Feature Selection

Data set: melodies of Irish dance tunes (of varying length). How can we extract an informative feature vector with constant length over all tunes?

- An n-gram is an ordered list of (in our case) n pitches of notes in a melody, ignoring accidentals.
- Initial features were counts of all 1- and 2-grams.
- Later, split melodies by measure into eighths and counted 1- and 2-grams separately for each measure.
- Finally, considered counts of notes of each length.

Steps taken to improve parsing robustness and quality:

- Unroll all repeat signs before further parsing.
- We consider only tunes with a number of bars in $\{16, 32, 64\}$: an individual A or B section usually has 8 or 16 measures, or 16 or 32 after unrolling repeat signs. We want to restrict to tunes that split evenly into two sections (e.g. not three sections, which could mean ABC, ABA, etc.).
- Decrease number of components via PCA to reduce possible bias and make metric learning computationally feasible.