+FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

OF HIGHER EDUCATION

ITMO UNIVERSITY

**Report**

**on the practical task No. 1**

**“Supervised learning algorithms”**

Performed by

Chernobrovkin T.V. (412642)

Academic group J4133c

Accepted by

Gladilin P.E.

St. Petersburg

2023

# Goal

Изучить supervised learning algorithms на примере решающих деревьев и случайного леса.

# Problem

1. Train 4 different classifiers using sklearn library to predict "Activity" (biological response of the molecule) field from the "bioresponse.csv" dataset:

- small decision tree;

- deep decision tree;

- random forest on small trees;

- random forest on deep trees;

Refer to ‘Random\_Forrest.ipynb’ and ‘Decision\_Trees.ipynb’ notebooks for examples. Split the data to train and test as 75%/25%.

2. Calculate the following metrics to check the quality of your models:

- precision;

- recall;

- accuracy;

- F1-score;

- log-loss;

3. Plot precision-recall and ROC curves for your models.

4. Train a classifier who avoids Type II (False Negative) errors and calculate metrics from p.2 for it. Recall for it should be not less than 0.95.

# Theory

Supervised learning) - one of the methods of machine learning, during which a test system is forced to learn by means of stimulus-response examples. In terms of cybernetics, it is a type of cybernetic experiment. There may be some dependence between inputs and reference outputs (stimulus-response), but it is unknown. Only a finite set of precedents - stimulus-response pairs, called a training sample - is known. On the basis of this data, we need to restore the dependence (build a model of stimulus-response relations suitable for prediction), i.e., build an algorithm capable of producing a sufficiently accurate response for any object. To measure the accuracy of responses, as well as in example-based learning, a quality functional can be introduced.

A classification task is a task in which there is a set of objects (situations) divided into classes in some way. There is a finite set of objects for which it is known to which classes they belong. This set is called a sample. The class belonging of the remaining objects is unknown. We need to construct an algorithm capable of classifying (see below) an arbitrary object from the initial set.

Cross-validation is a procedure of empirical evaluation of generalization ability of algorithms. Cross-validation is used to emulate the presence of a test sample that does not participate in training, but for which the correct answers are known.

A decision tree is a decision support tool used in machine learning, data analysis and statistics. The structure of the tree is represented by "leaves" and "branches". The edges ("branches") of the decision tree contain the attributes on which the target function depends, the "leaves" contain the values of the target function, and the remaining nodes contain the attributes that distinguish cases. To classify a new case, one has to go down the tree to a leaf and output the corresponding value.

Such decision trees are widely used in data mining. The goal is to create a model that predicts the value of a target variable based on several variables in the input.

Each leaf represents the value of the target variable changed as it travels from the root along the edges of the tree to the leaf. Each internal node is mapped to one of the input variables.

Random forest method is a machine learning algorithm that uses an ensemble of decision trees. The algorithm is applied to classification, regression and clustering tasks. The main idea is to use a large ensemble of decision trees, each of which by itself gives a very low classification quality, but due to their large number the result is good.

# Materials and methods

In this task, all calculations were performed on the student's personal laptop. The work was performed in the Python programming language.

# Results

1. Были обучены 4 различных классификатора, используя библиотеку sklearn, для прогнозирования поля «Активность» (биологическая реакция молекулы) из набора данных «bioresponse.csv». Данные для обучения и тестирования были разделены как 75%/25%. Код предоставлен в приложении.
2. Для small decision tree, deep decision tree, random forest on small trees и random forest on deep trees были посчитаны параметры precision, recall, accuracy, F1-score и log-loss. Результат представлен в таблице 1.

Таблица 1 – Результаты обучения моделей

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | small decision tree | deep decision tree | random forest on small trees | random forest on deep trees |
| precision | 0.79087 | 0.75719 | 0.71597 | 0.82718 |
| recall | 0.77757 | 0.78692 | 0.79626 | 0.79626 |
| accuracy | 0.75586 | 0.73454 | 0.70362 | 0.78891 |
| F1-score | 0.78417 | 0.77177 | 0.75398 | 0.81143 |
| log-loss | 8.79957 | 9.56809 | 10.68245 | 7.60836 |

1. Построили кривые precision-recall and ROC для каждой модели (рисунок 1–8).

Изображение выглядит как линия, диаграмма, График, текст

Автоматически созданное описание

Рисунок 1 – Кривая precision-recall для small decision tree

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 2 – Кривая ROC для small decision tree

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 3 – Кривая precision-recall для deep decision tree

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 4 – Кривая ROC для deep decision tree

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 5 – Кривая precision-recall для random forest on small trees

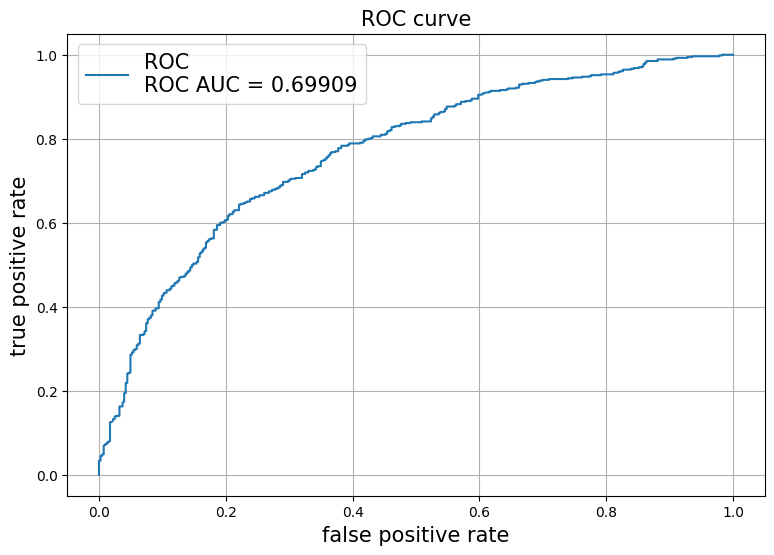


Рисунок 6 – Кривая ROC для random forest on small trees

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 7 – Кривая precision-recall для random forest on deep trees

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 8 – Кривая ROC для random forest on small trees

1. Обучили классификатор, избегающий ошибок типа II (ложноотрицательных), и рассчитать для него метрики из п.2. Recall должен быть не менее 0,95.

Подобрав гиперпараметры n\_estimators = 91, depth = 1, min\_samples\_leaf = 1, min\_samples\_split = 2, bootstrap = False, max\_features = 'log2' получилось достичь значения recall = 0.98505, accuracy = 0.62473, precision = 0.60505, F1-score = 0.74964, log\_loss = 13.52598.

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 9 - Кривая precision-recall для классификатора, избегающего ошибки 2 рода

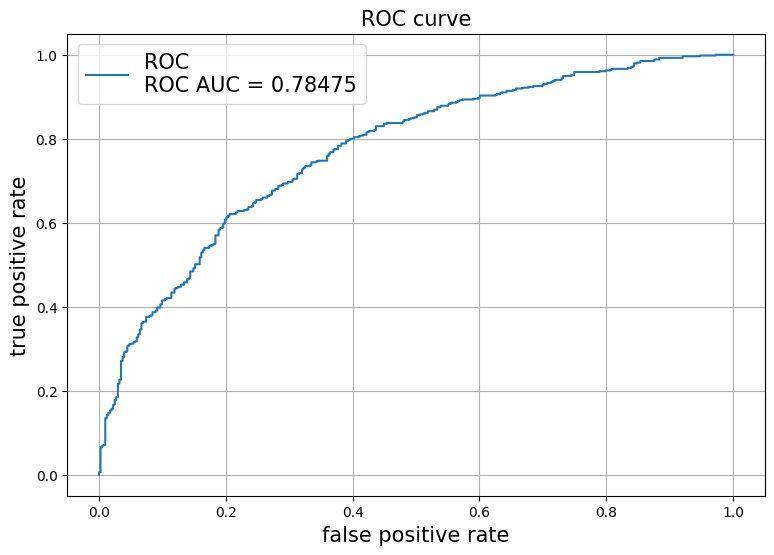


Рисунок 10 - Кривая ROC для классификатора, избегающего ошибки 2 рода

# Conclusion

В ходе выполнения данного задания были изучены аспекты supervised learning algorithms на примере решающих деревьев и случайного леса. Были обучены архитектуры малых и глубоких решающих деревьев и случайных лесов на малых и глубоких решающих деревьев. Для всех архитектур были выведены параметры accuracy, precision, recall, F1-score and log\_loss. При сравнении данных параметров лучшие результаты имеет случайный лес на глубоких деревьях. Для всех моделей были выведены графики кривой precision-recall and ROC.

Также была обучена модель классификатора избегающего ошибки 2 рода. Для такого классификатора значение recall должно быть больше 0.95. Путём подбора гиперпараметров удалось достигнуть значения recall 0.98505. Это значить, что нам удалось обучить классификатор избегающий ошибки 2 рода.

# Appendix

GitHub link: