Advanced Machine Learning 2022 **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 *spam* corresponds to the problem of recognizing if sms is spam message or regular (files: spam_train.csv, spam_test.data). In order to encode this data in tabular form, use document term matrix representation (create list of words found in training data and define features where x_j representing the number of occurences of j-th word in text).
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
spam	7879	4572	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 of the model 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_spam_prediction.txt*, posterior probabilities for validation data, for dataset *spam*.
   - *CODE_spam_features.txt*, selected features for dataset *spam*.

   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 MS Teams.
8. Projects are prepared in groups of two students.
9. It is necessary to test at least 4 feature selection methods.

10. Components of the final grade.

    - Predictive performance of the model will be assessed based on predictive power of the model (larger the better) and number of ORIGINAL features used by the model (smaller the better) for artificial dataset or just number of features (corresponding to words) for spam dataset. To evaluate the predictive power we use Balanced Accuracy measure
    $$BA = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}).$$
    We denote by m the number of features used in the model. A final score is defined below.

      – For Artificial dataset:
      $$Score(BA, m) = BA - 0.01(\frac{1}{5}m - 1)_+.$$
      Examples:

      (a) For $BA = 90\%$ and $m = 5$, the score is 90%.

      (b) For $BA = 90\%$ and $m = 20$, the score is $90\% - 3\% = 87\%$

      – For Spam dataset:
      $$Score(BA, m) = BA - 0.01(\frac{1}{100}m - 1)_+.$$

    Total points for the Score: (50 %).

    - Presentations summarizing the results (25 %). Each group has to provide a recording with project presentation (up to 5 minutes). Recording is obligatory, but additionally half of groups which didn't present the 1st stationary will be asked to present stationary the 2nd project. Dates of presentations will be known befor 1st project presentations. Stationary presentations are supposed to last around (not up to) 5 minutes.

    - Reports (max 4 pages A4) containing the description of the methods and results of experiments (25 %).

11. The project is for 20 points.

12. Please save all results to the ZIP file, named CODE.zip (where CODE is a code of the group). The archive should contain the following folders:

    (a) code (put all source codes in this folder)

    (b) report (put the report in pdf file in this folder)

    (c) results (put the results according to the instructions in item 6)

    (d) presentation (pdf or pptx)

    (e) presentation resording

13. Deadlines and presentations:

    - Deadline for BOTH groups: 1 June 2023.
    - Presentations in Group 1: 5 June 2023.
    - Presentations in Group 2: 12 June 2023.

14. Please send the results to norbertryciak(at]gmail.com.