Problem Set 09

Data Mining & Machine Learning

Wen Sheng - April 29, 2016

Approach Description

Lorem ipsum dolor sit amet, ligula suspendisse nulla pretium, rhoncus tempor fermentum, enim integer ad vestibulum volutpat. Nisl rhoncus turpis est, vel elit, congue wisi enim nunc ultricies sit, magna tincidunt. Maecenas aliquam maecenas ligula nostra, accumsan taciti. Faucibus at. Arcu habitasse elementum est, ipsum purus pede porttitor class, ut adipiscing, aliquet sed auctor, imperdiet arcu per diam dapibus libero duis. Enim eros in vel, volutpat nec pellentesque leo, temporibus scelerisque nec.

Sociis mauris in integer, a dolor netus non dui aliquet, sagittis felis sodales, dolor sociis mauris, vel eu libero cras. Faucibus at. Arcu habitasse elementum est, ipsum purus pede porttitor class, ut adipiscing, aliquet sed auctor, imperdiet arcu per diam dapibus.

2. Overall Model

Lorem ipsum dolor sit amet, ligula suspendisse nulla pretium, rhoncus tempor fermentum, enim integer ad vestibulum volutpat. Nisl rhoncus turpis est, vel elit, congue wisi enim nunc ultricies sit, magna tincidunt. Maecenas aliquam maecenas ligula nostra, accumsan taciti. Faucibus at. Arcu habitasse elementum est, ipsum purus pede porttitor class, ut adipiscing, aliquet sed auctor, imperdiet arcu per diam dapibus libero duis. Enim eros in vel, volutpat nec pellentesque leo, temporibus scelerisque nec.

Sociis mauris in integer, a dolor netus non dui aliquet, sagittis felis sodales, dolor sociis mauris, vel eu libero cras. Faucibus at. Arcu habitasse elementum est, ipsum purus pede porttitor class, ut adipiscing, aliquet sed auctor, imperdiet arcu per diam dapibus.



3. Experiments & Model Tuning

Lorem ipsum dolor sit amet, ligula suspendisse nulla pretium, rhoncus tempor fermentum, enim integer ad vestibulum volutpat. Nisl rhoncus turpis est, vel elit, congue wisi enim nunc ultricies sit, magna tincidunt. Maecenas aliquam maecenas ligula nostra, accumsan taciti. Faucibus at. Arcu habitasse elementum est, ipsum purus pede porttitor class, ut adipiscing, aliquet sed auctor, imperdiet arcu per diam dapibus libero duis. Enim eros in vel, volutpat nec pellentesque leo, temporibus scelerisque nec.

Sociis mauris in integer, a dolor netus non dui aliquet, sagittis felis sodales, dolor sociis mauris, vel eu libero cras. Faucibus at. Arcu habitasse elementum est, ipsum purus pede porttitor class, ut adipiscing, aliquet sed auctor, imperdiet arcu per diam dapibus.

Without special specification, the model used in experiments is specified here:

- Experiment Model Description Here:
- Embedding Layer (drop out 0.25)
- SimpleRNN layer
- Dense Layer(256 nodes, drop out 0.25, activation: relu)
- Output Dense Layer (32 batches, 10 epochs, early stop)
- Total number of top words: 5000
- Max length of each review: 3000

3.1. Choosing Max Length (number of words) for each review

- Experiment Model Description Here:
- Embedding Layer (drop out 0.25)
- SimpleRNN layer
- Dense Layer(256 nodes, drop out 0.25, activation: relu)
- Output Dense Layer (32 batches, 10 epochs, early stop)
- Total number of top words: 5000

We can see the pattern in this experiment: With the max length of each review increases, the classification also increases. Because we want the highest classification rate and using max length of 250 will also finish in acceptable time span, I choose 250 as the max length of each review.

Max Length	50	100	150	200	250
Classification rate	0.8104	0.8337	0.8375	0.8363	**0.8485**

3.2. Choosing Total number of top words

- Experiment Model Description Here:
- Embedding Layer (drop out 0.25)
- SimpleRNN layer
- Dense Layer(256 nodes, drop out 0.25, activation: relu)
- Output Dense Layer (32 batches, 10 epochs, early stop)
- Max length of each review: 250

The result shows when the total number of top words increases from 500 to 8000, the classification start to increase first, then it get to a steady rate around 0.835 - 0.845. Because of this pattern, I choose 5000 as the total number of words in the final model because it both has **high classification rate** in considerable **fast** in computing process.

Total Number	500	1000	2500	3000	4000	5000	6000	7000	8000
Classification rate	0.7901	0.8052	0.8384	0.8403	0.8371	**0.8436**	0.8192	0.8428	0.8389

3.3. Choosing types of RNN model

- Experiment Model Description Here:
- Embedding Layer (drop out 0.25)
- Dense Layer(256 nodes, drop out 0.25, activation: relu)
- Output Dense Layer (32 batches, 10 epochs, early stop)
- Max length of each review: 250
- Total number of top words: 5000

According to the result, the type of LSTM RNN model has the highest classification rate.

TYPE	Simple RNN(16)	LSTM(32)	GRU(32)
Classification rate	0.8446	**0.8538**	0.8461

3.4 Choosing Number of Nodes in RNN layer

Number of Nodes	4	8	16	32	64	128
Classification rate	0.8459	0.8608	0.8663	0.8538	0.8595	