Machine Learning Engineer Nanodegree

Capstone Proposal

Mario Henrique A. C. Adaniya

December 12nd, 2017

Proposal

Domain Background

Back in the days, we couldn't choose what to watch or listen, because we had a media centralized source. The Internet and the advances in technology came as a breakthrough into entertaiment industry. Now, we don't need to listen to what the DJ radio puts in the air, we can just open some music streaming app and choose what to listen. The same for entertaiment in TV, if we can't find nothing interesting to watch, just open a video streaming service and choose something. And to engage more the customer, these streaming services shows always something interesting to us. For example, if you watched a comedie movie, next time you open the movie streaming app, it will recommend some others comedies movies. In music streaming apps, if you listen to a genre, it will suggest other songs and bands from the same genre.

These are simple examples of **recommender systems**, a system that based on your profile, usage and others data, try to recommend similar items. The 11th ACM International Conference on Web Search and Data Mining (WSDM 2018) challenged the Kaggle ML community to build a better music recommendation system using a donated dataset from KKBOX. The challenge consist in build a music recommendation system. The challenge is that a person can listen to different kinds of genres, sometimes it just like one song from one singer of that genre, and like bands from another genre that is completely the opposite.

The Recommender System problem is discussed in several papers. In [1], the author discuss some algorithms and the importance to Netfix business to engage the customer. They show a prototype workflow, where briefly follows: training model -> evaluate the model -> if good results: test A/B else: generate hyphotesis. At the end, they still summarize some points that can improve the algorithm.

In academia, the recommendation problem, applied to music is also topic in researches: in [2], the authors evaluated the effects of combining contextual information with the Item-based Collaborative Filtering technique to recommend music in the Long Tail; in [3] the authors present a weighted hybrid recommender approach using history based similarity point, an additional serendipity measure based on complementary music to the list of recommended tracks and weighted hybrid recommender.

Interest

The main problem in recommendation is interesting because there are a lot of techniques that could be applied and mixed up to improve the results and the challenge is that listen to music is not that discrete and binary problem, because the music preference is a personal and subject. And in this challenge, the dataset is from real users, where some extreme or outlier scenarios might occur and mislead the algorithm.

References

[1] GOMEZ-URIBE, CARLOS A. and HUNT, NEIL **The Netflix Recommender System: Algorithms, Business Value, and Innovation** ACM Transactions on Management Information Systems, Vol. 6, No. 4, Article 13, Publication date: December 2015.

[2] Domingues, Marcos Aurélio; Rezende, Solange Oliveira. **The Impact of Context-Aware Recommender Systems on**

Music in the Long Tail. Procedings of the II Brazilian Conference on Intelligent Systems - BRACIS 2013.

[3] T. Hornung, C. N. Ziegler, S. Franz, M. Przyjaciel-Zablocki, A. Schätzle and G. Lausen, **Evaluating Hybrid Music Recommender Systems** 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, 2013, pp. 57-64.

Problem Statement

Nowadays, streaming services are present in our daily life. Video and music recommendations are common as well. If you listen to some genre, the probability that you would listen to another band or song of that genre is higher. But if you like "opposites" genres, it is more difficult to the streaming service suggest something for you. KKBOX, Asia's leading music streaming service and the 11th ACM International Conference on Web Search and Data Mining (WSDM 2018) created a challenge to build a better music recommendation system. The challenge is a classification problem to predict the chances of a user listening to a song repetitively after the first observable listening event within a time window was triggered.

Datasets and Inputs

The dataset is provided by KKBOX and is available at the WSDM - KKBox's Music Recommendation Challenge and the capstone repository. It have a training and a testing dataset where `target` is marked `1` if there are recurring listening event(s) within a month after the user's very first observable listening event, and marked `0` otherwise. Dataset consists of information of the first observable listening event for each unique user-song pair within a specific time duration. Metadata of each unique user and song pair is also provided by KKBOX.

The challenge provides the following files:

- `train.csv` : 7377418 entries, 6 columns: msno (object), song_id (object), source_system_tab (category), source_screen_name (category), source_type (category), target (uint8).
- `test.csv` : 2556790 entries, 6 columns: id (int64), msno (object), song_id (object), source_system_tab (object), source_screen_name (object), source_type (object).
- `songs.csv` : 2296320 entries, 7 columns: song_id (category), song_length (int64), genre_ids (category), artist_name (category), composer (category), lyricist (category), language (category).
- `members.csv` : 34403 entries, 7 columns: msno (object), city (category), bd (uint8), gender (category), registered_via (category), registration_init_time (datetime64), expiration_date (datetime64).
- `song_extra_info.csv` : 2295971 entries, 3 columns: song_id (object), name (object) isrc (object).

Solution Statement

A hybrid approach might help improve the results. As we have the target in training, we could use a unsupervised learning method and combine with a supervised learning method to find new groups into the groups we already discovered in the supervised. The unsupervised method applied will be the K-means clustering, depending on the results and groups created, it will be the input for a supervised learning method, ensemble method using random forest. As the ensemble is built from a sample and using the random forest, the final tree is constructed using the best split among random subset of the features trees, it could improve the final result. The general idea is create subsets from the data itself, using the unsupervised method. Thus, after some patterns were founded, them use a supervised method in each cluster to trying to improve results.

The input features to unsupervised will be the ones at `songs.csv`, and it will need some data preparation as well, such as cleaning, feature extraction, and others. The expected generated groups, therefore, will be combined with the `train.csv` and `members.csv` to be the input for the supervised method.

Benchmark Model

As the benchmark model will be very simple, the baseline will be created using a Dummy Classifier. Also, as a supporting metric I will compare my model results with the public leaderboard from Kaggle.

Evaluation Metrics

To evaluate the proposed solution, basics metrics will be used:

- False-positive rate = False-positive/Number of normal data
- True-positive rate = True-positive/Number of normal data
- Precision rate = True-positive/True-positive+False-positive

To help visualize the results and parameters, the Receiver Operating Characteristics (ROC) graph will be used as well. It is a technique to visualize the performance based on the parameters and demonstrated the better trade-off between false-positive rate and true-positive-rate.

To evaluate the clustering, the Davies-Bouldin, Dunn's indexes and Silhoutte will be applied.

Project Design

The solution will be following some steps:

1. **Data cleaning:** Methods to clean the dataset will be applied, such as outliers cleaning, nil or missing values might be excluded or some technique to fill the missing value might be applied.
2. **Summary data:** Statistical Inference will be applied, to analyse expected values, variability, distribution of dataset to conclude and infer some hypothesis.
3. **Unsupervised method:** The K-means will be applied to split the data in a first phase of the solution. It will create groups, these groups will be evaluated using some clustering metrics: Davies–Bouldin index, Dunn's index and Silhouette. Thus, after finding the best centers and groups, it will be the input for the unsupervised phase. The features which will be applied are under study, because the dataset consist in multiple files, and maybe some new features might be extracted as well.
4. **Supervised method:** Applied ensemble method using random forest, where each tree in the ensemble is built from a sample drawn with replacement from the training set. When the node will be splitted during the construction of the tree, the split that is chosen is the best split among a random subset of the features.
5. **Evaluating:** This step is the final step, after training the dataset into the supervised and unsupervised steps, the results of the classification will be summarized and organized to get a conclusion if the solution did a good job or not.