Case 1. Heart Disease Classification

Cognitive Systems for Health Technology Applications, Spring 2019

Type

Pair work or individual work, 15-20 hours

Deadline

Sunday 3.2.2019 20:30

Aim

The aims of this assignment are to

- learn to read data from external sources using pandas' read_csv -function
- use keras' neural networks to make an expert system to support in diagnostic decision making
- learn to test model architectures (number of layers, number of units, activation functions), solver optimizers and training settings (epochs, batch sizes, validation splits)
- use matplotlib's visualization tools to make graphical presentations of the training and validation results
- learn to document the results clearly and in easily readable format

Task

Your task is to read and preprocess the data in the folder: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/ and create and train a dense neural network to predict to classify the presence of heart disease.

The file `heart-disease.names` contains information about the attributes and the number of instances in the data files. Some of the files are already cleaned up, for example `processed.cleveland.data`, and some are original raw data (`cleveland.data, hungarian.data, long_beach_va.data`). Start with the processed Cleveland data.

Use Jupyter Notebook to solve the problem. Use the Markdown cells to wrap a document around your code cells.

Try different neural network architectures (number of layers, number of neurons within layer), batch sizes and number of epochs, and solver settings and compare them with each other. Try to find the simplest, fastest and smallest possible model and settings that solves the problem most accurately.

The recommended documentation structure is the following:

- Main heading (Title page) including:
  - The assignment name as the title/heading
  - Your name(s)

    - The last date, when you edited the document
    - Organisation name (Helsinki Metropolia University of Applied Sciences)
- Background (Objectives)
    - Why this document is created, and what were the main objectives
- Data
    - Introduction what the data contains and what are its characteristics (like size, attributes, missing values, descriptive statistics, etc.)
- Models and training
    - Description and code for the neural networks model(s)
    - How the training and testing were conducted (epoch, batch sizes, solver settings)
    - How much data was used for testing/validation purposes
- Results
    - Main results shown both graphically and explained textually
- Conclusions
    - What were the main observations and how well the objectives were achieved

Return

Save your results to your GitHub folder and provide a link to your Notebook in OMA.
Use OMA's hyperlink tool for providing the link.

Evaluation

The following categories are used for evaluation:

- Organisation (5 p)
    - The code is sequential and the code cells (parts of scripts) are in right order
    - The document follows a clear structure
- Clarity (5 p)
    - The document (and embedded code) is clear, polished, and easy to understand
    - The code follows good coding practices and contains sufficient comments
    - The document supports the code
- Contents (5 p)
    - The background and data preprocessing are well explained
    - The model is validated and tested
    - The results are reasonable
    - The conclusions are clearly stated and in a line with the results

max. 15 points. Late submission reduces the maximum achievable points.

Materials

- Reading data from a csv file
- Panda's tutorials
- Guide to the sequential model
- Preprocessing data
- Cross-validation: evaluating estimator performance