Seminar 2: Predicting Biodegradability of Chemicals

Intelligent Systems (IS)

November 24, 2022

1 Introduction

Chemicals are all around us. Studying their properties by the means of machine learning is an active research field; matching molecular patterns with their behavior can be a decisive factor in the creation of new materials, drugs, and more.

In this seminar assignment, your task is to explore the data and build machine-learning models that predict the biodegradability of chemicals.

2 Task

You will work with the data set compiled by Mansouri et al. (`https://www.openml.org/search?type=data&status=active&id=1494&sort=runs`). There are 41 features and one target feature (biodegradability). The target variable is encoded as ready biodegradable (1) and not ready biodegradable (2). The data set consists of 1055 instances. Features can be either symbolic or numeric.

IMPORTANT: Use the dataset provided on učilnica and NOT the one posted on the link above. It is minimally modified and split into train in test sets.

2.1 Exploration

Inspect the dataset. How balanced is the target variable? Are there any missing values present? If there are, choose a strategy that takes this into account.

Most of your data is of the numeric type. Can you identify, by adopting exploratory analysis, whether some features are directly related to the target? What about feature pairs? Produce at least three types of visualizations of the feature space and be prepared to argue why these visualizations were useful for your subsequent analysis.

(20%)

2.2 Modeling

Besides the baselines (majority classifier, random classifier), use at least three machine learning algorithms to model the target class. Be ready to argue why did you select specific algorithms and how did you find the best hyperparameters for them. Consider the following points when creating your models:

- Create your models using all features and subsets of them using various feature selection techniques.
- Certain models assume that data follows a particular distribution or may work better with other types of variables (e.g., categorical instead of numeric). Explore whether you can come up with feature transformations that are more appropriate for your models. Try to construct new features from existing ones. Try to explain the results and performance of different models.

(60%)

2.3 Evaluation

Given that the data set is not in the "big data" category, implement a cross-validation procedure based on five folds (approximately equal sized) of your data. Furthermore, repeat the experiment 10 times with different folds and average the results (include standard deviation). You are expected to report the following metrics:

1. F1
2. Precision
3. Recall
4. AUC

Comment on the performance of algorithms and visualize their final scores. How do they perform against the random baseline? What about the constant one? How do different learning scenarios impact the final score? Are the differences between the models statistically significant?

(20%)

3 Report and presentation

The assignment has to be submitted in the form of two files: a markdown file and a PDF file created from the R Studio markdown file (in RStudio → file - new file - R Markdown), where you write both the code, as well as the text of answers (echo = T option must be enabled for each code block). Markdown files can easily be exported to PDF using ("Knit") button in R Studio. If you are using Python, you can produce a similar report with Jupyter Notebook.