original data overview and attribute descriptions:

- https://www.kaggle.com/karangadiya/fifa19

and here is a better view of the information:

- https://sofifa.com/

Our goal is to classify each pair of players as having "Good" or "Bad" chemistry, based on the attributes of the two players. The data was split into train (60%), validation (20%) and test (20%) datasets. **The test dataset does not have "Chemistry" values.**

The train, validation and test datasets can be downloaded from the following links on CourSys:

https://coursys.sfu.ca/2019sp-cmpt-459-d1/pages/fifa-pairs-train
https://coursys.sfu.ca/2019sp-cmpt-459-d1/pages/fifa-pairs-valid
https://coursys.sfu.ca/2019sp-cmpt-459-d1/pages/fifa-pairs-test

Tasks

The guidelines that must be followed for this assignment are:

- Use either R (version 3.5.2) or Python 3 (version 3.6).
- Use either R Notebook or Jupyter Notebook.
- The assignment submission consists of two files:
    1) The HTML report exported from R Notebook or Jupyter Notebook.
        - Every notebook code cell must have its code displayed on the HTML report, along with the results.
        - After finishing and exporting the notebook to HTML, please make sure every cell is as expected, showing the code and result of your work.

2) A text file containing “Chemistry” prediction for the test dataset.
    - See [Task 2] for details on the format of the file.

Tasks are defined as follows:

[Task 1]: (70 marks) Considering that there is no need of interpreting the models, focus on using the dataset for prediction, **trying to avoid overfitting**:

- Feel free to select features and do any feature engineering desired, but please make sure to explain the reasons behind each transformation.
- Use the train dataset to train at least two models for predicting “Chemistry”.
- Use the validation dataset for tuning.
- Briefly describe your approach to **avoid overfitting.**
- Report the confusion matrix and accuracy for your **best model** on the **validation** dataset.

[Task 2]: (30 marks) Create a text file with the predictions of your best model on the **test dataset**. The file should contain “Pair.ID, Chemistry” as the first row, and “<number>, <predicted label>” on each subsequent row. See example below (without the quotes):

“
Pair.ID, Chemistry
1, Good
2, Bad
3, Bad
4, Good
“

The example above shows that:

- The file should not contain any quotes or double quotes.
- The first row is used for the feature names: “Pair.ID, Chemistry” (without quotes) separated by comma.
- Each row contains two values separated by a comma.

Marking will take into account two things:

- Task 1: The quality of your approach to training, in particular to the avoidance of overfitting.
- Task 2: The accuracy of your predictions on the test dataset.

Please note that text files with wrong format will receive 0 marks on [Task 2].

Submission

Two files need to be submitted:

- The HTML exported from your notebook, named: “PA4_<your_student_number>.html”
- The resulting text file for [Task 2], named: “PA4_<your_student_number>.txt”

After exporting the notebook to HTML, please **make sure that the content of every cell is as expected**, showing the code and the result of your work.

Please zip these two files and submit the resulting “.zip” file named “PA4_<your_student_number>.zip” in CourSys.