# Composer Classification with Score Data

Zach Sullivan

2017

# Getting Data

Github projects

- MutopiaProject (Lilypond)
- jshanley/bach-chorales (MusicXml)

# Dataset

Score data from:

| Composer | pieces | primitives |
|---|---|---|
| Bach | 192 | 55,706 |
| Beethoven | 23 | 45,713 |
| Horetzky | 60 | 10,542 |

# Preprocessing Data

A sequence of transformations:

$$\text{Lilypond} \rightarrow_1 \text{MusicXml} \rightarrow_2 \text{Internal Repr} \rightarrow_3 \text{Feature Sets}$$

1. Uses `python-ly` tool (lost: scores with Scheme code)
2. Requires a context-sensitive parser (lost: accidentals and chords)
3. randomly shuffles the data, output is a CSV

# Feature Sets

Dimensions of data:

- Key Signatures: $\{0, 1\}^{17}$
- Time Signatures: $\{0, 1\}^{7}$
- Note $\times$ Duration: $\mathbb{N}^{536}$
- Key Signature $\times$ Time Signature $\times$ Note $\times$ Duration: $\mathbb{N}^{560}$

# Confusion Matrix

| | Bach | Beethoven | Horetzky |
|---|---|---|---|
| Bach | 61 | 3 | 9 |
| Beethoven | 0 | 0 | 0 |
| Horetzky | 0 | 2 | 7 |

Figure: Logistic Regression Classifier on TimeSignature features. Correct: 68 out of 82; Accuracy: 0.829

# Confusion Matrix

| | Bach | Beethoven | Horetzky |
|---|---|---|---|
| Bach | 53 | 5 | 5 |
| Beethoven | 0 | 0 | 0 |
| Horetzky | 8 | 0 | 11 |

Figure: Logistic Regression Classifier on KeySignature features. Correct: 64 out of 82; Accuracy: 0.780

# Confusion Matrix

| | Bach | Beethoven | Horetzky |
|---|---|---|---|
| Bach | 60 | 1 | 1 |
| Beethoven | 0 | 3 | 0 |
| Horetzky | 1 | 1 | 15 |

Figure: Logistic Regression Classifier on (Note,Duration) features.
Correct: 78 out of 82; Accuracy: 0.951

# Confusion Matrix

| | Bach | Beethoven | Horetzky |
|---|---|---|---|
| Bach | 61 | 0 | 0 |
| Beethoven | 0 | 5 | 0 |
| Horetzky | 0 | 0 | 16 |

Figure: Logistic Regression Classifier on all features. Correct: 82 out of 82; Accuracy: 1.0

# Results

| Classifier | Correct | Incorrect | Percentage |
|---|---|---|---|
| Majority | 61 | 21 | 74.4 |
| Random Forest | 82 | 82 | 100 |
| Logistic Regression | 82 | 82 | 100 |
| Multinomial Naive Bayes | 82 | 82 | 100 |
| SVM | 81 | 82 | 98.8 |

Figure: Best results with different classifiers, all achieve using all of the features.

# Conclusions

- Bias towards composers with more pieces
- Different classifiers did not matter much
- Over sampling Horetzky to give him more primitives, did not change the results
- Under sampling Bach and over sampling Beethoven, to have around 60 pieces, increased the variance and decreased performance
- Putting all of the features together increased performance

# Future Work

- More data
- Add chords to features
- Markov Models
- More data

# Other Music ML

Problems

- Music Fingerprinting
- Score Transcription
- Pitch Detection

Data formats

- Audio
- MIDI