CSC 215-01 Artificial Intelligence (Spring 2023)

Project 1: Predicting Acute Kidney Injury after Liver Cancer Resection

Due at 3:00 pm, Monday, February 20, 2023

Demo: class time, Monday, February 20, 2023

1. Problem Formulation

Acute kidney injury (AKI) is a common postoperative complication among surgical patients. The incidence of postoperative AKI accounts for 18%-47% of total hospitalized AKI patients. Postoperative AKI can prolong the hospitalization period and increase the risk of both in-hospital mortality and chronic kidney disease.

In this project, we explore machine learning methods to predict the likelihood of acute kidney injury after liver cancer resection. We will model this problem as a BINARY classification problem. Compare the recall, precision and F1-score for AKI and non-AKI in each model, respectively. PLOT the confusion matrix and ROC curve for each model.

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

2. Dataset

Read carefully the following paper, which will be used as our baseline research paper in this project.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7047869/

Download the raw data file "peerj-08-8583-s002.csv" in the section of "Associated Data".
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7047869/#supplemental-information

You may want to refer to Table 1 in the paper for the meaning of each column in the data file. Note that in the data file, the column KDIGO (the Kidney Disease Improving Global Outcomes) indicates whether a patient develops a postoperative AKI or not after liver cancer resection (i.e., 0 means non-AKI while 1 means AKI).

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 (if there is any).
- Encode categorical features and normalize numeric features.
- You must use EarlyStopping and ModelCheckpoint when using Tensorflow.
- You must use TensorBoard to plot the training and test loss when using Tensorflow.
- Tune the following hyperparameters when training neural networks using Tensorflow to record how they affect performance in your report. **Tabulate your findings.**
  - **Activation:** relu, sigmoid, tanh
  - **Layers and 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.

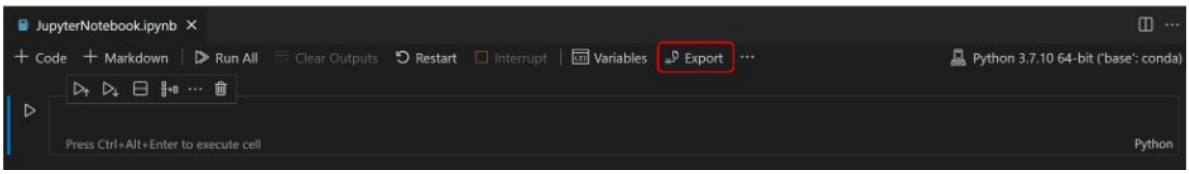
Use the evaluation form on Canvas as a checklist to make sure your work meets all the requirements.

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, February 20, 2023

NO late submissions will be accepted.

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

(1) 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/

(2) 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.

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.