Machine Learning Sprint 3: Data Cleanup

Michael Asfaw, Killian Lee, Michaelle Ramos
ENGG 408, Dartmouth College

1 Introduction

This report outlines the preparation of a curated music dataset derived from the FMA: A Dataset for Music Analysis GitHub repository, which sources its content from the Free Music Archive (FMA). The dataset was refined to support machine learning tasks such as genre classification. Through comprehensive metadata extraction and audio feature processing, we constructed a clean and consistent dataset of 1,480 songs. Key improvements include standardized genre labels, validated temporal metadata, and structured waveform representations, ensuring the dataset is reliable and ready for training and evaluation of classification models.

2 Metadata Standardization and Feature Extraction

A Python script (`strip_mp3.py`) was developed to process mp3 files from the FMA Small dataset, extracting metadata fields (artist, album, title, duration, bitrate, samplerate, channels, year, genre, track) and stereo audio channels (Ch0, Ch1) using TinyTag [3] and Librosa [2]. The extracted data were serialized to `music_database.pkl`, initially containing 1504 songs.

Further cleaning was conducted in a Jupyter notebook (`Data Cleaning.ipynb`), with the following steps:

- **Genre Standardization**: Inconsistent genre labels (e.g., “Folk” versus “Acoustic”) were resolved using `Genre_Mapping.csv`, which mapped fine-grained genres to 13 coarse-grained categories (e.g., Folk/Acoustic/Singer-Songwriter, Rock/Alternative/Punk) to ensure uniform classification labels.
- **Year Normalization**: The year field exhibited varied formats (e.g., “2008” versus “2008-11-26T00:00:00”). Regular expressions extracted 4-digit years, and 3 entries with invalid or missing years were removed.
- **Duplicate Elimination**: Duplicate song titles were identified, and 21 entries were eliminated, retaining the first instance to reduce redundancy.
- **Feature Extraction**: Stereo channels were converted to mono, and 13 Mel-Frequency Cepstral Coefficients (MFCCs), 12 chroma features, and 1 RMS energy feature were computed using Librosa, yielding a 26-dimensional feature vector per song.

The final dataset comprises 1480 songs with standardized metadata and extracted features. Verification confirmed no fully duplicated rows or empty columns, ensuring data integrity.

3 Dataset Suitability

The curated dataset is well-suited for machine learning due to its clean metadata and comprehensive feature set. Standardized genre labels enable accurate classification, while normalized years and removed duplicates enhance data reliability. The 26-dimensional feature vectors (MFCCs, chroma, RMS energy) capture essential spectral and temporal characteristics, making them appropriate for models such as neural networks or support vector machines. With 1480 songs spanning 13 genres,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

the dataset balances diversity and computational efficiency. The dataset and code are available in a repository [4].

References

References

[1] Defferrard, M., et al. (2017). FMA: A dataset for music analysis. *18th International Society for Music Information Retrieval Conference*. `https://github.com/mdeff/fma`.

[2] McFee, B., et al. (2015). Librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, 18–25.

[3] TinyTag. (2023). A library for reading music metadata. `https://github.com/devsnd/tinytag`.

[4] Asfaw, M., Lee, K., Ramos, M. (2025). Anonymized repository for music dataset curation. `https://github.com/mikaellum/ml-project-team-2`.