CSI 4106 Introduction to Artificial Intelligence

Assignment 1: Data Preparation

Marcel Turcotte

Version: Sep 16, 2024 16:50

🎯 Learning Objectives

- **Write** and **execute** a Jupyter Notebook.
- **Download** and **analyze** data in a cloud environment.
- **Prepare** data for a machine learning project.
- **Perform** an exploratory data analysis.

Data preparation is the foundational step in any machine learning project. It involves transforming raw data into a clean and structured format suitable for analysis and model training. The adage "garbage in, garbage out" aptly highlights the critical role of data quality in the success of a machine learning endeavor.

Furthermore, a thorough understanding of the data is essential for selecting appropriate machine learning algorithms and designing an effective experimental protocol. This comprehension aids in identifying relevant features and informs preprocessing decisions, ultimately enhancing model robustness and accuracy.

In this assignment, you will work with multiple datasets from the provided options. In the subsequent assignment, you will conduct an empirical study applying machine learning classification algorithms to a chosen dataset.

📤 Submission

- **Deadline:**
  - Submit your notebook by September 29, 11 PM.
- **Individual Assignment:**
  - This assignment must be completed individually.

- **Submission Platform:**
  - Upload your submission to Brightspace under the Assignment section (Assignment 1).
- **Submission Format:**
  - Submit a copy of your notebook on Brightspace. This copy will serve as the official timestamp of your submission.
  - Optionnally, your notebook **can** include a link to a Jupyter Notebook hosted on Google Colab, allowing the corrector to access and run the code cells. If you prefer a different platform than Colab, ensure that the corrector can access your notebook without the need to install additional software or copy data.

Important Notice: If the corrector cannot run your code, your submission will receive a mark of zero. It is your responsibility to ensure that your submission works from a different computer than your own and that all cells in your notebook are executable.

Requirements

1. Dataset Selection

In this assignment, you will work with multiple datasets from the provided options. All datasets are intended for multi-class classification tasks. Specifically, you need to implement code within your Jupyter Notebook that retrieves the datasets from the web.

1. Glass Identification dataset:
   - Number of samples: 214, Number of attributes: 9, Number of classes: 7 (type of glass like tableware, headlamps, vehicle)
   - www.kaggle.com/datasets/danushkumarv/glass-identification-data-set
2. Dermatology dataset:
   - Number of samples: 366, Number of attributes: 34, Number of classes: 6 (disorders – psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris)
   - www.kaggle.com/datasets/olcaybolat1/dermatology-dataset-classification
3. Maternal Health Risk:
   - Number of samples: 1013, Number of attributes: 6, Number of classes: 3 (risk level – high, medium, low)
   - archive.ics.uci.edu/dataset/863/maternal+health+risk
4. Car dataset:

- Number of samples: 1728, Number of attributes: 6, Number of classes: 4 (unacceptable, acceptable, good, very good)
- archive.ics.uci.edu/dataset/19/car+evaluation

5. Wine quality dataset:
   - Number of samples: 4898, Number of attributes: 11, Number of classes: 11 (0 to 10)
   - https://www.kaggle.com/datasets/yasserh/wine-quality-dataset
6. 16 personalities dataset:
   - Number of samples: 60K, Number of attributes: 60, Number of classes: 16 (personality type)
   - www.kaggle.com/datasets/anshulmehtakaggl/60k-responses-of-16-personalities-test-mbt
7. Credit Score dataset:
   - Number of samples: 100K, Number of attributes: 27, Number of classes: 3 (Good, Standard, Poor)
   - www.kaggle.com/datasets/parisrohan/credit-score-classification

The Kaggle datasets require a login for download, which adds unnecessary complexity to the assignment. To streamline the process, I have uploaded the data to a public repository on GitHub.

- github.com/turcotte/csi4106-f24/tree/main/assignments-data/a1

In your notebook, you can access and read the data directly from this GitHub repository.

2. Exploratory Analysis

1. **Analysis of Missing Values**: Examine the datasets to identify and assess missing values in various attributes. Missing values may be represented by symbols such as '?', empty strings, or other placeholders.

   1.1 In the list of options, what are the datasets that contain missing values? Specifically, which attribute or attributes has missing values?

   1.2 Describe the methodology used for this investigation, and provide the corresponding code.

   1.1 Data imputation involves replacing missing or incomplete data with substituted values to preserve the dataset's integrity for subsequent analysis. Propose imputation strategies for each attribute with missing values.

2. **Select and familiarize yourself with a classification task:** Choose one of the provided datasets for further investigation. It is advisable to select a dataset containing a sufficiently large number of examples, ideally around 1,000, to ensure robust results when applying machine learning algorithms in the subsequent assignment.

   2.1 What is the objective of the task? Is it intended for a specific application? Do you possess expertise in this particular domain of application?

3. **Attribute Analysis**:

   3.1 Determine which attributes lack informativeness and should be excluded to improve the effectiveness of the machine learning analysis. If all features are deemed relevant, explicitly state this conclusion.

   3.2 Examine the distribution of each attribute (column) within the dataset. Utilize histograms or boxplots to visualize the distributions, identifying any underlying patterns or outliers.

4. **Class Distribution Analysis**: Investigate the distribution of class labels within the dataset. Employ bar plots to visualize the frequency of instances for each class, and assess whether the dataset is balanced or imbalanced.

5. **Preprocessing**:

   5.1 For numerical features, determine the best transformation to use. Indicate the transformation that seems appropriate and why. Include the code illustrating how to apply the transformation. For at least one attribute, show the distribution before and after the transformation. See Preprocessing data.

   5.2 For categorical features, show how to apply one-hot encoding. If your dataset does not have categorical data, show how to apply the one-hot encoder to the label (target variable).

6. **Training and target data**: Set the Python variable `X` to designate the data and `y` to designate the target class. Make sure to select only the informative features.

7. **Training and test sets**: Split the dataset into training and testing sets. Reserve 20% of data for testing.

3. Documentation of Exploratory Analysis

Your report should comprehensively document the entire process followed during this assignment. The Jupyter Notebook must include the following:

- Your name, student number, and a report title.
- A section for each step of the exploratory analysis. Each section should contain the relevant Python code and explanations or results.

  - For sections requiring Python code, include the code in a cell.
  - For sections requiring explanations or results, include these in a separate cell or in combination with code cells.

- Ensure logical separation of code into different cells. For example, the definition of a function should be in one cell and its execution in another. Avoid placing too much code in a single cell to maintain clarity and readability.
- The notebook you submit must include the results of the execution, complete with graphics. In other words, your teaching assistant should be able to grade your notebook without needing to execute the code.

Evaluation

- **Overall Effort in the Report (20%)**
  - Ensure the writing is clear and descriptive, enabling the reviewer to easily understand what was done, how it was accomplished, and the underlying reasons.
  - Maintain good separation between text, code, and results for better readability.
  - Provide tests on various examples that the reviewer can easily execute.
  - Facilitate easy comparison between different approaches through visualizations using tables and/or graphs.
  - Ensure the report is detailed enough to allow reproducibility of the results.
- **Dataset Description (10%)**
  - Justify the choice of the dataset.
  - Provide a thorough description of the dataset.
  - Clearly define what the attributes and target represent.
- **Exploratory Analysis (60%)**
  - Analysis of Missing Values (10%)
  - Attribute Analysis (10%)
  - Class Distribution Analysis (10%)
  - Preprocessing (20%)
  - Training and Target Data (5%)
  - Training and Test Sets (5%)
- **Resources and References (10%)**
  - Cite any part of your code that is sourced from websites, including tutorial sites or Stack Overflow.
  - Reference any theory or algorithms utilized that are found in books, slides, or tutorials.

Resources

As previously mentioned, ensure that you cite any parts of your code that are derived from websites, textbooks, or other external resources.

Currently, many programmers leverage artificial intelligence to enhance their productivity, a trend that is likely to continue growing. To better prepare you for the job market, it is plausible to utilize these technologies. However, it is imperative that you fully understand the concepts upon which you are evaluated, as these tools will not be available during in-person evaluations.

If you do use AI assistance, thoroughly document your interactions. Include the tools and their versions in your report, along with a transcript of all interactions. Most AI assistants keep a record of your conversations. The recommended practice is to create a new conversation specifically for the assignment and consistently reuse this conversation throughout your work on the assignment. Ensure that this conversation is solely dedicated to the assignment. Submit this conversation transcript in the reference section of your Jupyter Notebook.

Questions

- You may ask your questions in the Assignment topic of the discussion forum on Brightspace.
- Alternatively, you can email one of the four teaching assistants. However, using the forum is strongly preferred, as it allows your fellow students to benefit from the questions and the corresponding answers provided by the teaching assistants.

Acknowledgement

This assignment utilizes a list of classification datasets curated by Caroline Barrière, and follows the general format of her instructional guidelines. Merci Caroline!