Predicting Popularity Scores by utilizing Recurrent Neural Network (RNN)

Project Description

In this project, you may need to predict the popularity score of movies, where the main objective is to improve the prediction accuracy as much as you can. In order to achieve this goal, you can implement the prediction model by using Recurrent Neural Network (RNN) model (such as Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU)), and also you can use the hybrid model (the combination of Convolutional Neural Network (CNN) + RNN).

Requirements

Dataset: MovieLens (http://files.grouplens.org/datasets/movielens/ml-latest.zip)

Package need to install: pandas, numpy,matplotlib, seaborn

Python version: 3.6

Machine learning libraries: tensorflow, keras, pytorch

Dataset property

This dataset (ml-latest) describes 5-star rating and free-text tagging activity from (http://movielens.org), a movie recommendation service. It contains 26024289 ratings and 753170 tag applications across 45843 movies. These data were created by 270896 users between January 09, 1995 and August 04, 2017. This dataset was generated on August 04, 2017.

Original dataset formatting and encoding (please check README.txt)

Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970.

Data cleaning process

You can download sample code from the Github link preprocess folder (https://github.com/kyithar/class)

It includes

1. Join_dataset.py (To join two csv files)
2. Preprocessing.py (To calculate the count and normalized count to make label)
3. Sort.py (Sort dataset with timestamp (second))
4. to_seq.py (To get sequences of inputs)

We will also provide sample code for prediction model with LSTM (https://github.com/kyithar/class)

Example prediction model

For the prediction model, you can consider four types as follows,

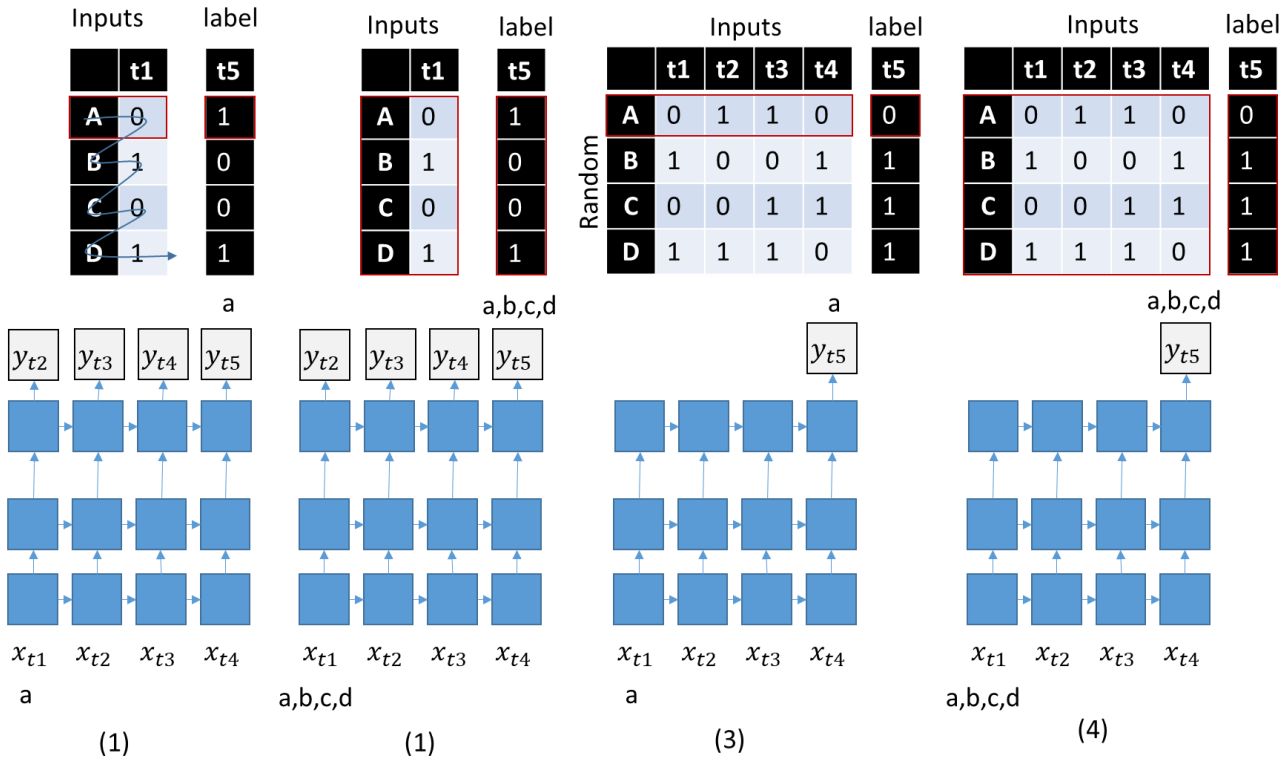

Figure 1 Prediction models

Details project description

1. Run sample data cleaning codes to understand data cleaning process. In this project, you need to use two csv files from the dataset 1) movies.csv, and 2) ratings.csv.
2. Do (Exploratory data analysis) EDA for the cleaned dataset to know which features are import.
3. Construct the prediction model by using open source machine learning libraries. (The goal is to improve accuracy).
    a. Please test with different type of RNN cells such as Long Short Term Memory, Gated Recurrent Unit and so on.
    b. Please test with different hyper parameters such as number of hidden layers, learning rate, number of time slots, mini batch size and so on.
    c. Please test with different optimizer (such as Adam, Stochastic Gradient Descent [SGD]), different types of loss function (such as mean squared error, root mean squared error, cross entropy and so on)
    d. Please show the results such as accuracy, prediction loss and graph of your prediction model.
4. Upload code and report to the piazza.