Advanced Machine Learning 2021 **Project 2**

1. The aim of the project is to compare different feature selection methods.
2. The goal is to propose methods of feature selection and classification, which allow to build a model with large predictive power using small number of features.
3. Datasets:
   - Dataset *artificial* is an artificial dataset in which some relevant features are hidden among large number of irrelevant ones (files: artificial_train.data, artificial_train.labels, artificial_valid.data).
   - Dataset *digits* corresponds to the problem of recognizing digits ("4" and "9") in which significant features are also hidden among large number of irrelevant ones (files: digits_train.data, digits_train.labels, digits_valid.data).
4. There are 3 files for each dataset: training data, labels for training data and validation data. Table 1 contains basic information about the datasets.

| Data | Features | Observations (training data) | Observations (validation data) |
|---|---|---|---|
| artificial | 500 | 2000 | 600 |
| digits | 5000 | 6000 | 1000 |

Tabela 1: Basic characteristics of the datasets.

5. Training data is used to train the model and select relevant features. The goal is to make a prediction for observations belonging to validation data. Each observation in validation data should be assigned posterior probability (for class '1'), i.e. $P(y = 1|x_1, \ldots, x_p)$.
6. Save the results to the files:
   - *CODE_artificial_prediction.txt*, posterior probabilities for validation data, for dataset *artificial*.
   - *CODE_artificial_features.txt*, selected features for dataset *artificial*.
   - *CODE_digits_prediction.txt*, posterior probabilities for validation data, for dataset *digits*.
   - *CODE_digits_features.txt*, selected features for dataset *digits*.

   CODE denotes the code of the student (first student from the group): 3 first letters of the first name + 3 first letter of the second name. In the first line of the file you should place the code of the student and in the following lines the probabilities or indices of selected features. Example files with the results: `JANKOW_artificial_prediction.txt` and `JANKOW_artificial_features.txt` for student 'JAN KOWALSKI'.
7. Datasets and example files can be found at: `https://pages.mini.pw.edu.pl/~teisseyrep/TEACHING/AML/`.
8. Projects are prepared in groups of two students.
9. It is necessary to test at least 4 feature selection methods.
10. Final grade will be based on :

- Predictive performance of the model, we will use balanced accuracy ($BA = \frac{1}{2}(\frac{TP}{P} + \frac{FN}{N})$) and the number of features in the model. Ranking will be based on balanced accuracy (the higher the value the better). In the case of non-significant differences in balanced accuracy, the number of features will decide (the smaller the better). (50 %).
- Presentations (5 minutes) summarizing the results (25 %),
- Reports (max 3 pages A4) containing the description of the methods and results of experiments (25 %).

11. Deadlines and presentations:
    - Group 1: 30 May (deadline), 2 June (presentations)
    - Group 2: 6 June (deadline), 9 June (presentations)
12. Please send the results to teisseyrep(at)ipipan.waw.pl.