Datamining & Statistics

Assignment 2: Classification

To get a 6 (complete chapter 1 & 2 in the Jupyter Notebook Template)

Find a data set of your own choice (different from the ones used in class). Use the different techniques used in class to make a prediction of a categorical variable.

Write your steps and findings in the Jupyter notebook. It should be possible for a student next year if given your notebook, that he/she can follow and understand what you've done. This includes therefor a brief explanation of a term, argumentation of some choices you make, how well the model performs (using some of the metrics), and what the results mean.

The file should be able to run from start to finish. Explain in the beginning which libraries might have to be installed first.

To get higher than a 6:

Add to the 'to get a 6', one or more of the following:

- + 1 Visualisation (chapter 3 in the template)
  If there are two features, you can visualise the decision areas of the different techniques. Add a function that will visualize the decision areas, and has as input parameters the model and the data to be used.

- + 1 Categorical features (chapter 4 in the template)
  The first model discussed was the categorical NB. That was actually the only model shown that uses categorical features. Create some other models using the categorical data. Use at least once the one-hot-encoding. You can do this with either your own dataset or the UFC data.

- + 1 Performance (chapter 5 in the template)
  The provided jupyter notebook has already a lot of different models to predict the winner of a UFC fight. The best performance score was the Quadratic Discriminant Analysis, with an accuracy score of 0.6747. It should be possible to improve on this by tuning one or more parameters of one or more of the models, or adding more features (or a combination of the two). Explore and show your result.

- +1 Something else (chapter 6 in the template)
  Add a machine learning method that is not shown in the provided jupyter notebook, with a thorough explanation. You can do this with either your own dataset or the UFC data.

A bonus point (if possible)
The student with the highest performance score will get an additional bonus. The performance score will be based on a different testing set than the one used in the provided jupyter notebook.

Deliverables

To submit:

- Jupyter notebook file
- Dataset file(s) (or link to this in case it's too large)

Method to submit:

- One zip or rar file containing the files. Include your student number in the name of the zip/rar file.
- Upload on Moodle.

FAQ's

Q1: Can we work together?
A1: No and yes. No, it is an individual assignment, yes you can of course always help each other. But no direct copy-pasting. Copy-pasting, or copy-pasting and changing just one or two words, either from fellow students, or the internet, books, etc. is considered fraud. This will be reported to the examination board.

But…: But some lines of code will be standard, and only a few parameters or variable names will be different…..
A-but: The copy-pasting is more on the text surrounding the code, and indeed some parts of the code will be copy-pasted. This is acceptable and probably normal in IT that you copy some code from the internet and make a few adjustments to make it work for your data.

Q2: I'm not clear about the '+1', can you explain?
A2: Sure. There are some criteria to get a 6. To get to a 7 add one of the three things described at a '+1', your choice which one. To get to an 8 add two, to get to a 9 add three, and for a 10 add all four.

Q3: Is it possible to get an 11?
A3: No. Indeed 6 + 5 = 11, but the +1 bonus mentions (if possible), so the maximum grade is a 10 and simply not possible to have a higher grade.