KAGGLE PROJECT REPORT

PENGCHENG JIANG AND GANG LI

Abstract. I finished the project and got good results.

Contents

Date: (None).
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. kaggle,slides,notebook.

1. Problem Definition

~~We are given a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company. Our target is predict total sales for every product and store in the next month Submissions are evaluated by root mean squared error (RMSE).This report introduces how to participate in~~ Last year, in the Toxic Comment Classification Challenge, you built multi-headed models to recognize toxicity and several subtypes of toxicity. This year's competition is a related challenge: building toxicity models that operate fairly across a diverse range of conversations.

Here's the background: When the Conversation AI team first built toxicity models, they found that the models incorrectly learned to associate the names of frequently attacked identities with toxicity. Models predicted a high likelihood of toxicity for comments containing those identities (e.g. "gay"), even when those comments were not actually toxic (such as "I am a gay woman"). This happens because training data was pulled from available sources where unfortunately, certain identities are overwhelmingly referred to in offensive ways. Training a model from data with these imbalances risks simply mirroring those biases back to users.

In this competition, ~~and gives the whole process of how to use python language for data cleaning, feature engineering extraction and model construction.~~

2. ~~Data Cleaning~~

~~Data Infomation The above table is some information about the datas.Next, let's look at the data types of the training set which will be conducive to the subsequent processing:~~ you're challenged to build a model that recognizes toxicity and minimizes this type of unintended bias with respect to mentions of identities. You'll be using a dataset labeled for identity mentions and optimizing a metric designed to measure unintended bias. Develop strategies to reduce unintended bias in machine learning models, and you'll help the Conversation AI team, and the entire industry, build models that work well for a wide range of conversations.

2. Text preprocessing

- ~~2935849 rows, 6 columns~~
- ~~21807 items,60 shops~~
- ~~data˙type~~
  - ~~data: object~~
  - ~~date˙block˙num: int~~
  - ~~shop˙id:int~~
  - ~~item˙id:int~~
  - ~~item˙price:float~~
  - ~~item˙cnt˙day:float~~

  ~~Next, let's look at the data types of the test set:~~
  - ~~214200 rows,3 columns~~
  - ~~5100 items,40 shops~~
  - ~~data˙type~~
  - ~~ID:int~~
  - ~~shop˙id:int~~
  - ~~item˙id:int~~

~~From here you can see a lot of stores, goods in training set are not in the test setNext, we do some common processing on the data:~~

- ~~Missing Value and Non Value:Find out whether there are empty values or missing values in~~ Count the total number of words contained in all texts, the maximum and minimum number of words contained in a text
- Check for missing data
- ~~Cartesian product:for items notsold during the month, you should add them and set them to 0(Find out all the stores and merchandise, and make cartesian product with sales˙trainz)~~ Change abbreviations to full:isn't -¿ is not(via dictionnary)
- ~~Data leakages:delete stores, goods in training set but not in the test set~~ clean˙numbers
- Find all non alphabetic characters and clean˙special˙chars
- ~~Data duplication:See if duplicate items exist in the dataset~~ Solve the problem of misspelling words
- ~~Outliers:Calculate the outliers of item˙cnt˙day and item price~~ lower

3. ~~Data analysis~~

~~First, look at the monthly sales of goods at figure 1~~

comment_text	comment_text
This is so cool. It's like, 'would you want yo...	this is so cool it is like would you want y...
Thank you!! This would make my life a lot less...	thank you this would make my life a lot less...
This is such an urgent design problem; kudos t...	this is such an urgent design problem kudos t...
Is this something I'll be able to install on m...	is this something I will be able to install on...
haha you guys are a bunch of losers.	haha you guys are a bunch of losers

Figure 1. ~~month˙total˙count~~data cleaning

~~Explain that the month is related to the sales volume of goods: the sales volume at the end of the year is increasingNext, take a look at the sales of each store in figure 2Finally, let's look at the sales of different kinds of goods in figure 3~~**~~Item and Shop Information:~~**~~large categories, small categories, we separate them, and code them separately to facilitate subsequent feature extraction~~**~~Shop information:~~**~~the city where the store is located, the type of store, which we separate and encode separately for subsequent feature extractionshop˙countitem˙category˙count~~

3. ~~Model~~

2.1. ~~decision tree.~~ ~~In machine learning, decision tree is a prediction model, which represents a mapping relationship between object attributes and object values. Each node in the tree represents an object, and each branch path represents a possible attribute value, while each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The decision tree has only a single output, if you want to have complex output, you can establish an independent decision tree to deal with different outputs.Decision tree is a frequently used technology in data mining,which can be used to analyze data, and also can be used for prediction.~~

2.1. ~~Candidate model~~.

3. Embedding

- ~~GBDT~~ What tokenizer does is actually very simple. It divides the words it sees into spaces, and then uses numbers to correspond one by one. Then we take the first num˙ Words is the word with the highest frequency, others are not recognized.
- ~~Xgboost~~ First learn the dictionary of the text, and then get the corresponding relationship between words and numbers, and then convert the text into a number string through this relationship, and then use the padding method to make up the number string to the same degree, then you can proceed to the next step : embedding
- ~~lightgbm~~ collections.counter,pytorch:torchtext.vocab,
- ~~neural network~~ The embedding layer is the same as word2vec. Whether it is skip gram or cbow model, they infer each other from the context and the current, so we consider the relationship between the preceding and the following.
- glove.42B.300d.txt

3.1. ~~Method One.~~ ~~The sales of the 34th month are regarded as the sales of the 35th month, Count the sales volume of each item in each store in the 33rd month and merge it with test The result is **RMSE=1.16777**~~

3.1. ~~Method Two.~~ ~~**Features:**~~

4. module

- ~~shop˙id~~Long short term memory (LSTM) is a special RNN, which is mainly used to solve the problem of gradient disappearance and gradient explosion in the process of long sequence training. In short, LSTM can perform better in longer sequences than ordinary RNN
- ~~item˙id~~ BiRNN:In practical problems, there are also problems that not only rely on the previous sequence, but also rely on the subsequent sequence for prediction. For those problems, we need to use bidirectional RNN (birnn)
- ~~item˙cnt˙month~~ embed˙size, num˙hiddens, num˙layers = 300, 100, 2

~~The model we choosed is lightgbm, and The result is **RMSE=**~~

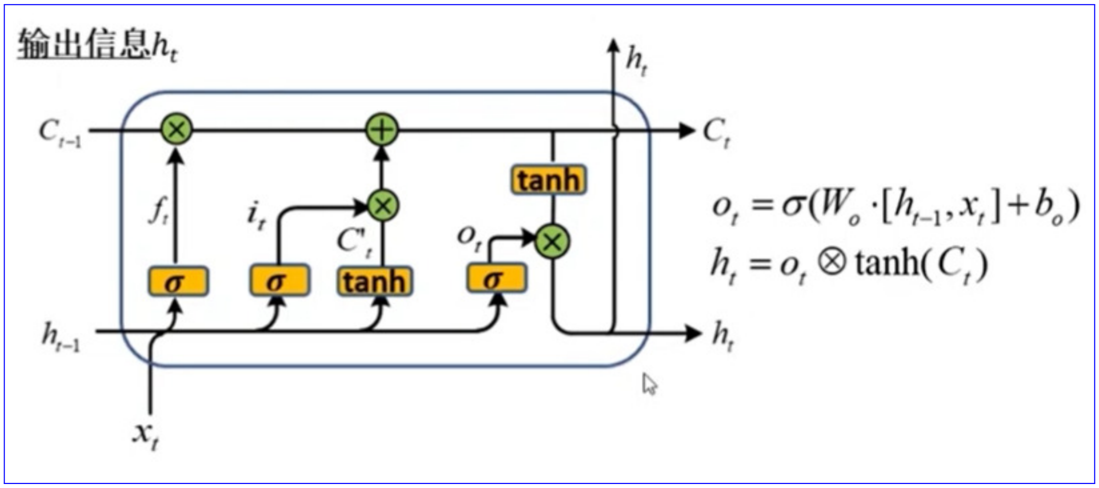

FIGURE 2. ~~final features~~LSTM

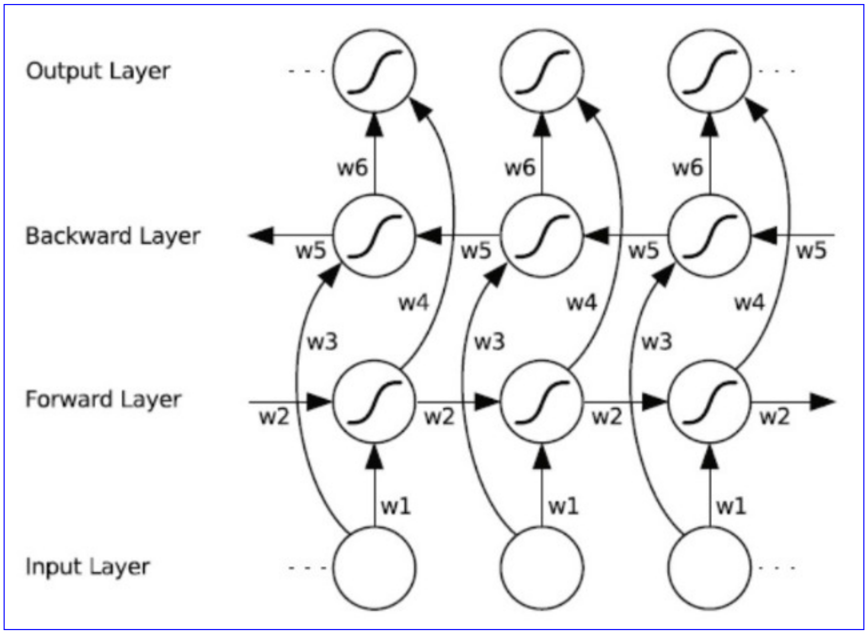

FIGURE 3. BIRNN

~~4.1. **Method Three.** Adding historical information is good for prediction. We can see the features after adding historical information in Figure 4. Finally, through experiments, it can be proved that the prediction results are improved. **training's rmse: 0.664209,valid'1's rmse: 0.880256**~~

5. Conlusion

Through ~~data processing and adding delay information~~text processing and glove embedding, the model achieves good results

(A. 1) School of Computer Science,, JiLin University, jilin 130000, China
Email address, A. 1: `pcjiang@tulip.academy`

(A. 2) School of Information Technology, Deakin University, Geelong, VIC 3216, Australia
Email address, A. 2: `gang.li@deakin.edu.au`