Baby Cry Prediction

February 29, 2020

Abstract

TODO - A one-paragraph abstract (tl;dr) is required. Summarize the main question and conclusions. Put this below the title but before the main text.

1 Problem Statement and Motivation

TAYLOR: Give a clear statement of the problems or questions the project addresses, their context, and a compelling motivation for why they are worth studying. The problems or questions your project addresses should be original, meaningful, and reflect a deep understanding of the subject and its subtleties.

Briefly review what is already known about the research questions and what techniques others have used to study these questions. Explain the scope of the project and how it fits into existing research.

2 Data

The data we use for this analysis is sourced from a corpus entitled the "Donate-a-Cry Corpus." This corpus, a compilation of audio recordings of babies crying, appears to have originally been collected as part of a campaign entitled the "Donate-a-Cry Campaign" in Sweden in 2015. More infomation about this campaign cannot be ascertained as the listed campaign website in the github repository has been converted into a website for Japanese escorts. Further work on this dataset, including removing background noise, standardizing sample rates, and eliminating outliers, was completed by researchers at the Royal Institute of Technology of Sweden in 2018.

This dataset contains 457 recordings (all in .wav format and all about 7 seconds long) of babies crying, each a different baby and each collected by the baby's mother. These recordings were each labeled with an assumed reason for the cry instance by the mother and then submitted to the corpus. The data contain 5 different label categories: belly pain, needing to burp, hunger, tiredness, and discomfort.

The two main issues to deal with in working with this cry corpus have to do with label representation. Over 350 of the audio files belong to the "hungry" category, comprising over 2/3 of the dataset, while none of the other four categories contained more than 26 audio files. To combat this, we first cut each audio file down to 3 seconds in length under the assumption that any information contained in the first half of a cry would also be contained in the second half. We discared the second halves of all of the hungry crying data, and kept the second halves of all other categories. This allowed us to effectively double the number of samples we had in each category. Second, we duplicated each non-hungry datapoint and added small amounts of noise proportional to each wavelength of the sample. This allowed us to create even more unique data corresponding to these labels while feeling safe that these new datapoints would contain the same features that the non-synthetic data contained. This process left us with exactly 400 unique datapoints to use in training and testing our models.

To extract features from this data, we took the Fourier transform of the sound files, and then normalizing the frequencies produced. This allowed us to eliminate the time dimension and analyze if the frequencies in a baby's cry contain the requisite information to differentiate the reasons for crying.

3 Methods

We began our analysis by testing baseline performance of several models to see how well a variety of models would handle our dataset. At this stage of our project, we are mainly concerned with which models fit our task the best. In further iterations, we will have chosen a model and will have made itentional choices of hyperparameters, architectures, etc. The models we initially used are as follows:

- PCA
- OLS
- SVM
- Random Forest
- Kmeans
- GDA
- LDA

4 Results

At this point in time, we have not successfully been able to identify a featureset/model/hyperparameter combination that performs our classification task any better than $\sim 70\%$ average accuracy. The models with most consistent performance were the Random Forest classifier and the LDA algorithm. The models that performed the worst were the QDA algorithm and the OLS algorithm. Surprisingly, running a PCA on our data, on both the frequencies gathered from the Fourier transform and on the raw wavelength data, suggested that much of the variance could be explained with just the first 25% of components.

5 Analysis

SIMPLE: Give a thorough analysis and a thoughtful discussion of the results and conclusions that can be drawn. Discuss the suitability and effectiveness of the different models and methods for the problems or questions treated.

6 Conclusion

Briefly summarize what you have done and describe the final conclusions that you draw from your computations, results, and analysis.