Aybüke Ekiz

MAIS 202 - PROJECT DELIVERABLE 2

Art Classifier: Piece-Composer Identifier

Problem Statement

I aim to train a model which will predict the composer of a newly heard (actually seen, because I am using MIDI files, not actual audio to train my model) piece.

Data Preprocessing

I merged two datasets together: Magenta Maestro Dataset [1] and Classical Music MIDI, which I found on Kaggle. I am not using the entire data though, I selected 6 composers who have the most pieces in those two datasets: Chopin, Liszt, Schubert, Bach, Beethoven and Rachmaninoff. They have 100-250 pieces each. Therefore, I have around 800 samples in total. I had to do a lot of preprocessing, because I had to extract relevant information from the MIDI files in a way that the model will be able to use. I followed the chapter from a Springer book "Computational Music Analysis" [2] I have found. They used jSymbolic software to extract several musical features from MIDI files. I extracted 11 different features, again following the method from Computational Music Analysis. I also labelled my data from Magenta Maestro Dataset because it was split into years, not composers; I labeled the samples with the composers by writing a simple Python program. As a result, I have a CSV file containing around 800 samples with 6 possible labels and 11 features. I plan to experiment more with the feature extraction and find more relevant features though, so that is a next step. (The extraction takes much longer than I expected, but I will have plenty of time to let my computer do various extractions in the March break)

Machine Learning Model

In the last deliverable, I stated that I will probably go with either Naive Bayes or Logistic Regression. Last week, Yang suggested that I compare performances of Logistic Regression, Naive Bayes, Decision Tree and SVM, so that is what I did. I used a software called Weka to compare them, which I again learned from the textbook chapter mentioned above. It automatically applies the chosen classification algorithm to the data, so it was very convenient to try out different algorithms and compare them. I used cross-validation for all my trials.

Here is a summary of their accuracy:

Naive Bayes: 22%	Logistic Regression: 33%	SVM: 23%
Rule Learner: 68%	Random Tree: 70%	Random Forest: 85%

I think Naive Bayes, Logistic Regression and SVM are underfitting and others are overfitting. I plan to implement one model from the underfitting cluster, and one model from the

overfitting cluster. This way, I can experiment with regularization in two different cases. Also, I think my dataset would benefit from some more samples, and since I am using 6 composers now, I can easily find/download MIDI files for them, and then my currently underfitting models may benefit a bigger sample size.

Preliminary Results

Since I focused on preprocessing and feature extraction; I have not implemented a model myself yet. I only used Weka's pre-implemented classifiers to experiment with different models. I will be implementing one underfitting and one overfitting model, using Scikit-Learn. I still plan to use the accuracy-precision recall and confusion matrix. I believe the project is feasible. Currently my best model is Logistic Regression (not counting the ones I believe to be overfitting), and it correctly classifies the composers 33% of the time, which is twice the random guess accuracy, since there are 6 possible composers. I aim for an >50% correct classification.

Next Steps

I will implement the chosen models myself, using Scikit-Learn, not using Weka.

I will experiment a bit more with the feature extraction, and try different features on the implemented models.

Then, I will fine tune my models with regularization and hyperparameters.

In the process, hopefully both will not overfit/underfit, and I will select the best performing model. At the testing, I don't think all 6 composers would perform equally well, so I may select the best performing 4 composers for the web page I will build.

References

[1] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang,

Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset." In International Conference on Learning Representations, 2019.

[2] Herremans, Dorien, et al. *Computational Music Analysis*,Springer, pp. 369–391.