Computational Medicine: 02-518/02-718

Carnegie Mellon University

Homework 3

Version: 1.0; updated 10/4/2020

Due: November 1 by 11:59pm

Hand-in: A single PDF to Gradescope that contains the following items:

1. A cover page that lists your name and Andrew id
    - If you worked on a team, indicate who your teammate(s) is/are by their name(s) and Andrew id(s).
        - Each team should have no more than 3 people.
    - Note: Each team should only hand in one pdf. It does not matter who does the actual upload. We will make sure that the grades are entered appropriately.
2. A PDF export of the Jupyter notebook for question 1
3. A PDF export of the response to question 2.

You can combine all three PDFs into one using Adobe Acrobat, or similar tool.

Overview

In this assignment, you will build models to predict mortality in hospital patients.

This assignment will give you the opportunity to apply methods related to the material covered during lectures 13-16 (weeks 7 & 8).

You can download the data for this assignment here: **HW3_Data.zip**. The file *HW3_Q1_Data.csv* in the zip archive contains data from 121,789 subjects. Each subject is represented by a feature vector with 18 features, plus a label indicating whether the patient lived or died.

The goals of the assignment are as follows:

1. To give you the opportunity to practice learning classifiers
2. To give you the opportunity to practice handling missing data
3. To give you the opportunity to practice handling imbalanced data

Question 1 (85 points)

The first question involves performing biomarker discovery on immunological data using multiple feature selection techniques.

Download the file **HW3_Data.zip** from the homework webpage. Extract the files from the archive. One of those files is named ***HW3_Q1_Data.csv***. That file contains a 121,789 by 19 matrix, plus a header row. The header row contains the names of the features, which include demographic features (sex, age, smoker), some status information (admitted to ICU, Intubated), comorbidities, whether they have been exposed to someone with, or have COVID-19, and whether the patient died in the hospital.

To complete this question, use the provided jupyter notebook.

Part 1.1 (15 points)

- (5 points) Select and apply a filter-based or wrapper-based feature selection method to the data.
- (10 points) Train a classifier using the selected features. Use 10-fold cross validation.

Part 1.2 (10 points)

- Select a learning algorithm that performs embedded feature selection.
- (10 points) Train a classifier using the selected features. Use 10-fold cross validation.

Part 1.3 (15 points)

- (5 points) Select and apply a data imputation method to handle the missing data.
- Apply the feature selection method you used in part 1.1.
- (10 points) Train a classifier using the selected features. Use the same classifier you used in part 1.1. Use 10-fold cross validation.

Part 1.4 (10 points)

- Apply a data imputation method to eliminate any missing values in the data. Use the same method you used in part 1.3.
- (10 points) Train a classifier. Use the same classifier you used in part 1.2. Use 10-fold cross validation.

Part 1.5 (10 points)

- Apply a data imputation method to eliminate any missing values in the data. Use the same method you used in parts 1.3 & 1.4.
- Select a learning algorithm that performs cost-sensitive learning.
- (10 points) Adjust the costs until you find a classifier that maximizes the F1-score, subject to the constraint that it achieves 95% sensitivity for the label 'Y'. Use 10-fold cross validation.

Part 1.6 (5 points)

- Find a classifier that achieves a weighted average F1-score of at least 0.74 using 10-fold cross validation.
    - The weighted average F1-score is computed as follows:
    - Let $F_{1,Y}$ be the F1-score computed using the label 'Y' as the true positive
    - Let $F_{1,N}$ be the F1-score computed using the label 'N' as the true positive
    - Let n_Y be number of instances with label 'Y'
    - Let n_N be number of instances with label 'N'
    - The weighted average F1-score is: $F_1 = n_Y/(n_Y + n_N)\ F_{1,Y} + n_N/(n_Y + n_N)F_{1,N}$
- You may use any method(s) you wish for this part.

Part 1.7 (20 points)

- (10 points) Create a ROC plot with the results from parts 1.1 to 1.6.
- (10 points) Create a table with the following performance metrics for the results from parts 1.1 to 1.6:
    - Accuracy
    - Sensitivity & Specificity
    - The positive and negative predictive values
    - F1-score
    - The Matthews Correlation Coefficient
    - AUC (Area under the ROC curve)

Question 2 (15 points)

In this question, you will create a Concept Map describing the relationships between terms and concepts relevant to the module on biomarker discovery (lectures 13-16).

The focus questions for this Concept Map are:

- "*What kinds of predictive models are used in Medicine?*", and
- "*How do we train predictive models from clinical data?*"

The process of creating a Concept Map begins with the creation of a list of concepts. We have provided a handful of concepts in the list below, to get you started.

2.1 (5 points) Create a list of at least 15 concepts.

2.2 (10 points) Create a concept map using at least 10 concepts from the list above and your response to part 2.1.

Use Powerpoint (or some similar tool) to create the concept map. Save the map as an image or pdf, and include it with your handin for HW3.