# Programming

## Winter 2020/2021

**Number 07, Submission Deadline: Jan. 12, 2021**

### 1. The diabetes data set (7P)

Scikit-Learn provides the following diabetes data set that has been published by

Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statistics (with discussion), 407-499.

The authors describe the data set as follows:

> Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

1. Inform yourself about Lasso regression. Briefly describe its key properties (2P)
2. Perform a regression analysis of the diabetes data set with Lasso. Split the dataset in a train (70%) and a test set (30%). (1P)
3. Describe and visualize your results (2P)
4. Compare your results with the test-dataset. (2P)

[22]:
```
from sklearn.datasets import load_diabetes
import pandas as pd

diabetes = load_diabetes()
data =pd.DataFrame(diabetes.data, columns=diabetes.feature_names)

data.head()
```

[22]:

| | age | sex | bmi | bp | s1 | s2 | s3 | \ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.038076 | 0.050680 | 0.061696 | 0.021872 | -0.044223 | -0.034821 | -0.043401 | |
| 1 | -0.001882 | -0.044642 | -0.051474 | -0.026328 | -0.008449 | -0.019163 | 0.074412 | |
| 2 | 0.085299 | 0.050680 | 0.044451 | -0.005671 | -0.045599 | -0.034194 | -0.032356 | |
| 3 | -0.089063 | -0.044642 | -0.011595 | -0.036656 | 0.012191 | 0.024991 | -0.036038 | |

```
4  0.005383 -0.044642 -0.036385  0.021872  0.003935  0.015596  0.008142

         s4        s5        s6
0 -0.002592  0.019908 -0.017646
1 -0.039493 -0.068330 -0.092204
2 -0.002592  0.002864 -0.025930
3  0.034309  0.022692 -0.009362
4 -0.002592 -0.031991 -0.046641
```

### 2. The breast cancer Wisconsin diagnostic data set (8P)

Another data set that Scikit-Learn provides is the *breast cancer Wisconsin diagnostic data set* that was first published by

W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

The data set comprises data of 569 patients and consists of features that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

[31]:
```
from sklearn.datasets import load_breast_cancer

breast_cancer = load_breast_cancer()
data =pd.DataFrame(breast_cancer.data, columns=breast_cancer.
 ↪feature_names)

data.head()
```

[31]:
```
   mean radius  mean texture  mean perimeter  mean area  mean smoothness ␣
 ↪\
0        17.99         10.38          122.80     1001.0          0.11840
1        20.57         17.77          132.90     1326.0          0.08474
2        19.69         21.25          130.00     1203.0          0.10960
3        11.42         20.38           77.58      386.1          0.14250
4        20.29         14.34          135.10     1297.0          0.10030

   mean compactness  mean concavity  mean concave points  mean symmetry  \
0           0.27760          0.3001              0.14710         0.2419
1           0.07864          0.0869              0.07017         0.1812
```

```
2           0.15990          0.1974              0.12790         0.2069
3           0.28390          0.2414              0.10520         0.2597
4           0.13280          0.1980              0.10430         0.1809

   mean fractal dimension  ...  worst radius  worst texture  worst␣
↪perimeter  \
0                 0.07871  ...         25.38          17.33           184.
↪60
1                 0.05667  ...         24.99          23.41           158.
↪80
2                 0.05999  ...         23.57          25.53           152.
↪50
3                 0.09744  ...         14.91          26.50            98.
↪87
4                 0.05883  ...         22.54          16.67           152.
↪20

   worst area  worst smoothness  worst compactness  worst concavity  \
0      2019.0            0.1622             0.6656           0.7119
1      1956.0            0.1238             0.1866           0.2416
2      1709.0            0.1444             0.4245           0.4504
3       567.7            0.2098             0.8663           0.6869
4      1575.0            0.1374             0.2050           0.4000

   worst concave points  worst symmetry  worst fractal dimension
0                0.2654          0.4601                  0.11890
1                0.1860          0.2750                  0.08902
2                0.2430          0.3613                  0.08758
3                0.2575          0.6638                  0.17300
4                0.1625          0.2364                  0.07678

[5 rows x 30 columns]
```

1. Inform yourself about decision tree classification with Scikit-Learn. Briefly describe the classification algorithms that Scikit-Learn provides (2P)
2. Perform a classification analysis of the breast cancer data set. In doing so,
    1. Use cross validation in your analysis; justify your choice(s) of the number of partitions (1P)
    2. Run the analysis for all decision tree algorithms that Scikit-Learn provides (2P)
    3. Evaluate the classification quality of the algorithms with your (justified) choice

of metric (2P)

4. Visualize your results. (1P)