Alma Mater Studiorum - University of Bologna
LM Informatica
Data Analytics Project

Cotugno Giosué [TOADD] - Pruscini Davide [1007343]

July 6, 2022

Contents

Chapter 1

Introduction

In this report we will discuss about the whole data pipeline of a project that uses different techniques of machine learning to classify some tabular data. In particular, we choose the first project proposal that aims to predict the average mark of a film from its features. The first step of the data pipeline requires to download, save and load in memory the MovieLens [1], TMDB [2] and IMDb [3] datasets. Consequently, the datasets were elaborated with the objective to generate a unified dataset that can be used as an input for some machine learning algorithms. Afterwards, during the modelling phase we built an MLP model thanks to the PyTorch framework [4]. In addition, we used other techniques like SVM, tree methods and naive bayes methods that are available into the Scikit-Learn library [5]. In order to find a good configuration during the performance analysis, it was mandatory to define a large enough hyperparameters space for all the models that we defined. Moreover, in the last phase, the cross validation technique was used to obtain more robust statistics results that has been compared between the trained models.

Chapter 2

Methodology

In the next chapter there will be the explaination of the data pipeline that the project followed. In particular, each subsection will focus on a specific task, except for the data visualization that has been used only when needed.

2.1 Data Acquisition

The used datasets are downloaded at runtime directly from the sources. These datasets come directly from MovieLens' page, which provides 6 different datasets:

Dataset	Features
ratings.csv	userId, movieId, rating, timestamp
tags.csv	userId, movieId, tag, timestamp
movies.csv	movieId, title, genres
links.csv	movieId, imdbId, tmdbId
genome-scores.csv	movieId, tagId, relevance
genome-tags.csv	tagId, tag

where most of these datasets provides information for approximately 60000 films. The links dataset provides two identifiers that allow to collect information from the IMDB and TMDB databases. Thanks to the links' features, it has been possible to collect some more information on the running times of the films that could provide more insight into them. Talking about the ground truth of the supervisioned models, the rating mean is missing. So the target feature will be computed during the pre-process phase thanks to the ratings dataset. Further information about the features usage and the cardinality of the datasets will be discussed in the Pre-Process section.

2.2 Data Pre-process

In this section, will be discussed the Pre-Process phase for each of the above presented dataset. In order to achieve a major clearity, the work on each dataset will be discussed in a specific subsection where the operation computed on them will be discussed.

2.2.1   movies.csv

Inside this dataset, the title and generes features contains multiple information. In particular, the title has been splitted in two part, where on one hand there's only the title name, and on the other, there's the year of the film production. Since, the title name is a string, that doesn't add more information to a classic machine learning model, this feature has been converted into its length meanwhile, the production year just constitutes a new feature. The genres feature contains a pipe separated list of a fixed possible values. Since the list is just saying if a film has a specific genre or not, to each film, all the fixed values has been added as a feature, and if a genre appears into the genres feature, that column will result into 1 that indicate the presence of that genre, 0 otherwise. At the end of this initial phase, the movies dataset looks like:

movieId	title_len	year	action	adventure	animation	...	Western	(no genres listed)
1	16	1995	0	1	1	...	0	0

In order to finish the data Pre-Process on this dataset, the data cleaning is required. First of all, there are some films where the year of the film production is missing. Analyzing the distribuition of these it was possible to see that the distribuition is right skewed, so it was possible to fill the missing values with the median.
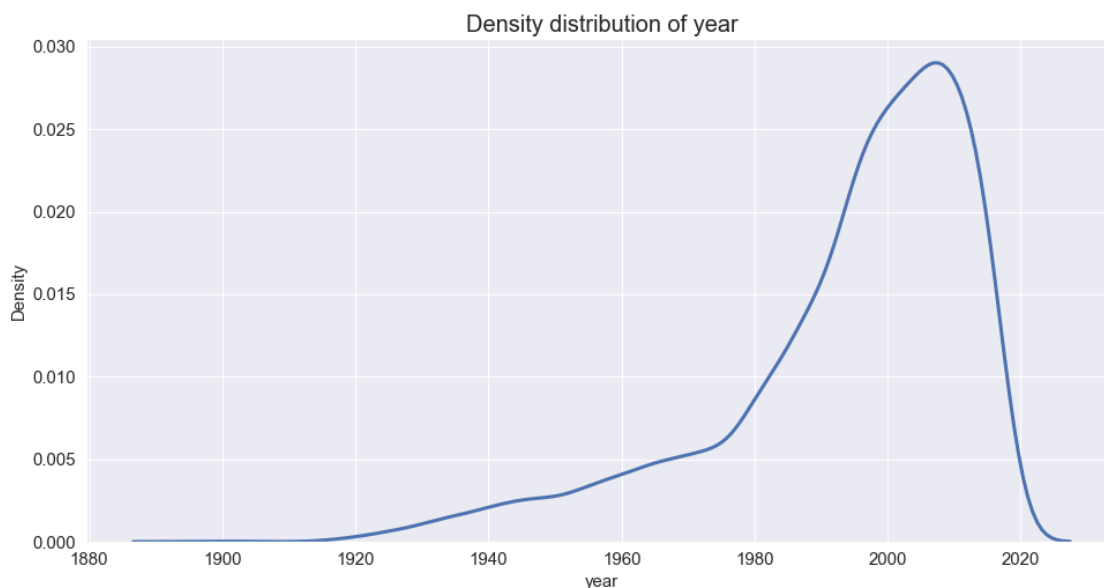

Figure 2.1: A boat.

2.3   Modeling

2.4   Performance Analysis

Bibliography

[1] GroupLens. MovieLens Latest Datasets. `https://grouplens.org/datasets/movielens/latest/`.

[2] TMDB Community. Api Overview. `https://www.themoviedb.org/documentation/api/`.

[3] IMDb.com. Imdb datasets. `https://www.imdb.com/interfaces/`.

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.