CSC 215-01 Machine Learning (Fall 2023)

Project 1: Heart Disease Detection using Neural Networks

Due at 3:00 pm, Monday, October 2, 2023

Demo: class time, Monday, October 2, 2023

1. Problem Formulation

Heart disease is one of the key contributors to human death. According to the WHO, 17.9 million people die each year due to heart disease. In this project, we explore machine learning methods to detect heart disease using tabular data, including both categorical and numeric features. We will model this problem as a BINARY classification problem. Compare the recall, precision and F1-score for each label (1 = heart disease, 0 = normal) in each model, respectively. PLOT the confusion matrix and ROC curve for each model.

- Nearest Neighbor
- Support Vector Machine
- Fully-Connected Neural Networks

For nearest neighbor and support vector machine, you may use **scikit-learn** API:
https://scikit-learn.org/stable/tutorial/basic/tutorial.html#

For fully-connected neural networks, use TensorFlow.

2. Dataset

Download the following dataset which combines 5 popular heart disease datasets over 11 common features.

https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

Examine the corresponding documentation to understand the dataset.

3. Requirements

- Apply train and test split. Use training data to train your models and evaluate the model using test data
- Check and drop any rows with missing values.
- Check and drop duplicates
- Encode categorical features and normalize numeric features.
- Use EarlyStopping and ModelCheckpoint when training with Tensorflow.
- Use TensorBoard to plot the training and test loss when using Tensorflow.
- Use KerasTuner to finetune the following hyperparameters and report the optimal combination of your selected hyperparameters.
  - Activation: relu, tanh
  - Neuron counts
  - Optimizer: adam and sgd

4. Grading Breakdown

You may feel this project is described with some certain degree of vagueness, which is left on purpose. In other words, **creativity is strongly encouraged**. Your grade for this project will be based on the soundness of your design, the novelty of your work, and the effort you put into the project.

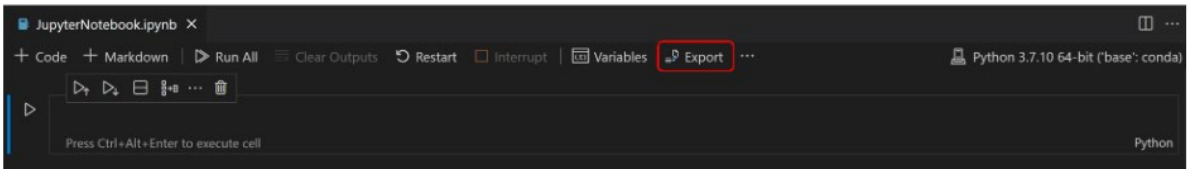
Use the evaluation form on Canvas as a checklist to make sure your work meets all the requirements.

5. Teaming

Students must work in teams of 2 people. Think clearly about who will do what on the project. Normally people in the same group will receive the same grade. However, the instructor reserves the right to assign different grades to team members depending on their contributions.

6. Deliverables

(1) The HTML version of your notebook that includes all your source code. In VS Code, you can export a Jupyter Notebook as an HTML file. To export, select the Export action on the main toolbar. You'll then be presented with a dropdown of file format options.



(2) Your report in PDF format, with your name, your id, course title, assignment id, and due date on the first page. As for length, I would expect a report with more than one page. Your report should include the following sections (but not limited to):

- Problem Statement
- Methodology
- Experimental Results and Analysis
- Task Division and Project Reflection
- Additional Features

In the section "Task Division and Project Reflection", describe the following:

- who is responsible for which part,
- challenges your group encountered and how you solved them
- and what you have learned from the project as a team.

In the section "Additional Features", you describe and claim credit for additional features.

To submit your notebook and report, go to Canvas "Assignments" and use "Project 1".

All the deliverables must be submitted **by team leader** on Canvas before

3:00 pm, Monday, October 2, 2023

NO late submissions will be accepted.

7. Possible Additional Features (5 pts per feature, 10 pts at most)

(1) Among all the features, can you identify the most important features (this is called **feature importance analysis**) and train models only on those most important features, e.g., top-5 most important features? What would be the benefits to do that?

Hint: One option is to use logistic regression to find the most important/influential features.

(2) Clustering algorithms can also be used for classification problems. Can you apply K-means clustering to the training set and use the returned two centroids to classify the records in the test set? Report the recall, precision and F1-score for each label in the test set in this approach.

(3) Can you create **a more balanced dataset** by using oversampling or undersampling to train your model so that you model will not be biased to the more frequent classes?

Hint: Read this article for a possible oversampling technique:

https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

8. In-class Presentation.

On the due day, each team has 5 minutes to present your work in the class. Explain your solutions by referring to your notebook. You do not have to prepare the PowerPoint slides for your presentation.