Course Project
Who tweeted that?

BIOM/SYSC5405 Fall 2024
Course Project Overview

20 Nov 2024

Bird song ecology

- Working with Chris Dennison and Rachel Buxton from Biology/Integrative Sciences, we will be recognizing birds from audio recordings
- Data collected 2021, 2022, 2023 nesting seasons in June in 10 locations:
  - years = [2021, 2022, 2023]
  - locations = ["BRY", "CAL", "FIO", "HAR", "KEA", "LAW", "LIF", "MCK", "PEN", "SYL", "WAT"]
- Each recording is truncated to 1 minute MP3 file
- Pre-computing spectrograms for each
- We are interested in detecting one-or-more calls (i.e., presence of bird type) for 10 different bird species within each recording
  - bird_species_list = ['BAOR', 'BLJA', 'COGR', 'GCFL', 'HAWO', 'HOWR', 'INBU', 'NOFL', 'REVI', 'SOSP']

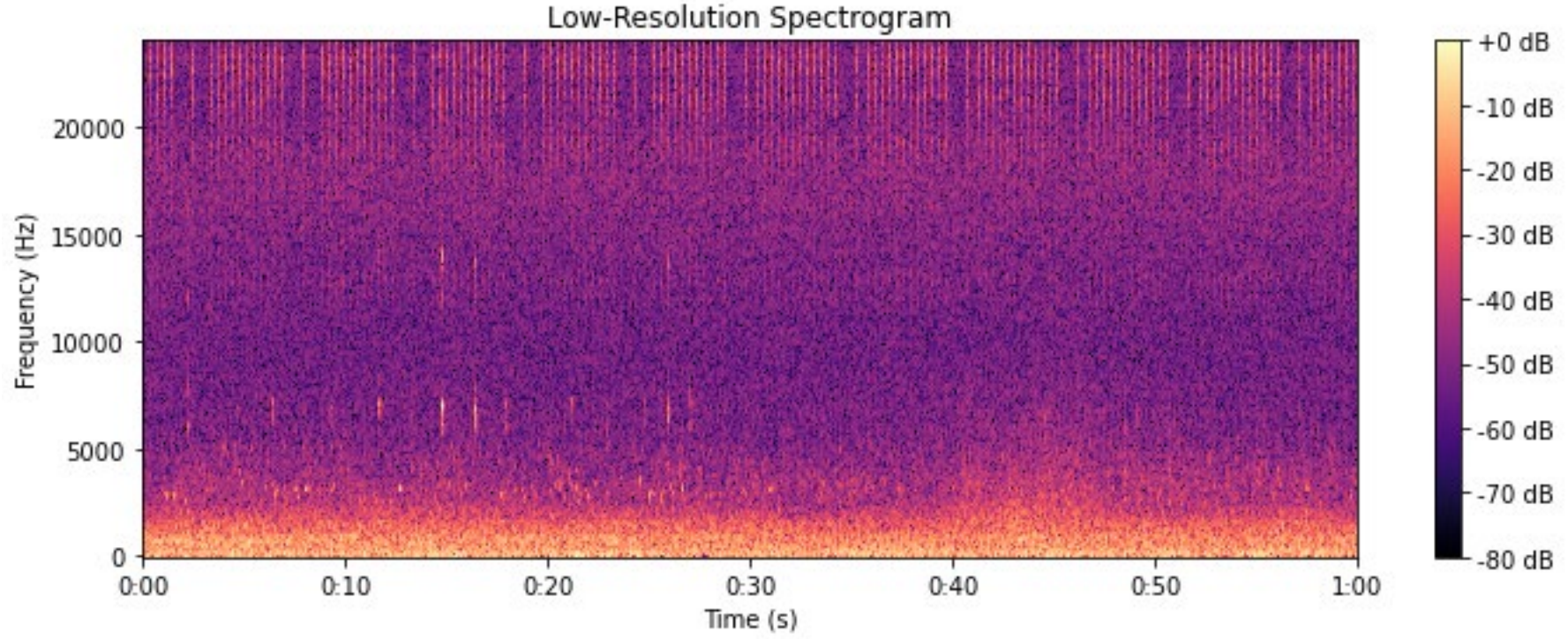
The Dataset

- Data will be structured as follows:
  - `Year`
    - `Location`
      - `labels_train.csv (X rows x 11 columns; cols = fname and 10 bird species)`
      - `MP3: sub-directory containing X MP3 audio files`
      - `SPEC: sub-directory containing X CSV files representing spectrograms for each MP3 file`
  - Each location-year (e.g., BRY-2021) will have a different number of recordings (X Samples), each with 10 classes (multiclass; non-exclusive)
    - Binary class data in `labels_train.csv` (i.e., X rows x 11 columns)
    - `Each row specifies a filename for the source MP3 file and the 10 classes as being either present (1) or absent (0)`
- Data will be released over the coming days. For now, you have all of 2021-BRY
  - Consider online learning?
- Use the data however you like… but watch your download and run time

The Dataset

- MP3 data
  - 1 minute long
  - 1.4 MB each
  - Play one:
- Spec data
  - Spectrograms are of size 257 x 2813.
  - 7MB each. Zipped = ~50%

BRY1_20210617_055000.WAV



The BLIND Test Dataset

- We will withhold some data, likely by year (e.g., part of 2023), as a blind test set.
  - The labels of these data will never be released.
  - The class imbalance in the train and blind test sets are similar
  - After class on Wednesday 4 Dec, you will receive access to the blind test data from which to make your final predictions
    - Watch your runtime!

Your Goal

- Classify each row/sample as one or more of the 10 classes
  - Again, non-exclusive, so a row can have multiple labels
    - Known as a 'multi-label' classification problem
    - Each class may be present (1) or absent (0) from the sample
- You will learn more about the relevance of this task to ecology at the *Awards Ceremony*

Additional Project Details

- You will be evaluated on:

  1) Prediction accuracy over test data set

  - as measured by micro-average F1 score across all samples and all classes

$$Score_{Accuracy} = F1\_microAvg$$

  - $F1\_microAvg$ is the mean F1 score computed over each sample and each of the 19 classes lumped together
    - Tutorials on micro vs. macro vs. weighted F1 score:

      https://www.mariakhalusova.com/posts/2019-04-17-ml-model-evaluation-metrics-p2/
      https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f

  - $F1score = \frac{2*Pr*Re}{Pr+Re}$

** The tutorials above are for the case where "each observation has a **single label**". That is not the case here…*

Additional Project Details

You will be evaluated on:

1) Prediction accuracy over test data set

$$Score_{Accuracy} = F1\_microAvg$$

2) How close your predicted $F1\_microAvg$ is to your actual test $F1\_microAvg$
   - Provide a mean and standard deviation σ

$$Score_{precision} = p(x = F1_{actual}), \text{ if } p(x) \sim N\left(F1_{pred}, \sigma^2\right)$$

Group and method selection – 20 Nov

- The project is completed in teams of 4
  - Simply add yourself to one of the 14 teams on Brightspace (12 teams of 4; 2 teams of 3)
- Select an approach that you will use for the project
  - It must be unique! Each team will take a different approach to solving the problem
    - Differences cannot be as simple as "their method plus some preprocessing", but it can be specific and unique a combination of methods
      - *(although each group may use meta learning, which may include stacking)*
  - Use the link on Brightspace to access the Google Sheets spreadsheet to claim your method.
    - Link becomes active at 1pm on 20 Nov
    - You can change your method any time before Monday 25 Nov at 7am
    - First-come, first-served
    - Don't cheat by editing another row or cell that another group has completed.
- Start looking at the data
  - 2021-BRY Data become available on Brightspace at 1pm

Project Proposal – 25 Nov

- A **project proposal** presentation detailing (4 required elements):
  - Introduce yourselves (introduce each other, not yourself)
  - The pattern classification approach that you plan to use
  - A source for an implementation of your chosen method
  - **A visualization of the data. Something you find interesting…**
- This will be a **5-minute** presentation with ~6 slides.
  - 14 groups * 5 mins + transition time ~= a full class
  - Be ready to go next!
  - All presentations to be submitted to Brightspace by 7am on Monday 25 Nov
- You will be evaluated on the quality of your presentation, the 4 required elements, and your progress to date (i.e., demonstrate that you've started working, have a software framework in place, understand the problem, etc.)
- Each group member should speak for roughly equal time

Develop your method

- Structure your investigation using the following steps:
  - Data pre-processing
    - Normalization, outlier detection, censoring of bad data, etc.
    - Handling of missing data, records of varying length, etc.
  - Feature extraction/selection
    - You may wish to generate new features from the data provided to you or to select only a subset in your classifier.
  - Partition data & establish experiment design
    - Train/validation/test sets, balancing classes (optional), etc.
  - Train classifier
    - What approach used, what hyperparameters required, how they were tuned, etc.
  - Testing & expected accuracy
    - What is predicted F1_microAvg, how was it computed, provide a standard error / standard deviation on your estimate (e.g. "the micro-average F1 score over the test data will be 0.43 ± 0.04")
  - Meta-learning approaches
    - Optionally, try at least one meta-learning strategy (e.g. CME-voting, bagging, boosting, stacking), and investigate its effect on performance.

Project Pitch – 4 Dec

- **The pitch** consisting of a presentation with ~6 slides describing your approach, your predicted accuracy, and how you computed it. Each group will be given **5 minutes** to pitch their method as being the best approach. At the conclusion of this class, all groups will be provided with the blind test data set. Slides should cover:
  a) Quickly review method/implementation
  b) Describe your experiment design
  c) Describe any pre-processing of the data
  d) Describe training/testing protocol
  e) Describe your meta learning strategy (**optional**)
     - However, your final classifier does not need to use meta learning
  f) Provide your estimated accuracy, measured using F1_microAvg (including the standard deviation of your estimate) and describe your methodology for estimating your "true" F1_microAvg
     (i.e. the F1_microAvg you should expect when applied to new test data).

Predict labels for blind test data

- Immediately after the Pitch presentations, the blind test data is released.
  - Keep in mind the size of the dataset (25K samples x 5k features)…
  - Beware of runtime issues (you have ~40 hrs to process all data!)
  - Recall that you will only be able to use one of the six feature sets
    - You will have to declare which file you wish to use, and I will set up permissions accordingly
- Submit a csv text file to BrightSpace indicating predicted classes for each row.
  - File format should follow `labels_train.csv`
  - **Please name your file `Group_##.csv`**

`e.g., Group_05.csv:`

fname	AMRO	BHCO	CHSW	EUST	GRCA	HOSP	HOWR	NOCA	RBGU	RWBL
BRY5_20210622_070000.MP3	0	1	0	0	1	0	0	0	1	1
BRY1_20230605_072000.MP3	0	0	0	1	0	0	0	0	0	0
LAW6_20220612_055000.MP3	1	0	0	0	1	0	0	0	0	0
…										

Order of presentations

14 groups, 5-6 slides,
5 minutes each + < 1min transition

#	Approach	Members	
1	**Recurrent neural networks**		11:40
2	**Feed-forward neural network**		11:45
3	**…**		11:51
4	**…**		11:56
5	**…**		12:02
6	**…**		12:07
7	**…**		12:13
8	**…**		12:18
9	**…**		12:24
10	**…**		12:29
11	**….**		12:35
12	**…**		12:40
14			12:46
	DONE!		12:52

Schedule

GOOD LUCK!!!

✓ **Wednesday 20 Nov:** Competition announced. **Method selection opens @ 1pm.**

Monday 26 Nov: Project proposal presentations

Wednesday 4 Dec: Pitch presentations given.

7am Friday 6 Dec: Final classification of blind data submitted on BrightSpace.

Friday 6 Dec: Results announced. Winners glorified. Prizes distributed.

Wednesday 18 Dec: Final reports submitted **electronically** via BrightSpace.