ARI 510 Lab 1: First ML Experiments

University of Michigan-Flint

Fall 2024

(See course Canvas page for due date)

1 Overview

In the lab-style assignment, you will have a chance to compare several ML approaches using the scikit-learn Python library. This is meant as a somewhat gentle introduction into using ML libraries and training classification models. If you have experience with ML, this may feel simple, but if it is your first time, plan to take some time to familiarize yourself with this process.

2 Instructions

For this assignment we will be working with this heart disease classification dataset from Kaggle. You should perform the following steps.

1. Download the dataset as a csv file from the Kaggle link or the course Canvas page.
2. Load the data in Python, storing it as a matrix of features X and a vector of labels y.
3. Split the data into train, development, and test sets.
4. Choose (at least) 3 different classification models from sklearn's Supervised Learning methods.
5. For each classifier, read about the hyperparameters that you can modify, and choose at least 5 different settings for each model. You are also welcome to transform the data or experiment with any other approaches you like here.
6. Run your various settings (at least $3 \times 5 = 15$ total combinations), training on the training set and testing on the development set each time, and record which combination of hyperparameters gave the best result for each model. Keep track of precision, recall, f1-score, and accuracy.

7. Choose the best configurations for each classifier (3 if you had 3 classifiers) and run the trained model (from the training set) on the test set, which you should not have used before this step. You should define what “best” means.

8. Write your report.

2.1 Deliverables

You should submit:

1. your written report (PDF)

2. the code you used for your experiments

2.2 Requirements

You will be assessed based on the completion of these requirements.

1. Written Report contains the following sections:

   (a) **Introduction:** which classifiers you chose and why you decided to use those.

   (b) **Approach**: how you selected your hyperparameters

   (c) **Results**: present your results (using a figure or a table) from your various configurations on the development set, as well as your final results on the test set

   (d) **Discussion**: how did you determine what “best” means in your experiments? which model worked the best, and why do you think that might be? what did you learn from completing this assignment? what challenges did you face (if any) and how did you overcome them?

   (e) [if you used chatgpt, copilot, etc.] **Generative AI Statement**: short statement about how you used generative AI and your observations about how effective it was.

2. Code:

   (a) make sure your code is readable and documented

   (b) you may either submit a file (.py or ipynb file) directly or link to a github repository containing your code

3 Other questions?

Please feel free to ask on Discord in the #homework-help channel.