Skip prediction using decision trees

Jasmijn Bookelmann
Radboud University
Nijmegen, Netherlands
jasmijn.bookelmann@ru.nl

***Abstract*—Skip prediction is predicting whether a user will skip a song or not. This is a good indication of how much the user likes this song. Therefore, it is an important component of recommender systems. Such as the spotify algorithm. We have created a skip prediction model using the spotify sessions database, which contains listening sessions and metadata of the songs in it. Our model predicts based on this initial session whether a user will skip the next song or not. For creating this model, we have first preprocessed the data by summarizing the session into unary variables. After this we apply decision trees to classify this data.**

I. Introduction

Automatic recommendations are getting increasingly popular with the digitalisation of music. They have an important role in the consumption of music nowadays. [something more about how important this is]

Predicting whether or not a user will skip a song is known as skip prediction. Skipping is a good method to know whether a user will like a song or not. Thus it is often used in recommender systems, such as the playlist creator or autoplay from Spotify.

Spotify released a database containing information about user's listening sessions. [This database has nearly 130 million entires]. These listening sessions contain the tracks in the order the user listened to. [They contain up to 20 tracks.] The first half of these sessions will be used to predict whether the tracks in the second half will be skipped or not.

The session entries have metadata such as: Whether the user has Spotify Premium or not, the action causing the listening session to start, the date etc.

The track entries contain data about their audio features, provided by the spotify API. In addition, the duration, popularity and release year.

A track is skipped if a user did not listen to the entire track. There are three metrics which measure this:

- `skip_1`: The track was only played very briefly
- `skip_2`: The track was only played briefly
- `skip_3`: Most of the track was played
- `not_skipped`: The track was played in its entirity

We will use `skip_2` as ground-truth.

[something more about the data]

II. Methods

A. Preprocessing

[issue of summarizing the data of a session]
[how we summarize the data of a session]
[how we summarize the other data]

B. Applying decision trees

[library used]
[metadata]
[something about cross-reference]

III. Results and Discussion

[results]
[what could have been better]

Acknowledgment

TODO

References

[1] TODO