Introduction to Artificial Intelligence
Warsaw University of Technology, Summer 2024
Lab 4: Regression and Classification

Filip Szatkowski

Task description

Solve regression or classification task on a given dataset, depending on the variant of the lab:

1. Write a model to predict house price on California housing dataset.
2. Write a model to predict disease progression on diabetes dataset.
3. Write a model to classify an iris flower based on its properties using iris dataset.
4. Write a model to predict if the person survived the Titanic crash using Titanic dataset.
5. Write a model to classify the wine type based on its properties on wine dataset.

Please follow the links for dataset description.

Instructions

1. **Data preparation.** Analyze the features in the dataset and decide which ones are useful for your task. Some features might be redundant, for example, a dataset might contain features that depend on each other, such as "pre-tax price", "tax", and "post-tax price", where the third one is just a sum of the previous two. In such cases, it's better to remove dependant features. If the dataset is not normalized, you should also perform normalization. You can learn more about data preprocessing on sklearn tutorial. Describe your thought process in the report.
2. **Data split.** Split the data into train and test sets. This is already done in the provided template.
3. **Model definition.** Choose two models to fit, depending on your task, eg.: linear regression, logistic regression, SVM classifier or regressor, decision tree, random forest, etc. You can use any models of your choosing, provided they are suitable to the task (regression or classification). Explain your choice of the models in the report.
4. **Model training.** Select an appropriate metric to evaluate your models. Try different model parameters for both models and compare their performance using cross-validation framework with 4 folds. Document the

parameters you tried for both models in separate tables. Compare the performance of both models with the best parameters you have found.

The code for loading the datasets is provided at `https://github.com/fszatkowski/EARIN_Lab4_template`. You can add other files and use any Python packages you like, but please update the `requirements.txt` with added packages.

Submission guidelines

1. The report should include a description of the task and dataset (no more than a few sentences).
2. Do not explain the code in detail in the report, only write about the code if you do something you feel is not standard. The code should be written to be easily readable and should contain the comments in places that are crucial to understanding it.
3. The report should include the thought process behind the data preparation and model choice. There is no clear right choice, so focus on explaining your decisions.
4. The solution must be implemented in Python, based on the provided template.
5. Please ensure that your code adheres to basic standards of coding (PEP8). If possible, use comments and unit tests.

Rules

To pass the lab, it is required to submit both the code and the final report and discuss your solution during the online assessment. The online assessment will take place during the labs and should take around 15 minutes. Please notify me on Teams when you submit the solution to schedule the exact time for the meeting. You should submit the code and a PDF report to the designated Gitlab repository at least a day before the online assessment of the exercise. Programs delivered after the deadline will not be assessed. If you have any questions, please contact me via MS Teams.

Assessment Criteria

You can get $[0, 5]$ points for the lab. The following criteria will be used to evaluate your work:

- Proper implementation of the algorithm: 1 point.

- Report: 2 points.
- Online assessment: 2 points.