Graduate Program in Software

SEIS 763: Machine Learning

Assignment #4 (100 points)

Due Date: June 20th

The dataset (CellDNA.xls) on the Blackboard contains various numeric measurements (i.e. size, center, etc) from thousands of bacterium under microscope. All the measurements are in different units. The last column with non-zero values indicate the bacterium are interesting enough for further study. Otherwise (i.e. last column with zero values), those bacterium are not interesting candidates for further study.

1. Use ****logistic + lasso regression**** with ****10-fold cross-validation**** to identify useful predictors. Plot a lasso plot with readable tick labels on the X and Y coordinates in your plot for easy visualization and verification. Missing clear and readable tick labels in your plot will cost you significant points for this assignment.
2. Which top ****Three**** predictors are you going to select to explain why a bacteria is an "interesting" candidates for further study?
3. What is the lambda (λ) value you choose in order to select the top 3 predictors you identified in the last question?
4. What are the $\boldsymbol{\theta}$ values for the 3 selected predictors at the lambda (λ) value you identified in the last question?

Please follow the instructions below to submit your assignment. Otherwise, your assignment won't be graded and *****NO***** points will be given to your assignment.

1. If you use Matlab, please name your MatLab program as **"a4.m"**. (or your programming language extension). E-mail the program file(s) to the instructor at clai@stthomas.edu before the class on the due date.
2. Have your program reads in your data from the "C:\tmp" folder.
3. Please submit a hardcopy on the due date to the instructor. Your hardcopy should include (1) your programming solutions (with excellent comments) ****AND**** (2) **a clear and readable screenshot** of your answer for ****EACH**** question. Please also staple all pages of your submission together!!! Instructor is not responsible for missing pages if your submission is not stapled together.