Computational Medicine: 02-518/02-718

Carnegie Mellon University

Homework 2

Version: 1.00; updated 9/26/2020

Due: October 18 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 perform feature selection to identify inflammatory biomarkers that can distinguish between one of three conditions in children:

1. Infection by the SARS-CoV-2 virus
2. A rare, but potentially deadly complication of COVID-19, called Multisystem Inflammatory Syndrome in Children (MIS-C)
3. Kawasaki disease, a potentially deadly syndrome of unknown cause.

This assignment will give you the opportunity to apply methods related to the material covered during lectures 9-12 (weeks 5 & 6).

You can download the data for this assignment here: HW2_Data.zip. The file *HW2_Q1_Data.csv* in the zip archive contains data from 1000 subjects. There are approximately equal numbers of cases of SARS-CoV-2 infection, MIS-C, and Kawasaki disease. The final column in the file is a label that indicates which condition the patient has at the time of data collection.

The goals of the assignment are as follows: 1) to give you the opportunity to perform feature selection analysis on clinically relevant data; 2) to have you devise, apply, and justify a method for selecting a small number of features from the 133 in *HW2_Q1_Data.csv*.

Question 1 (88 points)

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

Download the file HW2_data.zip from the homework webpage. Extract the files from the archive. One of those files is named ***HW2_Q1_data.csv***. That file contains a 1000 by 134 matrix, plus a header row. The header row contains the names of immunological markers and label indicating which of the three conditions the patient has. These variables will be discussed in-class.

To complete this question, use the provided jupyter notebook.

Part 1.1 (40 points)

Choose a **filter-based feature selection method** and apply it to the data using 10-fold cross validation.

Part 1.1.1 (8 points) List the top 20 features.

Part 1.1.2 (8 points) Create and plot a 20 by 20 correlation heatmap using the 20 features from part 1.1.1. What is the average of the values in the heatmap?

Part 1.1.3 (8 points) Train classifiers using the top 1, 2, 3, …, 20 features. Plot the 10-fold cross-validated accuracy of the classifiers as a function of the number of features.

Part 1.1.4 (8 points) Filter-based feature selection methods provide an overall ranking of all the features, but often provide no guidance as to which features are truly relevant to predicting the label, versus those with spurious correlations.

Devise a method for determining a threshold that can be used so separate truly relevant features from those that are spurious. Apply it to the data. Train a classifier using those features. Report the 10-fold cross-validated accuracy of a classifier trained on those features.

Part 1.1.5 (8 points) Cluster features you identified in part 1.1.4 into 10 clusters. Select a representative feature from each cluster and train a classifier. Report the 10-fold cross-validated accuracy of a classifier trained on those features.

Part 1.2 (24)

Choose a **wrapper-based feature selection method** and apply it to the data using 10-fold cross validation.

Part 1.2.1 (8 points) List the features that are selected in at least 8 out of 10 folds.

Part 1.2.2 (8 points) Create and plot a correlation heatmap using the features from part 1.2.1. What is the average of the values in the heatmap?

Part 1.2.3 (8 points) Train a classifier using the features from part 1.2.1. Report the 10-fold cross-validated accuracy of the classifier.

Part 1.3 (24)

Choose an **embedded feature selection method** and apply it to the data using 10-fold cross validation.

Part 1.3.1 (8 points) List the features that are selected in at least 8 out of 10 folds.

Part 1.3.2 (8 points) Create and plot a correlation heatmap using the features from part 1.2.1. What is the average of the values in the heatmap?

Part 1.3.3 (8 points) Train a classifier using the features from part 1.3.1. Report the 10-fold cross-validated accuracy of the classifier.

Question 2 (12 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 9-12). Here is an example of a Concept Map.

The focus questions for this Concept Map are:

- "*What are biomarkers?*", and
- "*How do we identify them 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.

Concepts:

- Filter-based methods
- Wrapper-based methods
- Embedded methods
- Disease
- Single Nucleotide Polymorphism

2.1 (4 points) Add at least 10 more concepts to this list.

2.2 (8 points) Create a concept map using at least 10 concepts from the list above and your response to part 2.1.

For this question, you will be graded on whether you demonstrate an understanding of the relationships between the concepts you include in your map (by adding connections and labels on those connections). Points will be deducted if you make "incorrect" connections.

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 HW2.