Software Engineering and Data Science

SEIS 763: Machine Learning

Assignment #5 (100 points)

Due Date: October 11th

The dataset on the Canvas (ML_HW_Data_CellDNA.csv) contains various **numeric** measurements (i.e. size, center, etc) from thousands of bacterium under microscope. The **non-zero** values in the **last column** are the target responses that indicate the bacterium (rows) that are interesting enough for further study. The 0s in the last column indicate the bacterium (rows) are NOT interesting candidates for further study. Convert this target dependent variable (last column) to **binary values** of either 0s or 1s for your **two-class** classification.

1. Use ****logistic + lasso regression**** with ****10-fold cross-validation**** to identify useful predictors.
2. **(This question is optional for Python Program)** Plot a lasso plot with readable tick labels on the X and Y coordinates in your plot for easy visualization and verification.
3. Which top ****THREE (3)**** remaining predictors (with non-zero theta values) are you going to select to explain why a bacteria is an "interesting" candidates for further study?
4. What is the lambda (λ) value in Matlab or the ***C*** value in Python you choose in order to select the top 3 predictors you identified in the last question?
5. What are the $\boldsymbol{\theta}$ values for the 3 selected predictors at the lambda (λ) value in Matlab or the ***C*** value in Python you identified in the last question?

Submission Guideline:

1. Please include the WORD document to include your answers (and clearly readable figures/screenshots) to the above questions. Please include **your name** on the top of your WORD document.
2. Please print your program (matlab or python) as **PDF** and include the **PDF** in your submission. Please name your program as "a5.m/.mlx/.py/.inpyb", depending on the programming language / environment you used.
3. Please also include your program in the formats like .m/.mlx/.py/.inpyb in your submission.
4. Prepare EVERYTHING mentioned in the guideline and submit them on **Canvas** no later than the due date.
5. Please carefully follow the submission guideline. Otherwise, the instructor may not be able to grade your assignment.