University of Colorado Denver

CSCI 4930/5930 Machine Learning, Fall 2022

Instructor: Ashis Kumer Biswas, PhD

Assignment 1

Background

Non-coding RNAs (ncRNAs) have a multitude of roles in the cell, many of which remain to be discovered. However, it is difficult to detect novel ncRNAs in biochemical screens. To advance biological knowledge, machine learning methods that can accurately detect ncRNAs in sequenced genomes are therefore desirable. In this assignment, you will be exploring the data with structural information of the RNA molecules to understand and evaluate 6 pre-trained classifiers stored in your project directory “models/” as “Model_1.pkl”, “Model_2.pkl”, “Model_3.pkl”, “Model_4.pkl”, “Model_5.pkl”, “Model_6.pkl”.

Dataset Description

In the “dataset/” you will find two csv files:

- “dataset/training.csv” contains 8 features (i.e., variables) about certain number of RNA molecules and the ground truth class labels (which is the last column in the csv file) being either 0 (not ncRNA) or 1 (ncRNA).
    - The entire training dataset was used to develop the 6 models (Model_1, …, Model_6).
- “dataset/validation.csv” contains the same 8 features about another population of RNA molecules, and their ground truth class labels as above.
- Both training and validation datasets are disjoint.
- Here is the list of 8 features:
    - Feature 1: deltaG_total value computed by the Dynalign algorithm
    - Feature 2: The length of shorter sequence
    - Feature 3: 'A' frequencies of sequence 1
    - Feature 4: 'U' frequencies of sequence 1
    - Feature 5: 'C' frequencies of sequence 1
    - Feature 6: 'A' frequencies of sequence 2
    - Feature 7: 'U' frequencies of sequence 2
    - Feature 8: 'C' frequencies of sequence 2
- More information about the 8 features and dataset can be found at the work by Uzilov et al [1]

Project Skeleton Description

You are given the project skeleton in the “Assignment-01.zip”. Besides the “dataset/” and “models/” directories which are containing the aforementioned items, the project also provides you with a jupyter notebook which contains anchors (i.e., placeholders) to 20 tasks for this homework. Your goal is to complete the 20 tasks. *Please write codes in a way to work only on the provided data and model files, and do not print constants to satisfy a requirement.*

Task name	Task that requires you to write code to do the following …	Points possible
A	Print total number of samples in the validation dataset: "dataset/validation.csv"	5
B	Print two numbers in the format: [n0, n1], where n0 represents number of class=0 (negative) samples and n1 represents number of class=1 (positive) samples in the validation dataset: "dataset/validation.csv"	5
C	Print standard deviation of the second feature: "The length of shorter sequence" of the validation dataset: "dataset/validation.csv"	5
D	Print median (i.e., 50% percentile) of the seventh feature: "'U' frequencies of sequence 2" of the validation dataset: "dataset/validation.csv"	5
E	Complete the function "confusion_matrix" partially defined that takes two arrays of target variable "y": y_actual and y_pred denoting ground truth class labels and predicted class labels for the N samples when N is the length of both the arrays. The function should return a list of 4 metrics: TN, FP, FN, TP (in this order).	5
F	You need to complete the accuracy function partially defined in the file that takes a confusion matrix, i.e. the list of the four metrics: [TN, FP, FN, TP], in this order, and return accuracy. In case of Division by Zero error, return -1.	5
G	You need to complete the precision function partially defined in the file that takes a confusion matrix, i.e. the list of the four metrics: [TN, FP, FN, TP], in this order, and return precision. It is also known as Positive Predictive Value (PPV). In case of Division by Zero error, return -1.	5
H	You need to complete the recall function partially defined in the file that takes a confusion matrix, i.e. the list of the four metrics: [TN, FP, FN, TP], in this order, and return recall. It is also known as Sensitivity, or True Positive Rate (TPR). In case of Division by Zero error, return -1.	5
I	You need to complete the F1 function partially defined in the file that takes a confusion matrix, i.e. the list of the four metrics: [TN, FP, FN, TP], in this order, and return F1. It is the harmonic mean of precision and recall. In case of Division by Zero error, return -1.	5

Task name	Task that requires you to write code to do the following …	Points possible
J	You need to complete the MCC function defined in the file that takes a confusion matrix, i.e. the list of the four metrics: [TN, FP, FN, TP], in this order, and return Matthews Correlation Coefficient (MCC). In case of Division by Zero error, return -1.	5
K	You need to complete the FDR function defined in the file that takes a confusion matrix, i.e. the list of the four metrics: [TN, FP, FN, TP], in this order, and return False Discovery Rate (FDR). In case of Division by Zero error, return -1	5
L	Print as a dataframe containing: {model_name,acc,prec,rec,f1,mcc,FDR} for each of the N models (listed in model_files) after predicting the target variables of the validation data: "dataset/validation.csv"	5
M	Print the model name with path which is performing superior among the 6 pretrained models in terms of accuracy, given the performance result dataframe from "L"	5
N	Print the model name with path which is performing the worst among the 6 pretrained models in terms of recall, given the performance result dataframe from "L"	5
O	Scale all the features of the validation set using the formula, z = (x-m)/s, where m = mean of a feature in the training set: "dataset/training.csv" s = standard deviation of the feature in the training set: "dataset/training.csv" # DO NOT SCALE the target feature. # At the end, return a tuple (X, y), with X being a numpy array of shape (N,8) and y is an N dim array and N is the total number of samples in the validation set: "dataset/validation.csv". Store the scaled dataset in a variable (preferably a dataframe) named "validation_scaled".	5
P	Print as a dataframe containing: {model_name,acc,prec,rec,f1,mcc,FDR} for each of the N models (listed in model_files) after predicting the target variables "y" (given) for "X" (the scaled validation dataset)	5
Q	Print the model name with path which is performing superior in terms of accuracy, given the performance result dataframe from "P"	5
R	Print the model name with path which is performing the worst in terms of recall, given the performance result dataframe from "P"	5

Task name	Task that requires you to write code to do the following …	Points possible
S	Flip the prediction of Model 1, and then compute and print as a dataframe containing: {acc,prec,rec,f1,mcc,FDR} on the original (i.e., not-scaled) validation dataset: "dataset/validation.csv"	5
T	Say, in a confusion matrix, the values of the four metrics are: TP=90, TN=1, FP=4, FN=5. Compute F1_original and MCC_original denoting the F1 and MCC scores. Now, flip the predictions, i.e., positives are now will be predicted as negative, and negatives are going to be predicted as positive. Then, compute F1_flipped and MCC_flipped, denoting corresponding F1 and MCC scores. Return the new {TP, TN, FP, FN, F1_original,MCC_original,F1_flipped, MCC_flipped, COMMENT} as a dataframe, where the COMMENT is a string that will be no longer than 200 characters but is going to be your comment about the F1 and MCC values for the two cases.	5
	Total	100

Six more tasks For Graduate Students only

Task name	Task that requires you to write code to do the following …	Points possible
U	This task is a follow up of Task P. There is always cost associated with misclassifications. For instance, if a model predicts a ncRNA (class=1) to be non ncRNA (class=0), further verification will then follow that includes going through the next generation sequencing of those samples costing USD 20 per sample. On the other hand, cost of predicting a non ncRNA to be ncRNA insignificant as most researchers do not care for ncRNAs. They might put it in a piece of paper as a note costing USD 1 per 5 samples. The same cost applies to correct predictions too. What is the cost of predictions with each of the six models?	G-5
V	Please comment on which of the six models is the best on the cost basis.	G-5

Task name	Task that requires you to write code to do the following …	Points possible
W	Please comment on which of the six models is the best overall. Explain your answer.	G-5
X	Again a followup of Tasks O and P: Please scale the given validation dataset with an alternate scaling technique you can think of and repeat Task P with the modified scaled validation.	G-5
Y	Print the model name with path which is performing superior in terms of accuracy, given the performance result dataframe from “X”	G-5
Z	Print the model name with path which is performing the worst in terms of recall, given the performance result dataframe from “X”	G-5

Implementation Requirements

- Please make sure to create a python virtual environment set up using the specific packages (with the noted version) listed in the “requirements.txt” and work in it.
- Other than numpy, pandas and the ones already provided in the notebook, you cannot import any other libraries while solving this assignment.
- Feel free to add as many cells in the jupyter notebook provided. But PLEASE DO NOT DELETE ANY OF THE PROVIDED CELLS.

Submission

In canvas, submit the jupyter notebook file after saving all the outputs from executing the entire notebook and nothing else.

Grading Rubrics

- Each of the 20 tasks has equal weights, which is 5 points. ***And, for the graduate students: the total score is 130 points (5 points each for the 26 tasks) and will be scaled to 100 after grading is complete which will be reflected in the gradebook.***
- With the exception in “T” (Task “T”), hard-coded implementations will be penalized. Please show your work by coding the algorithm, rather than returning only the answer.
- This is an individual assignment. The MOSS system is going to be used in functional level to check for collusion and plagiarism. Any red flag by the system will be reported to the University, and you will fail this course immediately.

Reference

[1] Uzilov, A. V., Keegan, J. M., & Mathews, D. H. (2006). Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC bioinformatics, 7(1), 173.