THE REPORT OF FLIP01 FINAL PRESENTATION

GUANZHANG HUANG

Abstract. This paper is divided into five parts, the first part mainly introduces the relevant introduction of this paper. The second part is the data preparation before the experiment, the third part is the experiment part, t he fourth part and the fifth part is the summary and analysis of this paper.

Contents

Date: (None).
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

1.1. Problem Statement.

Using 8 years daily news headlines to predict stock market movement

1.2. **Data List.** The Kaggle dataset contains date, label, and 28 columns of same-day news data. The data were divided into a training set and a test set to predict the relationship between news information and stock market movements on the day.

date: - Date information.
label: - 1 for trading, 0 for falling.
top x: - News ID.

2. Exploratory Data Analysis

2.1. Data Information.

In the table, there are date, label and 25 column top values, respectively representing the date, whether the stock market rose or fell and the news information of the day. Through the training of the data, in order to get the relationship between the rise and fall of the stock market and the news of the day.

Table 1. The head of the train data

	Date	label	Top1	...
0	2016-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	...
1	2016-08-11	1	b'Why wont America and Nato help us? If they w...	...
2	2016-08-12	0	b'Remember that adorable 9-year-old who sang a...	...
3	2016-08-13	0	b' U.S. refuses Israel weapons to attack Iran:...	...
4	2016-08-14	1	b'All the experts admit that we should legalis...	...
5	...	...	...	...

2.2. Text Processing.

Text data is data information that cannot be recognized by a computer. Only by digitizing text data can a computer be able to process the data. This chapter mainly introduces the processing method of text data.

- Get rid of the HTML tag
- Remove the punctuation
- Cut into the word /token
- stopwords
- Reorganize into new sentences

2.2.1. *Get rid of the HTML tag.* Get rid of the HTML tag is the first step in data processing to prevent information other than the data from interfering with the data.

2.2.2. *Remove the punctuation.*

To Remove the punctuation is to eliminate the interference of punctuation marks. In text messages, it is not necessary to recognize symbol information.

2.2.3. *stopwords.* In text data, there are many words that are necessary, but computer processing is not. For these words, t hey can be removed to mention the efficiency of computer recognition

3. Models

After processing natural language, it is necessary to train it. There are many training methods. In this paper, SVM,CNN and LSTM are mainly used for training.

- LSTM
- SVM
- CNN

3.1. **word2vec.**

Word2vector is a way to convert language text into a number vector. It contains the following steps:

- Word segmentation/stem extraction and morphological reduction
- Use the following five models to classify the preprocessed data
- Construct a tree structure
- Generates the binary code of the node
- Initializes the intermediate vector of each non-leaf node and the word vector in the leaf node
- Train the intermediate vector and the word vector

3.2. **model.**

Use the following five models to classify the preprocessed data

3.2.1. *LSTM.*

Train on 1611 samples, validate on 378 samples
Epoch 1/3
1611/1611 [==========] - 50s - loss: 0.6942 - acc: 0.5208 - val˙loss: 0.6935 - val˙acc: 0.5079
Epoch 2/3
1611/1611 [==========] - 51s - loss: 0.6722 - acc: 0.5971 - val˙loss: 0.6904 - val˙acc: 0.5106
Epoch 3/3
1611/1611 [==========] - 50s - loss: 0.5941 - acc: 0.7492 - val˙loss: 0.6831 - val˙acc: 0.5661
378/378 [==========] - 3s
Test score: 0.683092702633
Test accuracy: 0.566137566138
prediction accuracy: 0.566137566138

3.2.2. *CNN.*

Train on 1611 samples, validate on 378 samples
Epoch 1/1
1611/1611 [==========] - 14s - loss: 0.6900 - acc: 0.5317 - val˙loss: 0.6974 - val˙acc: 0.5079
352/378 [==========¿...] - ETA: 0sTest score: 0.69742725829
Test accuracy: 0.507936509829
prediction accuracy: 0.507936507937

3.2.3. *model score.*

The accuracy of each model is obtained through model training. Show in the table below.

Table 2. The score of models

	model	score
1	LSTM	0.566137566138
2	CNN	0.507936507937
3	svm	0.595238095238

4. Experiment and Analysis

[1] At present, we can only simply use the function call of the algorithm in NLP, and the parameter setting inside the function is still in the initial stage

5. Conclusion

[1] I need to master a variety of natural language processing methods

The next step is to improve the quality of the algorithm based on the internal principles of the function

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy