Software Engineering and Data Science

SEIS 763: Machine Learning

Assignment #9 (100 points)

Due Date: November 29th

The dataset on the Canvas (ML_HW_Data_CancerGene.xlsx) contains gene expression data from patients with 4 different types of breast cancer. The file has 4 sheets for training X (geneexpTrain), training Y (tumortypeTrain), testing X (geneexpTest), and testing Y (tumortypeTest). Training X has 63 records and 2308 gene-predictors, while testing X has 25 records and 2308 gene-predictors. Training Y has 4 different classes in the 63 records, while the testing Y has 5 different classes in the 25 records. All the predictors are numeric types.

1. Train a single decision tree from the training set to predict the breast cancer types in the testing set.
2. From **test data**, what is the precision / recall / F-measure for **EACH** class?
3. Do you see anything special from your observation in Question 2?

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 "a9.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.