

Yingfeng_Lou_EE660_H6W8_Type_1_Project_Proposal_F20_v1. pdf

by Yingfeng Lou

Submission date: 23-Oct-2020 12:23AM (UTC-0700)

Submission ID: 1424036055

File name: Yingfeng_Lou_EE660_H6W8_Type_1_Project_Proposal_F20_v1.pdf (207.83K)

Word count: 911

Character count: 4456



EE 660

Homework 6 (Week 8): Type 1 Project Proposal

Posted: Fri., 10/16/2020

Due: Fri., 10/23/2020, 5:00 PM PDT

This proposal form is for Type 1 projects: Solve a ML problem by implementing a ML system of your own design, using real-world data.

Please fill in both the Project Proposal form (pp. 1-2) and the Dataset Information Form (p. 3). This is required of everyone (each team submits one HW6 with all their names on it). *All fields except "other comments" are required. In each field, replace instructions (black text) with your descriptions. Preferred format is to enter your answers into the Word version of this form, then convert to pdf before submission. If you prefer to use another app instead of Word, then submit a typed version with each field labeled with its title ("Dataset", etc.), and submit as a pdf file.

Please note that this proposal will not be graded like a regular homework. The primary purpose is to give you some feedback on your project topic and plans; the scoring on this homework will be primarily based on whether you put in a reasonable effort and whether the content makes good technical sense.

*Hit Song Prediction

***Project team: Your name(s) and email address(es)**

Yingfeng Lou 3099544617

Email: louy@usc.edu

***Clear statement of the problem and/or goals.**


Hit Song Problem

To see if pop music is really formulaic, I would like to use different classification methods, like logistic regression, SVC, decision tree, to predict which song will be a hit song by seeing if it charts on Billboard Hot 100 Hits through 2000-2019 data (over 12000 data point) and if it is belonged to mainstream genre.

The goal is finding the best model that can predict if the song will be a hit song.

*A plan of preprocessing and feature extraction (if applicable)

Preprocessing:

1. Deal with the outlier. By evaluate the mean and std of these features, adjust the outliers' value.
2. missing data (only 16 data points, a little proportion of the dataset, have missing data, I plan to delete these directly);
3. Split the data to training set and test set, and the validation set will from training set but it is not used for training.
4. Standardization. Deal with some data of features are not on the same scale;  2

Feature extraction:

1. Plot each feature for coarse selection, also it can visualize the correlation among these features;
2. Perform PCA.

*A plan of your approach

1. Do preprocessing and feature extraction as described above;
2. Separate the dataset into the proportion 8:2 for training set and test set and extract the validation set from training set (not in the process of training, just for cross validation);
3. I plan to use 3 machine learning methods to this classification problem. They are: Logistic Regression, SVC, Decision Tree. (also consider l1 or l2 regularization) and compare the results among these algorithms.
4. Evaluate my system by the accuracy of the test set to see if my prediction will be right. Also, precision, recall or confusion matrix are important factors when evaluating my model.

*A description of any other work of yours that is related to your class project

None

*If yours is a team project, roughly describe how work will be divided

I will do it individually.

Other Comments

The challenge in this project can be from 2 parts: preprocessing and feature selection

There may be some challenge for me to assemble disparate type of features, not only numeric, also string, categorical types. Also, I find out I need to deal with the outliers after seeing the mean and std of the data.

Include one form for each dataset you plan to use. (For each dataset's form, you may continue onto an additional page if necessary.)

***Dataset or competition title:** The Spotify Hit Predictor Dataset

***Link:** <https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset?select=dataset-of-10s.csv>

***Problem type:**
classification/logistic regression



***Brief description of dataset and problem domain:**

A dataset consisting of features for tracks fetched using Spotify's Web API. And Billboard Hot 100 Hits is a convincing way to show the success of a song. So labeled tracks with hit('1') if they have been on Billboard Hot 100 Hits and belonged to mainstream genre, otherwise labeled "0".

***Number of data points:** 12272

***Number of features or input variables:** 18

***Feature or input-variable types:**

14 numeric features, among them 9 features are decimal type between (0,1), 2 features are decimal type larger than 1, 3 features are integer type. 1 feature is categorical (music mode), 3 features are string (tracks' name, artist, uri)

***Label (output) type:**
binary categorical

***If Label Type is Categorical, is the number of samples significantly unbalanced (maximal variation of more than a factor of 2)?**

No.

***Has Missing Data?**

Yes (give idea of how prevalent, if known)

Only 16 output labels are missing, which is little proportion of the whole dataset.

***Is the problem/dataset a Kaggle competition (current or past)?**

No

***Any other comments on the dataset:**

I think the challenge for me is how to assemble different kind of features together, because the features include string, numeric, categorical types and they are all contributes to if the song will be a hit song. Therefore, how to do the feature selection will be very important.

FINAL GRADE

GENERAL COMMENTS

Instructor

9/10

PAGE 1



Comment 1

Your proposal is ok. But

for this to become a good project, you must try several ML methods (along with hyperparameter selection) and you must add some exploratory ML analysis. For example, you can artificially create missing data and see how that affects performance. You can also delete labels and see how semi-supervised learning performs compared to the original supervised case.

PAGE 2



Comment 2

standardization should be trained from training set and applied on validation and test set

PAGE 3



Comment 3

logistic regression is an algorithm, not a problem type.

PAGE 4