Movie Rating Prediction Using Machine Learning

Grigor Keropyan*
Department of Mathematics and Mechanics
Yerevan State University
goqorkeropyan@gmail.com

Abstract

Movie rating prediction based on information available prior to theatrical release is important in order to understand how successful will be movie. This paper describes various Machine Learning methods to predict movie rating.

1 Introduction

Movie rating prediction is becoming a popular problem and various methods have been suggested. In this paper we will introduce Machine Learning (ML) methods to predict movie rating based on the data available prior to the theatrical release. We used IMDB open-source data such as posters and run-time of movies.

Although, it seems impossible to predict the movie rating based on the information available before theatrical release, we will see that ML methods predict them quite accurately. One of the explanation could be that people may like movies if their preferred actor is in the crew or films are produced by their favorite producers.

One of the goals of this work is provide a tool which can help producers to promote their films to be successful. Another one is that this work provide a good recommendation for people predicting their IMDB score.

2 Related Work

This work is based on the Stanford students paper [1] where they predict IMDB score of movies. In [1] authors used ML based methods to predict IMDB score. Another research which uses Bayesian approach [2]. This paper is based on [1] and uses new ML algorithms such as lightgbm to reach the better accuracy than [1]. Our preproccessing is almost the same as [1] and we have about 20K examples. We have shuffled the dataset and splited into training and test test correspondingly 80 and 20 precents. In the algorithm we have done K-Folds cross validation and after tuning the hyperparamters we have achieved better accuracy than [1].

3 Dataset and Features

We have used open source data, "Movie Genre from its Poster Dataset" [3] and "The Movie Dataset" [4]. The dataset is preprocessed like [1], taking only movies which original language is english and which produced after 1980. The features are divided into 3 groups: text (synopses), images (posters) and others (runtime, genre, director, actors). From poster we have extracted some features: such as

*https://github.com/grigor97

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Table 1: Dataset by Groups

Data	Type	Dimension	Example
Actors	Categorical	1671	Tom Hanks
poster	Numerical	13	Number of faces = 0
Director	Categorical	491	Howard Deutch
Runtime	Numerical	1	121 (minutes)
Genre	Categorical	23	Documentation
Synopses	Categorical	3712	Once

number of peoples in the poster, mean and standard deviation of blue, green, red, hue, saturation, brightness as in [5]. For summaries we have used count vectorizers and only left the word which appeared more than 20 times. Directors are have been made one hot and actors count vectorizers where we have only left top three actors in each movie. The example of result is shown in Table 1.

After filtering the english movies and released after 1980 it remained 18850 movies. Which we have splited in two ways. First: like [1] implementation, when it is separated train, validation and test sets correspondingly 70, 15, 15 percents. Second: which gave a better result it is separated into train and test sets correspondingly 80 and 20 percents.

4 Methods

These instructions apply to everyone.

4.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2018` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2018}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

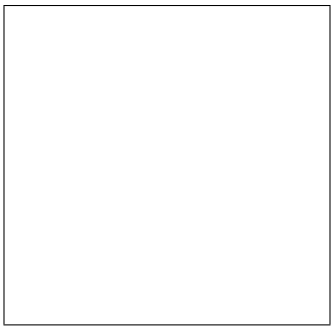

Figure 1: Sample figure caption.

Table 2: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

4.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[2] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[3]

4.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

4.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 2.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 2.

[2] Sample of the first footnote.

[3] As in this example.

5 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

6 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

6.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

[1] Yichen Yang et al., *"Predicting Movie Ratings with Multimodal Data"*, `http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26260680.pdf`.

[2] Y. J. Lim and Y. W. Teh, *"Variational bayesian approach to movie rating prediction"* Proceedings of KDD Cup and Workshop, vol. 7, 2007, pp. 15–21.

[3] KaggleInc, *"Movie genre from its poster,"* `https://www.kaggle.com/neha1703/movie-genre-from-its-poster` .

[4] Kaggle, *"The movies dataset,"* `https://www.kaggle.com/rounakbanik/the-movies-dataset` .

[5] F. B. Moghaddam, M. Elahi, R. Hosseini, C. Trattner, and M. Tkalci˘ c, *"Predicting movie popularity and ratings with visual features,"* in 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). IEEE, 2019, pp. 1–6.