Machine Learning – SS 2024 – Exercise Sheet 5

Machine Learning Libraries

Deadline: 13. May 2024, 23:55 (CET)

Important. Upload a single PDF file and a single .ipynb notebook with your homework solution to ILIAS by 13. May 2024, 23:55 (CET). We recommend typesetting your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it wont be graded and you waive your right to dispute that. You only need to submit homework problems. The in-class problems will be discussed in the tutorial but will not be graded.

Policy. Any resources that you use to solve the homework must be cited:

- When taking inspiration from online forums (like StackOverflow), you should include the URL where you found the information;
- If using large language models (like ChatGPT) for formulation/editing your answers, you should either provide the link when available (ChatGPT supports this feature) or detail how you did use it, whether for rephrasing, explaining an unclear concept, or seeking assistance in proving a point, among other uses.
- When taking inspiration from previous years' homeworks, you should include both the number of the exercise and the year of the homework along with how you did use the solution, e.g. "analogous exercise from the 2023 homework 4 ex. 2, differently from that we [...]. Similarly, to the exercise we [...]".
- If you discuss the homework with another team, you should include the name of the team and the members you discussed with – note that this does not mean that the solutions should be the same, just that you discussed the homework together. Write your own solution.

Homework Problems

California Housing dataset

Problem 1. In this assignment, you will apply the concepts and techniques covered in our lecture on machine learning libraries. You will be working with the California Housing dataset, available through sklearn.datasets, to practice data visualization, data preprocessing, model building, and hyperparameter tuning.
You can import the dataset with the following two commands.

```
from sklearn.datasets import fetch_california_housing
dataset = fetch_california_housing()
```

Objectives

Data Visualization Use `matplotlib` and `seaborn` to create visualizations that illustrate the datasets features and target variable distributions. Explore relationships between different features and so on.

Data Preprocessing Utilize `pandas`, `numpy` and transformations from `sklearn` for data cleaning and transformation tasks.

Model Building and Evaluation Implement machine learning models using scikit-learn. Set up pipelines that integrate preprocessing steps with a regression model of your choice (e.g., define pipelines for models we have across the course: linear regression, decision tree, k-NN).

Hyperparameter Tuning Conduct a search for the best hyperparameters for your model using techniques like `GridSearchCV` or `RandomizedSearchCV`. Describe the range of parameters considered and justify your choice of final model based on performance metrics.

Submission

- A Jupyter Notebook containing all code, comments, and visualizations.
- A brief report (from half a page to one page) summarizing your findings, methodologies, and insights from the model evaluation.

Deadline is 13. May 2024, 23:55 (CET). Upload a single PDF and .ipynb to ILIAS. LaTeX is recommended.