Machine Learning & Data Mining
Miniproject 1

Caltech CS/CNS/EE 155
Released February 7, 2023

1 Overview

Background

The search for life beyond Earth is a driving theme in the 2022 Planetary Decal Survey: to address the civilization-level question, "Are we alone?" In order to translate this search into well-defined planetary mission concepts, we draw insight from our current understanding of terrestrial life while accepting the fewest assumptions possible to preserve sensitivity to exotic forms. While the complexity, size scale, and biochemistry of potential extant life are impossible to predict, terrestrial environments suggest that our search should begin with simple, microscopic forms. One of many biosignatures exhibited by known bacteria and archaea is *motility*. Motility is a compelling biosignature that describes the purposeful movement of an organism, such as to search for nutrients, respond to stimuli, or avoid predators. By detecting the presence of this biosignature from microscopy data with onboard processing, we can prioritize scientifically valuable data to downlink back to Earth.

For this project, you will use data collected by the Caltech/NASA Jet Propulsion Laboratory's Ocean Worlds Life Surveyor[1] Digital Holographic Microscope to determine if a particle is exhibiting motility. While microscopy videos were originally collected, we have generated particle tracks with computer vision for this problem.

The task you are presented with (motility classification) is a part of a larger mission to create accurate and computationally efficient models that may be used in future spacecraft missions to Enceladus, Europa, and other ocean worlds to find evidence of life!

Competition

We will host the competition on Kaggle, a site for data science competitions. There can be 5 submissions per day per team, and the submission window will close on Tuesday, February 14th at 5:00 PM PST. **You may not use any additional data sources.**

In this competition, use the training data to come up with predictions for the test data. Feel free to use any machine learning methods that have been covered in the course so far as well as any other models you see fit. Based on your predictive accuracy on the test data, you will receive a ranking on the competition leaderboard within Kaggle.

The competition website, which includes the datasets and guidebooks describing the datasets, can be found here: https://www.kaggle.com/t/8022d24cc9934a49a70af852ebaf104b

There will be a benchmark submission added by the TAs. We encourage you to beat this benchmark. Jake, other instruction staff, and other JPL MLIA researchers may also compete as an additional challenge. There will be extra credit awarded to those on the top of the leaderboard.

Please follow the format in the sample submission files (`sample_submission.csv`) when generating your submissions to Kaggle.

[1] https://www.jpl.nasa.gov/go/owls

Your task:

Each row in train_basic_features.csv represents a different particle track. The label column indicates whether the particle is motile or not. Your task is to predict the target in test_basic_features.csv to the best of your ability.

Feature Engineering and Preprocessing

The features in train_basic_features.csv and test_basic_features.csv are basic measurements of a particle's motion: mean and standard deviation of the per-step speed, the total length of the track, the end-to-end distance (without considering the path taken), and the total time duration of the track.

There are certainly more useful features that could be calculated from the raw position/coordinate data to distinguish motility. The complete coordinates for each track have been provided as CSVs under train_csvs/ and as a single JSON file under train.json. Note that the Unique IDs (UIDs) identify each track. The notebook used to generate the original features from the JSON is also provided at https://www.kaggle.com/code/jakehlee/basic-motility-features-tutorial. **We strongly recommend that you dedicate some time to design more features**, as better features may perform better than better models. You may also use timeseries methods that don't require features at all, such as RNNs.

The training dataset also includes tracks from two sources: 1) From actual recordings taken by a microscope of microbes 2) From software written to simulate tracks of motile and non-motile particles. These are distinguishable by the prefix of their UID: lab and sim, respectively. **The test set only includes lab tracks from actual recordings**. This is a common problem for scientific applications when real data is expensive to obtain. You should consider whether to train on the entire training dataset, train on only the lab data, or combine them with different weights.

Strategy and Resources

You only have 5 submissions to the leaderboard per day, so while you can wait to see your test accuracy in these submissions, we recommend that you spend time developing a robust validation method to see what techniques work and what don't.

Since we allow for teams of up to 4 students, we recommend that you divide up the work between feature engineering and model development.

Some additional resources for research:

- See https://www.kaggle.com/c/2023-cs155-proj1/code for code tutorials on how the basic features were engineered as well as an example visualization of the track data.
- Dubay et al., "Quantification of Motility in Bacillus subtilis at Temperatures Up to 84°C Using a Submersible Volumetric Microscope and Automated Tracking" Frontiers in Microbiology, Volume 13, 2022. https://www.frontiersin.org/articles/10.3389/fmicb.2022.836808
- Manzo, Carlo and Garcia-Parajo, Maria F., "A review of progress in single particle tracking: from methods to biophysical insights" Reports on Progress in Physics, Volume 78, Number 12, 2015. https://iopscience.iop.org/article/10.1088/0034-4885/78/12/124601/

Performance metric:
Your model will be evaluated using the F_2-measure, a weighted combination of precision and recall that places a higher importance on recall:

$$F_2 = 5 \cdot \frac{\text{precision} \cdot \text{recall}}{(4 \cdot \text{precision}) + \text{recall}}$$

where precision is the ratio of true positives to all predicted positives, or $\frac{\text{TP}}{\text{TP+FP}}$, and recall is the ratio of true positives to all actual positives, or $\frac{\text{TP}}{\text{TP+FN}}$.

See `https://www.kaggle.com/competitions/2023-cs155-proj1/overview/evaluation` for more information.

2 Competition Logistics

- **The competition ends on Wednesday, February 9th at 5:00 PM PST**. Late hours cannot be used for the competition.
- You should work in groups of two to four students (or auditors). The task is feasible as an individual but groups are recommended. To join teams, each participant should sign up for the competition separately, then go to the `Team` tab and `Merge with other Kaggle teams`.
- You can make up to 5 submissions a day. There will not be any private test data, however there is a sufficiently large test set that the best way to optimize your score will to be to optimize your model, not to maximize your submission attempts.
- If you have questions, please ask on Piazza! We also encourage starting early and making a preliminary submission as quickly as possible. TAs will be more than happy to discuss ideas for model selection, feature engineering, and data cleaning.
- You can use open-source tools, concepts you learned in class, and other additional techniques you find online (except for existing code written to model this dataset, although this likely won't publicly exist) to get the best score that you can.
- **You may collaborate fully within your team, but no collaboration is allowed between teams.**
- **You may not search for additional data related to this task; you may only train and validate your models using the provided training set.**
- The top 10 groups on the leaderboard will receive a small extra-credit bonus (up to 5 points, based on ranking). However, we care more about your process and rationale than your model performance. The report and Colab are worth the large majority of your grade.

3 Written Deliverables

- **Report (90 points):** Additional details below. The report should be written to the length specifications. If you have additional insights, you can include them in the extra credit section. We encourage to use graphs in your report and Colab demo, as visualization is very helpful!

- **Colab Demo (10 points):** To help the class learn from each other, you should share a Colab notebook that runs your best-performing model and generates several of the visualizations included in your report. You can include exploratory data analyses, feature engineering, parameter curves, model ensembling, and/or more. Here is a particularly extensive example. Please share the public, read-only Colab link on Piazza in a public note with your team name, and attach the Piazza post link and the Colab link in your report. Don't copy code from other people's Piazza posts.
- **Due date: Wednesday, February 15th at 9:00 PM PST**, via Gradescope.
- **Please submit your report in groups rather than submitting it once per student.** You can see how to submit in groups here:
  https://help.gradescope.com/article/m5qz2xsnjy

4 Report Guidelines

We recommend that you use the LaTeX template provided to you and simply fill in the blanks. To collaborate on the report writing, we recommend using Overleaf (https://www.overleaf.com/edu/caltech), an online LaTeX editor. Caltech students can get a pro account for free using caltech.edu emails.

See our example file for guidelines. The structure is as follows:

1. **Introduction (15 points):** This section is purely for the TAs and should be brief. Maximum of 1 page.
   - Group members
   - Kaggle team name
   - Ranking on the leaderboard
   - F_2 score on the leaderboard
   - Colab link
   - Piazza link
   - Division of labor: Your team should collaborate so that each member has a similar workload. If there is a noticeable discrepancy in the division of labor, team members may receive differing grades.
2. **Overview (15 points):** Concisely summarize your attempts. Detailed explanations should go in the next section. Recommended length of half-page. Maximum of 1 page.
   - Models and techniques tried: What models did you try? What techniques did you use along with your models? Did you implement anything out of the ordinary?
   - Work timeline: What did your timeline look like for the competition?
3. **Approach (20 points):** This section should be a more detailed explanation of how you approached the competition. Maximum of 1 page.

   - Data exploration, processing and manipulation: Did you manipulate the data or the features in any way, such as data cleaning or feature engineering? Did you use the time series data or just the features? Why? What techniques and libraries did you use to accomplish such manipulation? Please justify your methodologies.
   - Details of models and techniques: Why did you try the models and techniques that you used? What was that process like? What are the advantages and disadvantages of using such methods?

4. **Model Selection (20 points):** This section should outline how you chose the best models. Maximum of 1 page.
   - Scoring: What optimization objectives did you use, and why? How did you score your models, and why? Which models scored the best?
   - Validation and test: How did you split your data? Did you use validation techniques? How did you test your models? What were the results of these tests, and what did the results tell you?
5. **Conclusion (20 points):** Summarize the report and your experience. Maximum of 1 page.
   - Insights: Please answer the following questions
     - Among all the features in the data, which features have the most influence on the prediction target? How did you determine which features were important? List the top 10 features. (Bonus points if you can analyze whether these 10 features positively or negatively influence the prediction target.)
     - Overall, what did you learn from this project?
   - Challenges: What could you have done differently? What obstacles did you encounter during the process?
6. **Additional insights (optional, up to 5 extra-credit points):** you can mention any insights you found interesting that you didn't have space to mention above. Be creative! Examples:
   - Why do we use F_2 as our Kaggle competition metric? Consider Type 1 and Type 2 error. Do you think there is a better metric for this project? Why, or why not?
   - Among the machine learning methods that your group used, are there any methods or pipelines that can be parallelized? If so, how? If not, why not? (You are not required to actually parallelize your code)
   - What additional data would be useful to collect for this problem, and why?
   - Did you make any new features from the time series data? If not, propose one that you intuitively think could be helpful.

5 Conclusion

Feel free to ask questions on Piazza and in office hours. Most of all, have fun!