Neural Networks for Health Technology Applications

29.1.2020, Class exercise 3 - Learn to preprocess the data

The aim of this exercise is to learn to preprocess the data using scikit-learn.preprocessing package and pandas DataFrame object methods to handle missing data. Check the lecture notes for more details.

1. Create a new empty Jupyter Notebook and for the first code cell write the following codes:

```
%pylab inline
import pandas as pd
from sklearn import preprocessing
```

```
Populating the interactive namespace from numpy and matplotlib
```

The last code line reads all the preprocessing methods from scikit-learn package.

2. Read in the same dataset as in previous class exercises, but now don't clean the missing values from the data.

```
filename = r'https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data'
df = pd.read_csv(filename,
                 index_col = None,
                 header = None,
                 na_values = '?')
df.tail()
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
298	45.0	1.0	1.0	110.0	264.0	0.0	0.0	132.0	0.0	1.2	2.0	0.0	7.0	1
299	68.0	1.0	4.0	144.0	193.0	1.0	0.0	141.0	0.0	3.4	2.0	2.0	7.0	2
300	57.0	1.0	4.0	130.0	131.0	0.0	0.0	115.0	1.0	1.2	2.0	1.0	7.0	3
301	57.0	0.0	2.0	130.0	236.0	0.0	2.0	174.0	0.0	0.0	2.0	1.0	3.0	1
302	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	1.0	NaN	3.0	0

3. Separate the data and labels.

```
data = df.loc[:, 0:12]
labels = df.loc[:, 13]
```

4. Build a simple neural network and try train the network with this data. You might also want to simplify the labels into two classes (healthy and disease) and try to use the binary classifier to predict the outcome. Notice, that the network can't handle the missing values.

5. Replace all missing values with zeros and check that the simple neural network works now. Make notes what the loss function and accuracy values were.

6. Preprocess the data using using scikit-learn's StandardScaler before training. Make notes how the loss function and accuracy changed when you preprocessed the data.

7. Now try different approaches for imputing the missing values and scale the data. Keep the neural network always the same and use similar settings for the training. Make notes how the accuracy and loss function are changing when you change the approach (strategy for preprocessing the data). Write a summary, which strategy worked best on this case.

Remember to save your exercise!