Book Rating Prediction

Project definition

- Many factors to determine the book rating
- Currently: Book rating system in Google
- Not accurate enough
  - Numbers of accounts created for rating
  - Affected by noise factors
  - Need to identify what are the main factors to rate the book
- Goal: Design a better predictor
  - More accurate book rating prediction

Model Design and Analysis

1. Regression Trees

- Check the specific record
  - Check “xxx” word ⟶ check the record ⟶ (Contained): Tend to the specific score
- Advantages:
  - Visualize the step to make decision
  - Filter out noise attributes
  - Easy to set the priority
  - Lazy method ⟶ no need the hypothesis

Model Design and Analysis

2. Artificial Neural Networks

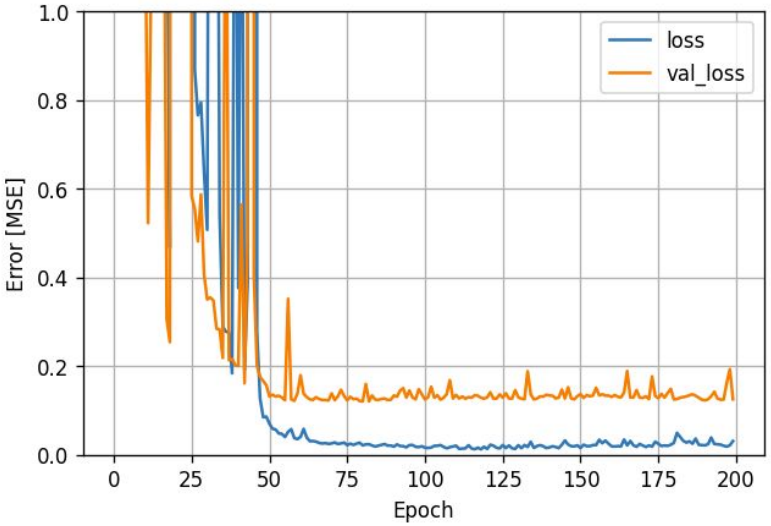
- Some data information has been set up
  - 4 layers in total
  - 1st layer: 128 neurons
  - Activation Function = RELU
  - Optimizer = Adam

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_11 (Dense)             (None, 128)               2044672
_________________________________________________________________
dense_12 (Dense)             (None, 64)                8256
_________________________________________________________________
dense_13 (Dense)             (None, 64)                4160
_________________________________________________________________
dense_14 (Dense)             (None, 1)                 65
=================================================================
Total params: 2,057,153
Trainable params: 2,057,153
Non-trainable params: 0
_________________________________________________________________
```

Model Design and Analysis

2. Artificial Neural Networks

- 80% Training, 20% for validation data
- Epochs: Set as 150 ⟶ Stabilize the mean-squared error ⟶ Accurate results (Actually after 100 starts to become stable)
- Advantages:
  - Have the ability to learn by themselves
  - (e.g. Non-linear and complex relationships)
  - Can predict the unseen data by the initial input



Data Preprocessing

12 attributes (Include target)

4 types of data

Attributes	Type	Example
publication_date	Date	10/1/2006
bookID	Identifier	28193
isbn	Identifier	380818957
isbn13	Identifier	9780380818952
Average_rating (Target)	Numerical	3.76
num_pages	Numerical	167
ratings_count	Numerical	1840
text_reviews_count	Numerical	245
title	Text / Categorical	When Santa Fell to Earth
authors	Text / Categorical	Cornelia Funke/Paul Howard/Oliver G. Latsch
language_code	Text / Categorical	eng
publisher	Text / Categorical	Chicken House / Scholastic

Data cleaning

- Remove extra white spaces of column's name
  - E.g. “ num_pages”
- Remove the record if
  - target (average_rating) is not a float number
  - target is not within 0-5
  - numerical data type attributes contain non-integer

Attributes	Type	Example
isbn	Identifier	380818957
average_rating (target)	Numerical	3.76
num_pages	Numerical	167
ratings_count	Numerical	1840
text_reviews_count	Numerical	245

Data preprocessing - publication_date

Data range from “1/1/1900” to “9/30/2006”

Since timestamp counts the date starting from 1 Jan 1970.

To avoid -ve value, this project will transform to

Attributes	Value
Year	2006
Month	9
Week	39
Day	30

Data preprocessing - duplication check

The International Standard Book Number (ISBN) is a numeric commercial book identifier which is intended to be unique.

- Use “isbn” to check the uniqueness
- Remove identifier’s attributes

Attributes	Type	Example
bookID	Identifier	28193
isbn	Identifier	380818957
isbn13	Identifier	9780380818952

Data preprocessing - language code

Some language labels contain similar meanings

- Original
  - “en-US”, “en-GB”, “en-CA”, “eng”
- Transform to
  - “eng”

Data preprocessing - one hot encoding

Convert categorical data into multiple columns depending on the categories in a column.

Only contain 1 or 0 correspondings to the categorical label

BookID	Before encoding		english	chinese	japanese
1	english	Encode->	1	0	0
2	chinese		0	1	0
3	japanese		0	0	1

Data preprocessing - one hot encoding

3 attributes adopted one hot encoding

Attributes	Type	Example
authors	Text / Categorical	Cornelia Funke/Paul Howard/Oliver G. Latsch
language_code	Text / Categorical	eng
publisher	Text / Categorical	Chicken House / Scholastic

For “authors”, “publisher”, first split by “/”

BookID	Before encoding	Encode->	Tommy	Judy	Mary
1	Tommy /Judy		1	1	0

Data preprocessing
Book title - count frequency

Break the title into Bag Of Words, 4 methods tried

1) count frequency

BookID	Before encoding	Encode->	intro	to	nn	ml	data
1	Intro to NN to ML		1	2	1	1	0

Data preprocessing
Book title - TF-IDF by library sklearn

Find out the importance of a word in book title over all the words across different book titles

2) by sklearn without stopwords

2.1) by sklearn with stopwords (English)

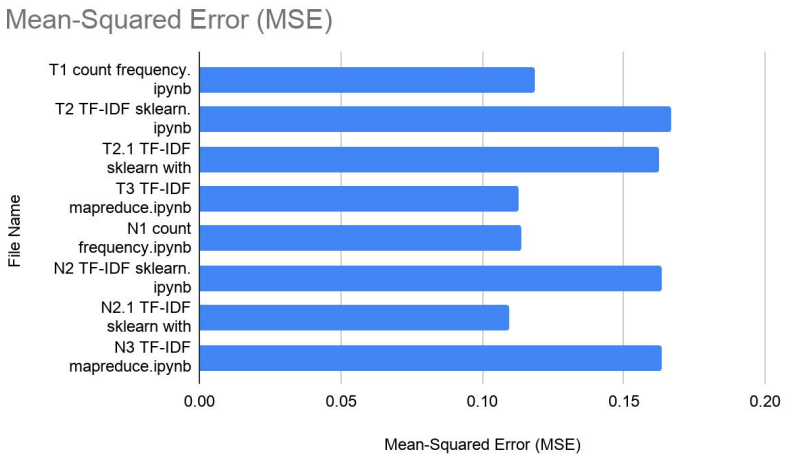
Data preprocessing
Book title - TF-IDF by MapReduce

3)

	Input/Output value	Example
Mapper input	{<bookID>: <title>}	{“25260”: “horizon problem”, “24394”: “one hercule poirot #23”}
Mapper output / Reducer input	{<word>: {<bookID>: <word count>}}	{'one': {24394: 1}, 'hercule': {24394: 1}}
Reducer output : tfidf	{<bookID>: {<word>: {<TF>, <IDF>, <TFIDF>}}}	{24394: {'one': {'TF': 1.0, 'IDF': 1.0, 'TFIDF': 1.0}, 'hercule': {'TF': 1.0, 'IDF': 1.0, 'TFIDF': 1.0}}

Performance evaluation

- Create 8 files for testing
- mean-squared error ⇩ : Better output
- Find the lowest mean-squared error (MSE)
- N2.1 TF-IDF sklearn with stopwards : using method of preprocessing → BEST algorithm



<model used><preprocessing method of "title"> xxxxx.ipynb
T = Regression tree, N = Neural network

Summary and Discussion

- Cost Function:
  - Neural Networks are better than Regression Trees
- Calculation/Running time:
  - Library’s TF-IDF is faster than using MapReduce.
- Models Training Time:
  - Regression Trees are better than Neural Networks
  - Tree method is lazy learner

~ END ~