Supervised learning final project guideline.
Andres M. Cifuentes-Bernal 2023

Supervised Methods in Machine Learning.
Guideline: "Supervised learning final project"

Please use this document as a guideline for your team submission of the ***third*** assignment. This assignment has a total mark of ***50*** points.

Your submission for the third assignment should be comprehensive, demonstrating a thorough understanding and practical implementation skills related to supervised learning.

Key points

- This assignment must be performed in teams (2 to 3 people).
- Each student must upload their team files in the classroom.
- Marking is individual and it strongly depends on individual demonstration (presentation).
- **No email submission.**
- **No submissions after deadline (Nov 22 – 2023, no exceptions).** Mind your time.

SUBMISSION FILES

Final Report (PDF)

The report should be structured to cover all the critical elements of supervised machine learning you have learned in the course. It should be written in a clear, professional manner, and can be in English or Spanish. Include the following sections:

1. **Introduction (up to 1 pg.)**: Briefly introduce supervised learning and its relevance in the field of data science and machine learning engineering.
2. **Theoretical Framework (up to 2 pg.)**: Explain the basic concepts of supervised learning, including types of problems (regression, classification), algorithms, training and testing models, overfitting, underfitting, cross-validation, and performance metrics.
3. **Methodology (up to 2 pg.)**: Describe the dataset(s) used, the preprocessing steps, the choice of algorithms, and the rationale behind these choices.
4. **Implementation**: Discuss how the supervised learning models were implemented, detailing the programming techniques and any challenges encountered.
5. **Results and Discussion**: Present the results of your models and interpret them. Compare different models and discuss their performance based on appropriate evaluation metrics.
6. **Conclusion**: Summarize the findings and their implications for a position as a data scientist or machine learning engineer.
7. **References**: Cite all the resources and tools used in your project.

Presentation

Create a concise and informative presentation that highlights the most important aspects of your project:

1. **Introduction Slide:** Introduce the topic and its significance.
2. **Supervised Learning Concepts:** Briefly explain the concepts and techniques used.
3. **Methodology Slide(s):** Outline the steps taken in the implementation process.
4. **Results Slide(s):** Showcase the key results and findings using graphs, tables, or other visual aids.
5. **Conclusion Slide:** Recap the main points and takeaways from your project.
6. **Q&A Slide:** Anticipate potential questions and prepare to address them.

Problem Definition: Supervised Machine Learning for Cancer Gene Expression Analysis

Cancer diagnosis and prognosis often rely on understanding the underlying genetic factors and their expression patterns. Supervised machine learning (SML) offers a powerful toolkit for building predictive models that can classify types of cancer and predict patient outcomes based on gene expression data.

a. Problem Statement

The core challenge of this project is to develop and apply supervised machine learning models to a dataset of cancer gene expressions, with the aim to accurately classify different types of cancers and predict clinical outcomes. This task involves training models on a labeled dataset where the expression levels of genes are used as features and the type or stage of cancer as the target variable.

b. **Requirements**.

Implementation and explanation of the following steps of a data science project are expected: data wrapping, EDA, feature engineering, training of methods, validation, and comparison of performance.

All suitable methods for classification discussed in class should be implemented and compared. (i.e. KNN, logistic regression, decision tree, etc.).

Data Preparation and Model Setup for Cancer Genomics Analysis

Step 1: Installing the DynamicCancerDriverKM Package

The foundation of your project requires the installation of the **DynamicCancerDriverKM** package. You can find the package at https://github.com/AndresMCB/DynamicCancerDriverKM. Follow the instructions provided in the repository to install the package correctly in your R environment. This package will be integral for the later stages of your analysis.

Step 2: Creating a Unified Gene Expression Matrix

Create a single matrix by merging the data from TCGA_BRCA_Normal.rdata and TCGA_BRCA_PT.rdata. Ensure you retain the 'sample_type' column in the merged dataset.

Step 3: Filtering Gene Expression Data

Refine your dataset by removing genes not expressed in at least 20% of the samples. This step is crucial to focus on genes that are likely to be more biologically and clinically relevant.

Step 4: Selecting Genes Based on the PPI Network

Load the PPI.rdata file to access the PPI network. Extract the top 100 genes with the highest connection degrees—these genes are central in the network and will be the predictors for your models.

Step 5: Building Machine Learning Models

The target variable for prediction should be *'sample_type*'.

a. Use the 100 genes you identified as predictor variables in your machine learning models. This approach will test the genes' ability to classify the samples accurately.
b. Use as predictors the top 100 genes (ranked by geneScore) inferred by the *DynamicCancerDriverKM* package when using the PIK3R1 gene as target.
   To do so, adapt the script in demo/Test_DynamicCancerDriverKM(Bulk).R by changing the line 69 to *target <- AMCBGeneUtils::changeGeneId("PIK3R1")*.

Step 6: Evaluate and Compare Models

After training your machine learning models, evaluate them using clear metrics such as accuracy, precision, recall, and the area under the curve (AUC) for the ROC and Precision-Recall curves. Cross-validation should be used to ensure that your models generalize well to new data. Note which models perform best and if there's a significant difference in their performance.

Step 7: Results Discussion

Discuss what the model results mean for identifying cancer from gene expressions. Highlight the predictive genes and their possible biological importance in breast cancer. Discuss differences and similarities of the results from 5.a and 5.b

---

REMARKS AND ADDITIONAL INFORMATION

A. Grading criteria

The following aspects are going to be taken into consideration for this assessment:

- **Understanding and progress**: Answers reflect a proper understanding of the topic, and the discussion is clear with enough details.
- **Presentation**: Document is well structured and follows the guidelines in a professional manner (referencing, professional language, technical discussion, visual aids, etc.).
- **Content**: The content is accurate and replicable.

B. Marking equivalence

For this assignment the maximum mark corresponds to **50 points**, equivalent to **5.0** in the standard marking scheme **0.0 – 5.0**. Approbatory mark corresponds to **30 points**, which is equivalent to **3.0** in the standard marking scheme **0.0 – 5.0.**

Any mark in the range **0.0 – 1.0** will become **1.0.** Following the university internal guidelines.