CSC 371 Fall 2019

Project 3

Due: Oct. 29, 5:00PM

You work at a startup to that receives blood samples from physicians to profile patient cancers to tailor their immunotherapy. You are provided microRNA (miRNA) profiles for a number of cancer patients, drawn from The Cancer Genome Atlas (TCGA) repository[1]. The tissue samples provided represent six different kinds of cancers:

- Breast invasive carcinoma
- Kidney renal clear cell carcinoma
- Lung adenocarcinoma
- Lung squamous cell carcinoma
- Pancreatic adenocarcinoma
- Uveal melanoma

The data corresponding to each cancer type is in its own folder. Each folder contains numerous text files where each file contains the miRNA profile of a single patient. A typical file is formatted as follows (this is a breast cancer example):

```
miRNA_ID        read_count   reads_per_million_miRNA_mapped    cross-mapped
hsa-let-7a-1      64530            15083.167788                      N
hsa-let-7a-2      64740            15132.252946                      Y
hsa-let-7a-3      64296            15028.472898                      N
...
```

The left-most column indicates the name of the miRNA marker — these are the names of our features. The `reads_per_million_miRNA_mapped` column is a normalized value[2] indicating the level to which the specified miRNA was expressed in this patient's tissue — these are the values of the features. Your task on this assignment is to create a tuned classifier for predicting the cancer type, given an miRNA profile like the one above. Thus, one patient's information comprises one example in the dataset. **You must experiment with at least two different classes of models in your work.**

[1] The data is hosted at: `https://portal.gdc.cancer.gov/`

[2] Note that this column has been normalized in a *biological* sense, and not in the mathematical sense.

Unlike our previous work, the data provided to you for this project is not in a format that is "ready for consumption". The relevant information is distributed across multiple files and you'll need to write some scripts to consolidate this into a usable format. In particular:

- You'll need to generate the target labels. Currently, this is encoded by the directory structure of the dataset (for example, all the uveal melanoma examples are in the same subfolder).
- You'll need to verify that the miRNA ID column entries are aligned across files, i.e., that each patient file contains the same number of rows and in the same order. (And if not, you'll need to account for this when consolidating the data.)
- There is a `MANIFEST` file in each cancer folder; this file contains a listing of the file names in that subfolder. This information may be handy when performing the data aggregation.

Finally, here's the paper co-authored by Phillipe Loher et al that inspired this project: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5389567/`

The Write-Up

you need to be sufficiently precise with your writing and include enough detail that a competent reader could reproduce your results. Here are some specific things to address in your report, in no particular order. This is *not* meant to be an exhaustive list.

- Did you perform any preprocessing on the data? If so, describe these steps (and cite sources if appropriate).
- What classification models did you build? How do they compare in terms of performance? What was the best performing model, and how did it do?
- Has your approach been used for this problem before[3]? If so, how do your numbers compare to the published results? If not, how do your results compare to the most similar/related approaches?
- What was your model-building and tuning regime? How did you address overfitting? How did you make hyperparameter choices?

Your primary deliverable for this assignment is a PDF report that provides an accurate representation of your model(s) to doctors that are interested in using your service. This is similar to your model report for the postal service project, with a slight change in audience.

[3] Answer: given that you are attempting to replicate published work, yes!

Recommended Timetable

Here's a recommendation for how to budget your time over the next couple of weeks as you work on this assignment.

- **Oct. 10–12:** Explore the dataset, think about feature engineering, build your first models.
- **Oct. 16–22:** Run more thorough experiments (hyperparameter tuning, further feature engineering, etc.), analyze your results and iterate, search the literature for related work on the problem, begin writing about your data exploration procedure.
- **Oct. 23–25:** Complete experiments, take a step back and think about your report's narrative, write about your experiments and models.
- **Oct. 26–29:** Wrap-up any pending experiments, conclude your witing, revise and proof-read the entire report and submit your final version.