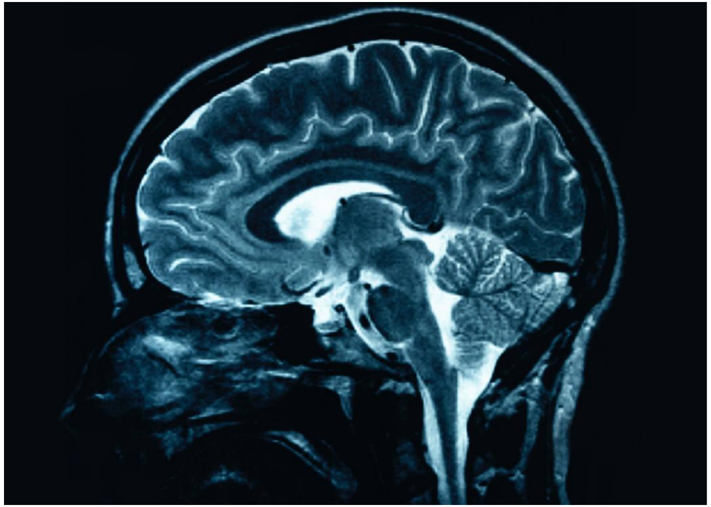
Practical exercises

Advanced Machine Learning

Task 1: Predict a person’s age from brain image data



MRI image

MRI

- Magnetic Resonance Imaging (MRI) is a key technology in medical imaging

- Non-invasive + Non-radiative investigation of sensitive organs (brain)

- Switzerland: 45.0 MRI units per 1.000.000 inhabitants (2016)
  West-Africa: 0.22 MRI units per 1.000.000 inhabitants (2018)

https://data.oecd.org/healtheqt/magnetic-resonance-imaging-mri-units.htm
http://www.panafrican-med-journal.com/content/article/30/240/full/

MRI processing

Raw brain scans are difficult to handle

- 3D brain scans are ~ 200x200x200 ~ 10^7 features/voxels (3D pixels ~ $1mm^3$)

- 3D structure + individual brain shapes → difficult to recognize disease patterns

MRI feature extraction

We are using ~200 anatomical features for this project

- Informative features derived from image data (with Freesurfer)
- No need to process big images (6 GB) → csv sheet (3 MB)
- No need for image analysis (feature extraction)

- Information loss

Task 1: Age prediction

- We have modified the derived input data in three ways:

1. Irrelevant features

2. Outliers

3. Perturbations (e.g. missing values, etc.)

Description about the dataset

File description

We provide the following 3 files:

- train.csv: the training set, including the features and labels
- test.csv: the test set (make predictions based on this file)
- sample.csv: a sample submission file in the correct format

Task description

Subtask 0: Filling missing values

Background

There are some missing values in the data, originally they are set to NaN values. Most of the methods can not handle then automatically. There are different strategies how to impute them: mean, median, most frequent etc.

Task requirement

We require that students
fill missing values in the training and the test set.

Subtask 1: Outlier Detection

Background

In the training set, there are some outliers. If the resulting model is not robust enough, it may be sensitive to the outliers. In this case, outliers deletion can be expected to lead to better results.

Task requirement

We require that students build an outlier detection model to make classification for samples in the training set i.e. whether they are outliers.

Subtask 2: Feature selection

Background

To make the task a bit more challenging, we add some manual features to the FreeSurfer processed dataset.

Feature selection approach is thus used for the following reasons:

- Simplification of models to make them easier to interpret
- Shorter training times and avoiding the curse of dimensionality
- Enhanced generalization by reducing overfitting

Task requirement

We require that students

use feature selection methods to label the features as selected features and unselected features.

Here, unselected features includes irrelevant features and redundant features.

Main task: Age Prediction

Background

After primary preprocessing and dimensionality reduction, now we finally arrive at the regression task. Tuning hype-parameter and parameter to achieve better regression performance.

Task requirement

We require that students
use suitable regression methods to predict the age of a person from his/her brain data.

Evaluation metric

Coefficient of Determination *R^2*

is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). The best score is 1 and it can be negative. Prediction of always the same constant, equals to the expected value of y, will give score of 0.

How to compute it in Python:

```
from sklearn.metrics import r2_score

score = r2_score(y_true, y_pred)
```

Q&A