INFOB2DA | Practical Assignment 4

Classification methods and model evaluation (Total 100 Points)

Utrecht University | Visualization and Graphics Group
Dr. Michael Behrisch
Alister Machado dos Reis, Elio Verhoef, Kalee Said, Sacha Vermeer, Vincent Haverhoek

Submission deadline:
Sun, 29.10.2023, 23:59.

General information

- You must form groups of 3 students. Individual submissions are only accepted in special cases. Each group member must understand the entire assignment, including the code which you will create!
- Submission deadline: Sunday, 29.10.2023, 23:59pm. You will have to present your submissions on Monday, 30.10.2023 in the regular exercise/werkcollege slot. If you are unable to present your work there, send us an email.
- If you have questions or need more information, you can always use the Questions channel on Teams, or our Office Hours on Friday the 20th of October, and 27th of October, from 09.30 to 10.30.
- You are only allowed to use the Python programming language in your code.
- For this practical assignment, you can achieve 100 points in total. Your overall practical assignments grade will be determined by the sum of your points in PA1-4.

Handing in your assignment

Hand in the Jupyter Notebook, dataset, notebook checkpoints (the folder which is automatically created in your project by Jupyter) and presentation slides as a ZIP-file on MS Teams. Make sure you press the submit/inleveren button after uploading your files in Teams, otherwise the submission is not completed.

Introduction

The aim of this practical assignment is using the classification for answering the following question: "*How to create good classifications using the dataset at hand?*" The dataset contains data gathered by a Blood Transfusion Center in Taiwan. The main question the Center is trying to answer is whether, from the history of someone's blood donations, it's possible to predict whether they'll be coming back to make another blood donation next month.

After this assignment you will be able to …

1) … apply standard practices for evaluating classifiers and understand their differences
2) … select an appropriate classification algorithm given a dataset
3) … reason on a classifier's performance given the data quality
4) … compare the performance of various classification approaches
5) … have a **structured approach to solve classification problems**.

Dataset overview

The blood_transfusion.csv file contains data from 748 blood donors. The blood donation center is trying to learn how to predict whether people will come back to donate blood by analyzing their blood donation history.

In order to do so, the blood donation center looked at historical data. They decided on a cutoff point in time, and calculated some information for each registered donor that had ever been in their facility (see table below) with data from before the cutoff point. They then determined which ones of these donors came to donate blood again in the month following the cutoff point. After the data collection process, they are left with the following dataset:

Feature	Description
months_since_last_donation	Float, number of months since someone's last blood donation.
total_number_of_donations	Float, number of times a person has donated blood.
total_blood_donated	Float, total volume of blood donated in ml .
months_since_first_donation	Float, number of months since someone's very first donation.
class	Integer, **class label**, whether the person indeed came back to donate blood in the month following the cutoff point.

Task 0: Setup Environment (0 Points)

Result: a ready to go Jupyter notebook.
Recommend software and libraries: Python, Visual Studio Code, Jupyterlab

This assignment uses the same setup as specified in the first assignment.

Task 1: Get dataset on screen (10 Points)

Goal: After you have achieved this task, you are able to explore the dataset with basic functions and visualize the importance of features.
Graded result: The result of question 1.1 must be visible in both the notebook and the presentation. The result of question 1.2 must only be included in the presentation.
Recommended libraries: Pandas, Plotly

- Use Pandas to import the dataset "blood_transfusion.csv". This dataset can be downloaded through MS Teams

1.1 Explore the dataset similarly to the first assignment, so you have a good understanding of its content and features. Come up with an interesting one-minute story about the dataset. (6 points)

1.2 Create a visualization, which shows the characteristics of tuples of class 1, compared to tuples of class 0. (4 points)

Task 2: Preprocessing (0 Points)

Goal: After you have achieved this task, you are able to transform the data into a suitable input for classification algorithms.
Result: A preprocessed dataset.
Recommended libraries: Pandas, Sklearn

- Think about the shape of the data that is used as input for classification algorithms. Use preprocessing techniques, which you deem necessary, to apply to this dataset.

Task 3: Creating a train and test set (5 Points)

Goal: After you have achieved this task, you are able to create a train and test set using a function from sklearn.
Graded result: The result of question 3.1 must be visible in both the notebook and the presentation.
Recommended libraries: Sklearn

3.1 Look for a function in the Sklearn documentation which can create train-test splits. Use this function to create two train-test splits of different sizes. *Throughout the entire PA keep comparing the different splits and the impact they have on different tasks*. (5 points)

Task 4: Classification algorithms (25 Points)

Goal: After you have achieved this task, you are able to successfully implement classification algorithms or use existing algorithm implementations.
Graded result: The results of questions 4.1-4.4 must be visible in both the notebook and the presentation.
Recommended libraries: Sklearn

- For the next questions, use the created training sets of both splits from question 3.1. Use the complementary test sets to predict the new cases.
- Predict a few cases that you pick at and compare the predicted labels with the actual labels. You will conduct a thorough evaluation in the task 5. Show that you have developed a feeling for the classification results.

4.1 Manually implement a KNN classifier. You are not allowed to use any predefined functions or libraries, except for auxiliary functionalities, such as the square root function of the Math library. **(10 points)**

4.2 Use Sklearn's Naive Bayes Classifier. Predict a few new cases and compare the predicted labels with the actual labels. **(5 points)**

4.3 Use Sklearn's Support Vector Classifier. Predict a few new cases and compare the predicted labels with the actual labels. **(5 points)**

4.4 Use Sklearn's Multilayer Perceptron (Neural Network) Classifier. Predict a few new cases and compare the predicted labels with the actual labels. **(5 points)**

Task 5: Evaluation of classification methods (15 Points)

Goal: After you have achieved this task, you are able to evaluate classification models using different evaluation measures.
Graded result: The results of questions 5.1-5.3 must be visible in both the notebook and the presentation. Show that you are able to interpret the results.
Recommended libraries: Sklearn

- For the next questions, use the **already fitted/learned models** from the previous task for the splits you have created in Task 3.

5.1 Manually implement a confusion matrix to evaluate the results of the classification models. You are not allowed to use any predefined functions or libraries. **(8 points)**

5.2 Use Sklearn's classification report function to evaluate the results of the classification models. **(3,5 points)**

5.3 Use Sklearn's fbeta score function to evaluate the results of the classification models. **(3,5 points)**

Task 6: Cross-validation (15 Points)

Goal: After you have achieved this task, you are able to understand the use of cross-validation and what its impact is on different classification algorithms.
Graded result: The results of questions 6.1 and 6.2 must be visible in both the notebook and the presentation.
Recommended libraries: Sklearn, Plotly

6.1 Apply Sklearn's K-Fold cross-validation on both splits of one of the classification algorithms to optimize at least one parameter of the algorithm you have chosen. Use scores from task 5 when evaluating each fold. Afterwards, refit the classifier, this time with the best parameter setting. Render plots showing the evaluation scores for each fold. Also render plot(s) showing the difference between not using cross-validation and using cross-validation for this algorithm with both splits. **(15 points)**

Task 7: Reflection (30 Points)

Goal: Practice to communicate your findings to domain-experts and domain-novices.

Graded result: Presentation (***max 15 minutes per group including Q&A from TAs; we accompany more presentation slots to achieve the full 15min/group***) and presentation quality/structure, PDF with screenshots along with descriptive text (e.g., annotations, captions, bullet points to highlight specific findings), data and performance tables and charts, code quality and cleanness, knowledge of each team member of the code (make sure to have the code running during the presentation/Q&A).

Guiding Questions (meant solely as a loose guideline on the kind of questions you should answer in your presentation; list is non-exclusive; does not impose any order; answering all questions is not obligatory for full points; we intentionally give less guidelines from PA2 onwards):

- What defines a good train-test split?
- Compare the different algorithm parameter choices.
- How do you effectively evaluate classification algorithms? Compare the different choices.
- What is the impact of cross-validation?